UNIVERSIDAD DE GRANADA

Departamento de Economía Financiera

y Contabilidad

TESIS DOCTORAL

Una perspectiva webmétrica del estudio de empresas.

Aplicación al estudio de variables financieras en empresas con presencia en la Web

Tesis presentada por:

Esteban Romero Frías

Dirigida por:

Prof. Dr. Lázaro Rodríguez Ariza

Prof. Dra. Liwen Vaughan

Granada, 2010

UNIVERSITY OF GRANADA

Department of Accounting and Finance

PhD THESIS

A Webometric Approach to Business Studies. Application to the Study of Accounting and Financial Variables in Companies with Presence in the Web

Author:

Esteban Romero Frías

Supervisors:

Prof. Dr. Lázaro Rodríguez Ariza

Prof. Dr. Liwen Vaughan

Granada, 2010

AUTORIZACIÓN PARA PRESENTACIÓN DE TESIS

D./Dña: LAZARO RODRÍGUEZ ARIZA & LIWEN VAUGHAN

Director/es de la Tesis: UNA PERSPECTIVA WEBMÉTRICA DEL ESTUDIO DE EMPRESAS. APLICACIÓN AL ESTUDIO DE VARIABLES FINANCIERAS EN EMPRESAS CON PRESENCIA EN LA WEB.

de la que es autor D./Dña.: Esteban Romero Frías

Programa de Doctorado: Doctorado en Contabilidad (realizado en la Universidad de Valencia)

AUTORIZA la presentación de la referida Tesis para su defensa y mantenimiento de acuerdo con lo previsto en el Real Decreto 56/2005, de 21 de enero, emitiendo el siguiente informe:

La tesis de referencia alcanza de forma altamente satisfactoria todos los requisitos requeridos para proceder al acto su lectura y defensa, tanto por la novedad del tema elegido, revisión de la bibliografía realizada, metodología empleada y conclusiones alcanzadas.

Y para que conste y surta sus efectos en el expediente correspondiente, expido la presente en

Granada, 6 de ABRIL de 2010.

Fdo.: LAZARO RODRÍGUEZ ARIZA & LIWEN VAUGHAN

A mi madre,

a mi padre,

a mi hermana,

a mi Mamu.

«Heródoto viaja con el fin de encontrar una respuesta a su pregunta de niño: ¿cómo es que en el horizonte aparecen naves? ¿De dónde han salido? ¿De qué puerto han zarpado? O sea que lo que vemos con nuestros propios ojos, ¿no es aún el límite del mundo? ¿Hay otros mundos todavía? ¿Cómo son? Cuando crezca, querrá conocerlos. Aunque más vale que no crezca del todo, que conserve un poco de ese niño curioso que es, pues sólo los niños plantean preguntas importantes y de verdad quieren aprender.

Y Heródoto, con su entusiasmo y apasionamiento de niño, parte en busca de esos mundos. Y descubre algo fundamental: que son muchos y que cada uno es único.

E importante.

Y que hay que conocerlos porque sus respectivas culturas no son sino espejos en los que vemos reflejada la nuestra. Gracias a esos otros mundos nos comprendemos mejor a nosotros mismos, puesto que no podemos definir nuestra identidad hasta que no la confrontamos con otras.

Por eso, después de hacer este descubrimiento -otras culturas como espejo en que mirarnos para comprendernos mejor a nosotros mismos-, cada mañana a la salida del sol, incansablemente, Heródoto reanuda su viaje.»

Ryszard Kapuściński

Viajes con Heródoto *(2004/2008: 296-297)*

«Vínculo a vínculo construimos sendas de entendimiento a través de la red de la humanidad. Somos los hilos que cohesionan el mundo.»

Tim Berners-Lee

Tejiendo la red *(1999: 189)*

Inventario de escenarios y agradecimientos

Los escenarios. Los meses en Bruselas; la pasión política; los libros de Kapuściński; la música de Los Secretos, U2, Coldplay, Aute o Antonio Vega; los sueños de Paul Auster; los girasoles ciegos y los corazones helados; los conciertos; Madrid; las tierras nevadas de Ontario; los puertos y ciudades de Malta; una Sicilia torrencial; las Midlands británicas; el invierno en Praga; los reencuentros en Valencia; y, siempre Granada como el lugar al que acabar volviendo, escenario de una vida que tiene por telón de fondo montañas nevadas y una vega donde naufragan los atardeceres.

Los agradecimientos. Quiero agradecer con las siguientes palabras, inevitablemente breves, a aquellas personas que, en la cercanía y en la distancia, en cada uno de los diferentes escenarios de estos años, han sido más pacientes y sabias que yo, habiendo sido capaces de estar a mi lado y de contribuir a hacer que el improbable día de hoy se convierta en una realidad.

En primer lugar, agradezco a mis directores de tesis, Lázaro Rodríguez Ariza y Liwen Vaughan, su ayuda y consejo que me ha permitido hacer realidad este proyecto. Quiero reconocer expresamente, no solo su paciencia para coordinar mi trabajo, sino también su espíritu constructivo y su flexibilidad para dejarme coordinar las aportaciones de ambos, confiando en mi criterio para enlazar sus contribuciones transatlánticas. Lázaro ha sido siempre efectivo en su motivación, pragmático y efectivo en sus perspectivas desde el área de la contabilidad y las finanzas, así como sincero en sus juicios, animándome en la aventura de caminar por sendas que a ambos nos eran desconocidas. Liwen, por su parte, merece una mención especial por su participación desde Canadá. Agradezco profundamente su generosidad y valentía para aceptar trabajar conmigo y codirigir esta tesis. Pasé un curso entero leyendo artículos sobre investigación en Internet hasta que, finalmente, descubrí sus trabajos. Llegaron entonces los correos electrónicos, mi estancia en la University of Western Ontario y un año de trabajo conjunto. Mi relación con Liwen representa una de las mejores lecciones en el plano académico y personal que me llevo de estos años de tesis, principalmente por lo inesperado de la misma. Desde el área de Ciencias de la Información, ella me ha enseñado el camino de la investigación y de la publicación,tratándome como a un igual al tiempo

que con la paciencia y sabiduría de las mejores maestras. Sin Lázaro ni Liwen, cada cual a una orilla del océano, esta tesis no habría sido posible. Mi gratitud hacia ellos es inmensa.

En el camino de esta tesis quiero agradecer también a Mike Thelwall su generosidad por haberme abierto las puertas de su grupo de investigación en la University of Wolverhampton (Reino Unido). Agradezco también a los miembros del tribunal el haber aceptado la invitación de venir a Granada para examinar esta tesis doctoral, así como a Sarah Cooper, de la University of Edinburgh, y a Lúcia Maria Portela de Lima Rodrigues,de la Universidade do Minho, por su evaluación positiva de este trabajo. Gracias también a Jorge Chica por aclarar mis dudas econométricas y a Nacho Belda por ayudarme, siempre tan dispuesto, en el diseño de esta tesis.

Doy las gracias a mis compañeros y compañeras del Departamento de Economía Financiera y Contabilidad por su apoyo, ayuda y comprensión a lo largo de estos cuatro años que hemos compartido juntos. De manera especial, quiero nombrar a Ramón García-Olmedo, profesor, cuyas clases hoy en día siguen iluminando las mías, y compañero y amigo, que con esa influencia involuntaria que ejercen los verdaderos maestros me ha acercado a la contabilidad internacional y a la fiscalidad y me ha abierto la puerta de proyectos y de nuevas amistades. Mis primeros pasos en la Universidad de Granada se han debido a él, que me presentó a Antonio López, a quien agradezco el haberme apoyado en mis primeros pasos en el Departamento. Por último, mencionaré a José Manuel Aguayo. Se trata de un reconocimiento que se remonta a tiempos más recientes. Su talante, su capacidad de escucha, su apoyo desde la Dirección del Departamento han sido indispensables en este último año. Con todo, no es lo institucional lo más importante para mí, sino una amistad silenciosa que ha hecho que, en días de cansancio y agotamiento, una conversación, con cualquier excusa de por medio, me permitiera seguir adelante.

Gracias también a mis profesores del Departamento de Contabilidad de la Universidad de Valencia y de manera muy especial a mis amigos Gregorio Labatut, Rafael Molina y Miguel Arce. Gregorio fue mi director de tesina junto con Rafa. Me ha brindado su amistad y me ha enseñado la honestidad intelectual y el valor inestimable de la bondad en la labor académica y en la dimensión personal.

También quiero agradecer a Araceli Mora el haber confiado en mí para participar en distintos proyectos y por haberme dado siempre referencias útiles para mi trabajo. Una de ellas fue la de José Luis Arquero, de la Universidad de Sevilla, al que quiero reconocer su espíritu siempre positivo, así como su disponibilidad para trabajar juntos, para pasarlo bien trabajando juntos. Precisamente a estos proyectos compartidos en educación les debo el haber descansado de la tesis con el siempre efectivo método de asumir más trabajo.

Por último, quiero recordar también a mis compañeras que ahora se encuentran en plena batalla por realizar su tesis. Mención especial hago a Paco Alcaraz, generoso compañero de despacho e inmerso también en la lucha de la tesis. Les diría que nunca se dejen llevar por la desazón de no percibir progresos en su trabajo, cada paso, aunque aparentemente sea hacia atrás, nos lleva más cerca de la meta.

No quiero olvidar a mis compañeros y compañeras que, ya doctores, han sido una fiel compañía estos años; entre otros, Javier Delgado, Vanesa Barrales, Mª Nieves Pérez, por compartir comidas, libros, música, penas y alegrías, en definitiva. Entre ellos, un reconocimiento especial para mi amistad con Ana, con quien me unen horas de inquietudes y confidencias, siendo profunda mi admiración.

Deseo acordarme de mi alumnos y alumnas, a los que espero deberme muchos años. Ellos han sido la inspiración y el motor de proyectos como *Descuadrando.com* u otras innovaciones docentes para animar su estudio y motivar mis clases. Me considero afortunado de poder enriquecerme con la docencia y de poder compartir los limitados conocimientos que poco a poco voy alcanzando.

Quiero agradecer la presencia a muchos amigos y amigas que a lo largo de estos años han hecho que el trabajo valga la pena, a los que han estado siempre, a los que he recuperado, a los que se han distanciado, a los que veo poco, a los recién llegados, a los clásicos de Granada, a los valencianos, a los de Bruselas (esto es, a los de todas partes), a los de London y Wolverhampton. Nobleza obliga a pronunciar el nombre de tres de ellos: Natalia, Antonio y Anca.

A Natalia Reyes, por ser compañera del día a día, aquí y allí y en todas partes, porque la amistad existe y nosotros estamos convencidos de que hemos de vivir para contarlo, aunque sólo sea el uno al otro.

A Antonio Zugaldía, que es una de esas pocas personas que constituye una influencia ineludible en mi vida. Estar junto a él me ha llevado a conocer más el mundo de la universidad, me abrió las puertas de la tecnología y de la Web 2.0, animó mi pasión política y fue instigador de nuestra aventura europea en Bruselas. Esta tesis y mucho de lo que hago y de lo que llegaré a hacer se lo debo.

A Anca Adamescu, que fue testigo desprevenida del ensimismamiento académico, del proyecto totalitario en el que de algún modo parece convertirse la tesis doctoral. Ella me mostró, y quizá también aprendió junto a mi, los límites de la vida personal y de la profesional, haciéndome ser mejor persona.

Me reservo para el final el agradecimiento a mi familia en un sentido amplio y el recuerdo para mis abuelos que ya no están y que hoy estarían orgullosos de verme alcanzar este logro.

Esta tesis, con sus esfuerzos y desvelos, está dedicada a mis padres, a mi hermana y a mi Mamu, que son los pilares de mi vida.

Nada habría sido posible sin mi madre ni mi padre. No me refiero a la tesis, sino a ninguno de los objetivos que he podido alcanzar en mi vida ni a los tropiezos de los que me han ayudado a levantarme. De ellos he aprendido el valor del compromiso, de la dignidad y de la honestidad, la justicia y el talante, la mesura y la decisión, la responsabilidad y la humildad. Ellos han sido siempre testigos de mis preocupaciones, de mis ausencias y de mi mal humor. Quizá no hay modo de compensar esto; ellos tienen todo mi agradecimiento y amor.

A mi hermana, que es mi mejor compañera de viaje, por compartir tantas esperanzas y melancolías, mientras escuchamos a Los Secretos, comemos sushi en el centro de Madrid o bajamos en autobús urbano por la Castellana soñando conque algún día la línea tenga parada en Granada.

A mi Mamu, con quien me une un amor inquebrantable. Con toda mi alegría por su compañía y por todo lo que nos queda por vivir.

A todos ellos les he dicho muchas veces que me quedaba ya poco para acabar este

trabajo y poder dedicarles más tiempo. Yo tampoco pensaba que llegaría este momento, pero ahora que ha llegado quiero que sepan que es más suyo que mío, porque sin ellos yo no lo quiero para nada.

Índice de contenidos

Índice de tablas

Índice de figuras

1. INTRODUCCIÓN

«La armonía invisible es mayor que la armonía visible.»

Atribuida a Heráclito de Éfeso

La Web no ha existido siempre.

Pocas creaciones humanas han conseguido en apenas dos décadas convertirse en un elemento tan característico de nuestra cultura; llegando a alcanzar un grado de invisibilidad propio de invenciones que nos han acompañado durante un tiempo relativamente largo. Los orígenes de Internet, hace 40 años, y de la Web, 20 años atrás, son desconocidos para la gran mayoría, olvidando que hace tan sólo medio siglo todo este desarrollo era un sueño poco más que imposible. La Web e Internet han dejado de ser un "invento" para convertirse en un contexto, en espacio ubicuo donde ocurren cosas, donde transcurre nuestra vida, a través de todo tipo de pantallas, de un portátil, de un móvil o de un televisor, entre una variedad de dispositivos cada vez más amplia. Aunque creamos que no estamos conectados, aunque nos empeñemos en no participar en las redes, la redes interactúan con nosotros: nuestro yo virtual y la presencia digital de empresas e instituciones siguen activos aún cuando cancelamos nuestra conexión. Que la Web no ha existido siempre no es una mera obviedad, sino un ejercicio de extrañamiento, una declaración de intenciones para salir de este espacio que lo abarca casi todo con el

objeto de observarlo, desde fuera, pero también desde dentro.

En la última década, tras la famosa burbuja financiera de las empresas *puntocom*, la Web no ha dejado de crecer a la par que lo han hecho sus posibilidades de uso y el número de usuarios, atraídos por una rica oferta de servicios de todo tipo y por una reducción generalizada de los costes de conexión y de los equipos informáticos. La Web, principalmente a lo largo del último lustro, ha dado el salto, del ordenador personal a una amplia gama de dispositivos (teléfonos, libros electrónicos, reproductores de música y vídeo, videoconsolas, etc.) que no pueden permitirse quedar fuera de la red si quieren seguir siendo comercialmente atractivos. El poblamiento de la Web ha ocasionado que una gran mayoría de actividades sociales, económicas, políticas, educativas, de toda índole, que únicamente se desarrollaban en el mundo físico, encuentren un fiel reflejo en ella o hayan desarrollado manifestaciones genuinamente *online*. Sin este espectacular desarrollo a lo largo de los últimos 20 años, no podríamos explicar la sociedad en la que actualmente vivimos: las formas de interacción social, el trabajo, la transformación de los sectores económicos, el reajuste de poder entre distintos países del mundo o las nuevas brechas sociales en el contexto de la sociedad del conocimiento. No entenderíamos importantes amenazas para nuestra sociedad, tales como el terrorismo o nuevas formas de acoso y chantaje; pero tampoco contaríamos con poderosas armas de transformación social, de protesta, de movilización y solidaridad. Hablar de la World Wide Web, red de redes desarrollada sobre Internet, es hablar de nosotros mismos, que dejamos nuestros rastros cada vez que nos conectamos frente a nuestro portátil, con el teléfono móvil o desde cualquier otro dispositivo multimedia. Las redes sociales forman parte de nuestra rutina diaria, nos informamos a través de nuevas fuentes, consultamos enciclopedias que no tienen un autor conocido. Lo antiguo vuelve a ser vigente, pero de otra forma. De nuevo reconocemos formas de autoridad basadas en el mérito y la confianza, aunque en este tiempo puede que jamás hayamos visto a la persona en cuestión. La "aldea global", sugerida por Marshall McLuhan, se puede desarrollar a través de formas nunca antes imaginadas. Los medios de masas permitieron abrir y agrandar nuestro mundo, haciéndolo accesible a los ojos de millones de ciudadanos; Internet y la Web, por su parte, han permitido que esos

ciudadanos dispongan hoy en día de medios tecnológicos con capacidades análogas a las de los medios de masas. A un coste muy reducido, Internet constituye un vehículo de desarrollo para muchos de los excluidos de este mundo.

De acuerdo con los datos del informe *La Sociedad en Red. Informe anual 2008*, elaborado por el ONTSI (Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información, 2009), en 2008 existían 1.596 millones de usuarios de Internet en el mundo, lo cual representaba una penetración del 23,8%. En Norteamérica la tasa de penetración de Internet alcanzaba el 74,4%. Asia, por su parte, constituía el continente que aportaba el mayor número de internautas en términos absolutos, un 41,2% del total. En el ámbito empresarial, tanto el número de firmas con presencia en la Web como el número de transacciones económicas realizadas se han ido incrementando significativamente en los últimos años. Las empresas no han quedado al margen del fenómeno de la Web e Internet. En España (ONTSI, 2009), el 95% de las empresas de 10 ó más empleados tiene acceso a Internet. De ellas el 57,5% dispone de página web, incrementándose el porcentaje conforme aumenta su tamaño (por ejemplo, un 89,2% de las empresas de más de 250 trabajadores). El volumen de actividad económica realizada en la Web ha crecido notablemente debido al aumento en el número de usuarios de Internet, lo cual implica un mayor número de clientes potenciales. Desde hace poco más de un lustro, la presencia de millones de nuevos internautas ha sido alimentada por el desarrollo de la Web 2.0 (O'Reilly, 2005), que facilita el acceso a contenidos multimedia, la interacción de los usuarios y la creación de contenidos sin necesidad de poseer grandes conocimientos tecnológicos. Algunos de los servicios 2.0 más populares son los blogs, redes sociales, wikis, entre otros. La facilidad para generar contenidos ha provocado que el volumen de información disponible en la Web se haya multiplicado, constituyendo una potencial fuente de información que debemos aprender a utilizar para obtener nuevos conocimiento en los más diversos campos: la ciencia, la política, el gobierno electrónico y el mundo de la empresa, por poner unos ejemplos.

Las líneas anteriores pretenden poner de relieve el interés que nos ha motivado a explorar una realidad tan fascinante como relevante en nuestro tiempo. La Web es,

sin embargo, un espacio cambiante y esquivo, difícil de aprehender y de observar. Por ello, hemos buscado enriquecer nuestro acervo metodológico mediante el empleo de nuevos métodos de investigación en nuestra área de conocimiento. Partimos del convencimiento de que la interdisciplinariedad es prácticamente una apuesta obligada en nuestro tiempo para abrir nuevos espacios de conocimiento y diálogo. Especialmente cuando el objetivo perseguido no es reincidir en el estudio de un espacio acotado sino realizar una pequeña aportación en el vasto dominio de una realidad que no entiende de parcelas ni fronteras académicas. Los investigadores llevan dos décadas prestando atención a la Web, desde distintas perspectivas y puntos de vista. La Web, como contaremos en el Capítulo 2, nace directamente vinculada a la ciencia, como un instrumento de comunicación del conocimiento. La Web e Internet en su conjunto, en tanto que objetos de estudio, permiten investigar tanto fenómenos completamente nuevos, que no existirían sin estas redes, como fenómenos que, existiendo en el mundo puramente físico, encuentran reflejo y se desarrollan cada vez más significativamente en estos entornos virtuales. Si bien en ocasiones se sigue confrontando "mundo real" frente a "mundo virtual", la separación entre ambos se nos antoja ficticia: lo *online* y lo *offline* son dimensiones de una misma realidad, de un único mundo.

Un componente fundamental en toda esta amalgama de información disponible en la Web son los hiperenlaces, que, de una forma discreta, tejen, utilizando la metáfora de su creador Tim Berners-Lee (1999), las conexiones que entrelazan las distintas páginas o documentos de todo tipo existentes en la Web. Representan un elemento básico para comprender las posibilidades de la investigación en la Web y, en concreto, de la investigación basada en la aplicación de técnicas webmétricas. Un hipervínculo, hiperenlace o enlace, como emplearemos de forma más frecuente, se puede definir como una referencia que una página web establece a una sección de esa misma página o a otra página web completamente distinta. Los enlaces conforman la estructura de la Web (Berners-Lee, 1999), ya que sin ellos resultaría imposible acceder a los contenidos alojados en las páginas, a no ser que conociéramos la dirección específica de dicho recurso (conocida como su URL). Si bien algunas voces han señalado la naturaleza caótica y desordenada de la Web (Keen, 2007), la ausencia de estructura es sólo aparente: cada enlace constituye

una fuente potencial de información sobre el modo en que se haya distribuida. Un buen ejemplo de ello es la industria de los motores de búsqueda o buscadores (Battelle, 2006; Machill, Beiler y Zenker, 2008), que tienen al análisis de hiperenlaces como una de las bases para la detección y ordenación de la información. Éste es precisamente el origen del éxito de Google, líder mundial del sector, cuyo sistema de ordenación de la información en función de su relevancia, el *PageRank* (Brin y Page, 1998), se basa fundamentalmente en dos principios: el primero es que cada enlace que se establece a una página web constituye en cierta forma un voto que se otorga a la misma, una llamada de atención, una cita o referencia; el segundo es que no todos los enlaces tienen el mismo valor. Los enlaces provenientes de páginas relevantes, con una buena reputación, tendrán más valor que aquellos que proceden de páginas poco importantes.

A pesar del carácter distribuido y no estructurado de la Web, el desarrollo de herramientas para agregar y analizar las numerosas acciones de miles de individuos nos permite extraer información que puede ser de gran utilidad para las empresas (Choi y Varian, 2009). En este contexto, la Webmetría ofrece una serie de técnicas, principalmente inspiradas en la Bibliometría, que permiten llevar a cabo un estudio cuantitativo de la Web.

1.1. La Webmetría

La Webmetría ha sido definida por Björneborn (2004) como el estudio de los aspectos cuantitativos de los recursos, estructuras y tecnologías de la información en la Web empleando un enfoque principalmente bibliométrico. Los estudios webmétricos toman como principales variables de estudio la cuantificación de términos de consulta (por ejemplo, el número de páginas web que mencionan el término "Repsol YPF") y el número de enlaces que contiene o que recibe una página web (por ejemplo, el número de páginas que enlazan la página web http://www.repsol.com). La medición del número de hiperenlaces es la variable utilizada con mayor frecuencia, ya que permite delimitar claramente el objeto de

estudio. Además, la creación de un hiperenlace es, por lo general, resultado de un acto deliberado y significativo.

La idea, mencionada anteriormente, de que un enlace a una página web representa de alguna manera un voto a favor de la misma se encuentra ya en la base de los sistemas de citación y de elaboración de *rankings* de impacto en el campo de la investigación científica (Garfield, 1979). De ahí que sea desde el campo de ciencias bibliométricas desde donde, de forma natural, se han desarrollado algunos de los primeros estudios que se centran en la investigación cuantitativa de la Web. En ellos se han puesto de manifiesto relaciones entre variables *online* y *offline*, como son, por ejemplo, los enlaces recibidos por una página o sitio web y diversas medidas de desempeño de la actividad investigadora. Así, se han estudiado las relaciones entre enlaces recibidos por sitios web de universidades con los niveles de investigación alcanzados por las mismas (Smith y Thelwall, 2002), enlaces a sitios web de facultades con su productividad investigadora (Chu, He y Thelwall, 2002), y enlaces a sitios web de revistas con las calidad de las revistas (Vaughan y Hysen, 2002; Vaughan y Thelwall, 2003).

En el ámbito empresarial destacan fundamentalmente los trabajos de la profesora Liwen Vaughan, que ha explorado el empleo de los enlaces que apuntan a los sitios web de empresas como indicadores del desempeño o de la posición financiera de dichas entidades.

1.2. Motivación y objetivos: preguntas de investigación

Los trabajos de la profesora Vaughan han inspirado de manera decisiva esta tesis doctoral, que pretende, desde del campo de los estudios de empresas, profundizar en las evidencias obtenidas hasta el momento, así como abrir y proponer nuevas líneas de investigación. Las preguntas de investigación que nos hemos planteado se resumen en cinco formulaciones generales.

1.2.1. Pregunta de investigación 1: ¿En qué medida se relacionan las variables financieras clave de una empresa con su presencia en la Web, medida a través del número de enlaces que reciben sus sitios web?

Las investigaciones llevadas a cabo hasta el momento (Vaughan, 2004a, 2004b; Vaughan y Wu, 2004) han encontrado evidencia de correlaciones significativas y positivas entre el número de enlaces que recibe el sitio web de una empresa y sus variables financieras, tanto de posición como de desempeño financiero. Sin embargo, estos trabajos se han centrado casi exclusivamente en el sector de tecnologías de la información, cuyas empresas, por la propia naturaleza de su actividad, tiene una presencia en la Web más intensa que empresas de otros sectores. Por ello, hemos explorado sectores de diversa índole, tales como la construcción, la hostelería y restauración, los bancos o las grandes cadenas de distribución (Apartado 4.1).

Por otra parte, los trabajos llevados a cabo previamente se han concentrado geográficamente en empresas de Canadá, China y Estados Unidos, sin haberse explorado todavía la situación en Europa. Nuestro trabajo (Apartado 4.1.2) busca también solventar esta carencia. Adicionalmente, la regulación comercial en la Unión Europea nos permite disponer de datos públicos de empresas no cotizadas, con lo cual es posible comprobar qué ocurre con la vinculación entre variables cuando estudiamos empresas más pequeñas que no están sujetas a la estricta regulación de los mercados de valores.

1.2.2. Pregunta de investigación 2: ¿Cuáles son las variables financieras que explican el número de enlaces que recibe el sitio web de una empresa?

Los trabajos anteriormente mencionados se han centrado fundamentalmente en la

identificación de relaciones entre variables, mediante el empleo de correlaciones simples. En la tesis (Apartado 4.1.3) pretendemos identificar, por primera vez, cuáles son las variables que explican la presencia en la Web de las empresas, medida a través del número de enlaces que reciben sus sitios web. Para ello, se ha llevado a cabo un análisis de regresión múltiple sobre empresas de España y Reino Unido pertenecientes a diversos sectores de actividad.

1.2.3. Pregunta de investigación 3: ¿Cómo se interrelacionan empresas pertenecientes a sectores de actividad distintos analizadas a través de su estructura de enlaces en la Web?

Una línea importante de investigación webmétrica en empresas son los análisis competitivos de sectores basados en los co-enlaces que vinculan los sitios web de las empresas en estudio. Los co-enlaces son páginas web que incluyen simultáneamente enlaces a dos sitios web contemplados en la investigación. Hasta ahora los estudios realizados (Vaughan y You, 2006; 2008; 2009) han seleccionado conjuntos de empresas pertenecientes a un único sector, generalmente vinculados a actividades tecnológicas, y han visualizado en un mapa las posiciones relativas de las mismas en función del número de co-enlaces que las relacionan. Cuanto más próximas se encuentran dos empresas mayor elementos en común comparten. Esta similitud, entendida en un contexto de organizaciones pertenecientes a un mismo sector, se interpreta como una medida del grado de competencia entre ambas: cuanto más próximas entre sí se hallen, más intensa será su relación competitiva.

En este caso, nos preguntamos qué ocurre cuando empresas de distintos sectores, de naturaleza heterogénea, se analizan conjuntamente. Para ello, en el Apartado 4.2.1, hemos investigado las empresas que componen los principales índices bursátiles del mundo, detectando patrones y proporcionando explicaciones que pretenden avanzar en la comprensión de la evolución de los diversos sectores de actividad conforme se va desarrollando una economía cada vez más basada en la información y el conocimiento.

1.2.4. Pregunta de investigación 4: ¿Podemos emplear la información sobre enlaces para estudiar eventos económicos determinados?

Para contestar a esta pregunta, hemos llevado a cabo un análisis de co-enlaces complementado con el empleo de palabras clave, que tienen por objeto filtrar los resultados obtenidos y refinar el tipo de información que se recopila para el análisis. Se trata de un método ensayado previamente por Vaughan y You (2008) con éxito. El objetivo (Apartado 4.2.2) es analizar la evolución de la crisis financiera en los bancos de los Estados Unidos.

1.2.5. Pregunta de investigación 5: ¿Cómo se pueden combinar distintos tipos de análisis webmétrico para ofrecer una visión global de un sector empresarial?

En los trabajos previos sólo se había empleado un único procedimiento webmétrico en el análisis, ya fuera el análisis de impacto de enlaces o el basado en co-enlaces. Con el fin de avanzar en la integración de ambos tipos de análisis se han estudiado los principales bancos mundiales desde ambas perspectivas (Apartado 4.3), obteniendo resultados que permiten ofrecer una imagen más completa del sector.

1.3. Estructura de la tesis

La tesis se divide en cinco capítulos. El primero, en el cual nos encontramos, presenta el contexto general en el que se enmarca la investigación, así como las preguntas que han constituido el motor de la misma. Son cinco preguntas de investigación para las cuales se han desarrollado diversos estudios independientes, aunque vinculados entre sí por la temática y la contribución general que realizan a

los objetivos de la tesis. Estos objetivos principales son: avanzar en el análisis webmétrico de empresas, profundizar en las evidencias obtenidas previamente, generar nuevas preguntas y líneas de trabajo, y, por último, presentar un conjunto de técnicas de investigación que, a nuestro juicio, pueden proporcionar nuevas herramientas para los estudio de empresa, especialmente en temas vinculados a la Web e Internet, así como para abordar cuestiones relativas a la economía del conocimiento y la identificación, medición y gestión de elementos intangibles de esta naturaleza en la empresa.

El segundo capítulo aborda la importancia social de Internet y la Web, haciendo un repaso histórico a sus orígenes y su desarrollo hasta llegar a la actualidad. Se trata de un capítulo que pretende poner en valor el significado y las repercusiones de la Web, desde un punto de vista social, histórico, tecnológico y cultural. Se trata de una parte importante de esta tesis, en tanto que supone el reconocimiento de la existencia de un vasto espacio para la investigación, algo que va más allá de la consideración de un área de estudio acotada y claramente delimitada. Se pretende también subrayar que la perspectiva webmétrica es sólo una de otras muchas posibles, tantas casi como problemas y fenómenos se producen en la Web e Internet. Por último, se trata de un capítulo que pretende persuadir e invitar a otros investigadores al estudio de este fascinante contexto.

El tercer capítulo aborda la Webmetría como disciplina, abordando su origen, técnicas, forma de obtención de los datos para la investigación y limitaciones, entre otras cuestiones. Finalmente, se incluye la revisión de la literatura, incidiendo fundamentalmente en la investigación webmétrica de empresas.

El capítulo cuarto aborda la investigación empírica. El Apartado 4.1 comprende diversos trabajos basados en el análisis de impacto de enlaces en la Web. El Apartado 4.1.1 examina las correlaciones entre variables de enlaces a sitios web de empresas y variables financieras en diversos sectores de los Estados Unidos. El Apartado 4.1.2 incluye un estudio similar para empresas de España y Reino Unido pertenecientes a diversos sectores. En este caso, la mayoría de empresas no son cotizadas, a diferencia de lo que ocurre en Estados Unidos. Por último, el Apartado

4.1.3, lleva a cabo un análisis de regresión múltiple basado en los datos del apartado anterior, con el fin de identificar las variables que explican el número de enlaces que recibe el sitio web de una empresa.

La Sección 4.2 incluye dos estudios basados en el análisis de co-enlaces. El Apartado 4.2.1 recoge un trabajo basado en el análisis de empresas pertenecientes a distintos sectores de actividad. Se analizan las empresas que integran los principales índices bursátiles de Estados Unidos y Europa. El Apartado 4.2.2 explora las posibilidades de utilizar el análisis de co-enlaces, complementado con componentes cualitativos, para filtrar las consultas a los motores de búsqueda y poder investigar eventos concretos de índole económica. En concreto, la investigación pretende examinar si la información de enlaces en la red nos puede proporcionar información relevante sobre la evolución de la crisis financiera en los bancos estadounidenses.

Finalmente, el Apartado 4.3 incluye un trabajo que analiza el sector bancario internacional desde una perspectiva webmétrica que integra, tanto el análisis de impacto de enlaces como el análisis de co-enlaces.

Por último, el capítulo quinto incluye una visión global de la tesis previa a una exposición de las conclusiones, aportaciones realizadas, limitaciones y líneas futuras de investigación. Al margen, al final de la tesis, se adjunta la bibliografía y una sección con diversos anexos.

2. INTERNET Y LA WEB. LA BIBLIOTECA DE BABEL

«Ya se sabe: por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias.»

Jorge Luis Borges

La Biblioteca de Babel *(1944/2005a: 466).*

«Al principio se creyó que Tlön era un mero caos, una irresponsable licencia de la imaginación; ahora se sabe que es un cosmos y las íntimas leyes que lo rigen han sido formuladas, siquiera en modo provisional.»

Jorge Luis Borges

Tlön, Uqbar, Orbis Tertius *(1944/2005b: 435)*

Los hiperenlaces no han existido siempre.

Vivimos en una etapa de continuos desarrollos científicos y tecnológicos que han acabado por saturar nuestra capacidad de asombro. Antes de ser capaces de entender y asimilar un nuevo avance lo vemos superado por el siguiente. Al referirnos a las "nuevas tecnologías", concepto útil aunque de una vaguedad manifiesta, olvidamos frecuentemente que el libro que leemos, el lápiz con el que escribimos, el papel que empleamos sin mayor aprecio, fueron en su tiempo

desarrollos tecnológicos revolucionarios en buena medida. La postura intelectual de considerar que cada elemento con el que convivimos no ha existido siempre facilita un proceso de distanciamiento que nos devuelve una dimensión única e inesperada de lo que nos rodea. Nuestro tiempo fagocita rápidamente sus progresos más audaces, como Cronos devoraba a sus hijos.

Inmersos en este ritmo frenético, es probable que en los últimos 50 años no haya habido ninguna revolución tecnológica tan importante como la protagonizada por el desarrollo de la informática y por la creación de las redes de comunicaciones a escala global; especialmente me refiero a la invención y desarrollo de Internet y posteriormente de la World Wide Web. Esta tesis doctoral, y el camino investigador que inicia, parte de la perplejidad y la curiosidad ante estos avances que de manera tan rápida y radical han transformado cómo nos relacionamos, trabajamos, comerciamos, cómo somos en definitivas cuentas y hacia dónde vamos. El esfuerzo que ha movido la realización de esta investigación nace de la necesidad personal y social de entender este gran cerebro global que, para la Humanidad, representan Internet y la Web.

En mi opinión, hay un elemento que sustancia de manera inmejorable el porqué, el cómo y el destino de nuestro tiempo: el hiperenlace. Los hiperenlaces o, simplemente enlaces, que encontramos en la Web como sendas que continuamente se bifurcan, representan una llamada de atención sobre el otro, sobre lo otro, puertas a nuevo conocimiento o puntos de vista, una invitación al diálogo y a salir de uno mismo. Los hiperenlaces permiten la creación de hipertextos, que componen la World Wide Web, el hiperespacio. Todo esto no es, sin embargo, un producto demiúrgico o un fruto del azar, sino que es el resultado de diversas tradiciones, sueños y utopías que se enraízan siglos atrás, por lo que, de algún modo, suponen una realización cultural colectiva. En este capítulo pretendemos justificar la importancia social y científica de la Web, para lo cual convertimos las siguientes páginas en una forma de reconocimiento a grandes inventos, como el hiperenlace, y en una forma de homenaje a aquellas personas que han convertido este sueño en realidad, ahora que se cumplen los 20 primeros años de vida de la Web.

La World Wide Web se ha convertido en el laboratorio social de la Humanidad, un espacio en el que tienen reflejo todo tipo de actividades humanas que dejan su rastro, abriendo posibilidades casi ilimitadas de investigación. La inmensa amalgama de datos e información existente nos devuelve a la metáfora de la biblioteca borgiana. La Web se convierte en una gigantesca base de datos en la que es preciso identificar patrones, extraer conocimiento, visualizar la información. Se trata de una labor ardua y expuesta a dificultades y peligros. Entre ellos, destacan los problemas éticos a los que nos enfrentamos, la privacidad, los derechos de autor, la libertad de los ciudadanos, etc.

Baker (2009: 14) señala:

> "The exploding world of data, as we'll see, is a giant laboratory of human behavior. It's a test bed for the social sciences, for economic behavior and psychology. [...] These streams of digital data don't recognize ancient boundaries. They're defined by algorithms, not disciplines."

La minería de datos en Internet constituye un campo de investigación cada vez más potente. El propio nombre es descriptivo de la ardua labor que supone distinguir "los ecos de las voces". En adelante, nos serviremos de la metáfora de la Biblioteca de Babel borgiana (Borges, 1944/2005a: 466).

> "El universo (que otros llaman la Biblioteca) se compone de un número indefinido, y tal vez infinito, de galerías hexagonales, con vastos pozos de ventilación en el medio, cercados por barandas bajísimas. Desde cualquier hexágono se ven los pisos inferiores y superiores: interminablemente."

En uno de sus pasajes se narra la difícil tarea de la búsqueda del sentido, de lo inteligible, en la inmensidad de símbolos de la Biblioteca. Dice el relato, en relación con algunos de los libros encontrados en sus infinitos anaqueles (Borges, 1944/2005a: 465):

> "Uno, que mi padre vio en un hexágono del circuito quince noventa y cuatro, constaba de las letras M C V perversamente repetidas desde el renglón primero hasta el último. Otro (muy consultado en esta zona) es un mero laberinto de letras, pero la página penúltima dice Oh tiempo tus pirámides. Ya se sabe: por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias."

2.1. El origen de la Biblioteca de Babel o el sueño del cerebro global

Si bien la Biblioteca de Babel soñada por Jorge Luis Borges existe *ab aeterno*, la metafórica imagen de una biblioteca total, infinita en potencia, ha constituido para muchos una eficaz representación de las inabarcables regiones del ciberespacio, como un universo en continua expansión una vez que décadas atrás asistimos a su *big bang* creador.

La Biblioteca de Babel es un homenaje a la biblioteca mítica por antonomasia, la Biblioteca de Alejandría, monumento al conocimiento de la antigüedad que ha alimentado los sueños ilustrados de un saber recopilado y ordenado como herramienta del progreso de la Humanidad. La Biblioteca de Alejandría, en su tiempo, contenía gran parte de todo el conocimiento existente. Si bien el gran volumen de información que se genera en nuestros días hace inviable el mantener un registro unificado de toda ella, el sueño de organizar este conocimiento con el objetivo último de hacer un mundo mejor, libre de la barbarie y la guerra provocada por la ignorancia, ha estado en la base de buena parte de los antecedentes que expondremos en las siguientes páginas.

En la Edad Media, el conocimiento estaba restringido a estamentos privilegiados que eran los únicos capaces de leer y escribir. Por contra, el Renacimiento, con el revolucionario invento de la imprenta y, posteriormente, la Ilustración, con su sueño enciclopédico, hacen que el conocimiento, unido a un incipiente desarrollo científico, se multiplique y difunda entre capas sociales privilegiadas, pero cada vez

más amplias. Las bibliotecas viven un floreciente desarrollo y surgen las bibliografías universales, que buscan proporcionar acceso al conocimiento recopilado. Con todo, la Enciclopedia sigue siendo el producto de una élite cultural para una élite cultural, en una sociedad en la que la mayor parte de la población es analfabeta.

A lo largo del siglo XX con el creciente desarrollo tecnológico, la utopía del acceso universal a la información se manifiesta como un proyecto viable en la imaginación de diversos científicos y escritores. Al compás de los grandes enfrentamientos bélicos del siglo, se va generando la idea de que la recopilación y acceso al conocimiento por parte de todos los ciudadanos constituye la mejor vacuna contra nuevas guerras. Este pensamiento utópico se compagina con el objetivo instrumental de organizar el conocimiento científico de una manera más eficiente, al tiempo que con la carrera militar para generar nuevas tecnologías que permitieran alcanzar ventajas competitivas frente al enemigo. Todas estas líneas confluirán en el desarrollo de Internet a finales de los 60 y en la creación de la Web a finales de los 80. De alguna manera, ambas creaciones encierran en sí la semilla del sueño ilustrado; si bien, en nuestros días, la Web constituye una muestra destacada del caleidoscopio de la posmodernidad.

En las próximas páginas vamos a realizar un recorrido por aquellos discursos que vislumbraron o sirvieron de inspiración en el desarrollo de los hiperenlaces, el hipertexto, Internet y la Web. Para ello, empezaremos con dos autores, Paul Otlet y H.G. Wells, que, tras la Primera Guerra Mundial, plantean dos proyectos para organizar el conocimiento global, el *Mundaneum* y el *World Brain*, respectivamente (Torres-Vargas, 2005). Tras la Segunda Guerra Mundial encontramos un artículo de referencia, "As we may think", de Vannevar Bush (1945), donde plantea la organización del conocimiento a través de una máquina bautizada como *Memex*. Posteriormente, asistimos al desarrollo de las computadoras y a la invención del hipertexto y de Internet, en el contexto de la Guerra Fría y de los esfuerzos de los Estados Unidos por imponerse al bloque comunista. Aún habrían de pasar 20 años hasta que Tim Berners-Lee, creara la World Wide Web con el objetivo inicial de organizar el conocimiento que los científicos generaban en el CERN, sede de la

Organización Europea para la Investigación Nuclear.

> "De esas premisas incontrovertibles dedujo que la Biblioteca es total y que sus anaqueles registran todas las posibles combinaciones de los veintitantos símbolos ortográficos (número, aunque vastísimo, no infinito) o sea todo lo que es dable expresar: en todos los idiomas. Todo: la historia minuciosa del porvenir, las autobiografías de los arcángeles, el catálogo fiel de la Biblioteca, miles y miles de catálogos falsos, la demostración de la falacia de esos catálogos, la demostración de la falacia del catálogo verdadero, el evangelio gnóstico de Basilides, el comentario de ese evangelio, el comentario del comentario de ese evangelio, la relación verídica de tu muerte, la versión de cada libro a todas las lenguas, las interpolaciones de cada libro en todos los libros, el tratado que Beda pudo escribir (y no escribió) sobre la mitología de los sajones, los libros perdidos de Tácito."
>
> Jorge Luis Borges, *La Biblioteca de Babel* (1944/2005a: 467-468).

2.1.1. Paul Otlet y el Mundaneum

Paul Otlet (1868-1944) nació en Bruselas y es considerado como el padre de la Ciencia de la Documentación. Entre sus mayores logros destacan la creación del Instituto Internacional de Bibliografía, el Repertorio Bibliográfico Universal, la Clasificación Decimal Universal y la elaboración de diversas obras que sirvieron de base para el desarrollo de las ciencias documentales. Como autor inmerso en las corrientes científicas de su tiempo, defendía los principios del positivismo y la evolución, lo cual derivó en su concepción acerca del relativismo del conocimiento y de la formación histórica de los conceptos e ideas.

Al margen de las contribuciones referidas anteriormente, Otlet ideó el proyecto del *Mundaneum*, cuyo origen se remonta a 1910. Se trata de una idea que surge en el contexto de internacionalismo imperante tras la Primera Guerra Mundial. Un mundo organizado racionalmente debería contar también con un centro internacional

donde todo el conocimiento pudiera concentrarse y organizarse para el aprovechamiento y progreso de la Humanidad. El *Mundaneum* se concibe como un proyecto arquitectónico, sede de los recursos bibliográficos que se organizarían a través del Repertorio Bibliográfico Universal. Se situó en un primer momento en el *Palais du Cinquantenaire* en Bruselas, siendo finalmente trasladado en los años 30 a un inmueble de la ciudad belga de Mons. En 1929 Otlet encargó al arquitecto Le Corbusier su diseño, que habría de construirse en Ginebra (Suiza). Sin embargo, el proyecto nunca se llevó a cabo.

El *Mundaneum* se concibe como un instrumento para organizar y coordinar la información mundial, sirviendo de base para un orden internacional racional y en paz. La forma de alcanzar este objetivo es contando con una sociedad civil fuerte, con capacidad para la crítica y la opinión, bien informada, asistida por un sistema internacional de documentación. El Repertorio Bibliográfico Universal sería el instrumento encargado de registrar toda la información generada a lo largo del mundo. Consistiría en un inventario, clasificado por temas y autores de todos los libros y publicaciones de todos los países, de todos los periodos y referentes a todos los temas (Otlet, 1996: 405).

La organización del conocimiento en el proyecto de Otlet se entiende de una manera centralista. Su sistema de información mundial se materializaría en un lugar físico en el que se almacenarían todos los documentos que sirven de soporte al conocimiento humano. La importancia del componente arquitectónico en el proyecto se remonta al poder simbólico de la Biblioteca de Alejandría o de la propia torre de Babel, como monumentos a la diversidad humana. El *Mundaneum* constituiría el edificio principal de una Ciudad Internacional, que representaría un nuevo monumento al saber humano, una ciudad de los libros.

El utópico proyecto se basa en una idea de acceso universal al conocimiento; si bien, su materialización cuenta con limitaciones estructurales, en tanto que la acumulación de documentos en un único lugar impediría precisamente la consecución del objetivo del acceso universal y libre a la información.

2.1.2. H. G. Wells y el World Brain: The Idea of a Permanent World Encyclopaedia

Herbert George Wells (1866-1946) era un escritor inglés, graduado en Biología en 1890 en la London University. Inspirado en sus conocimientos científicos, imaginó algunas de las obras de ciencia ficción más populares de todos los tiempos, entre ellas, *The Time Machine* (1895), *The Invisible Man* (1897) y *The War of the Worlds* (1898). En 1937, Wells publicó también un artículo con el título "World Brain: The Idea of a Permanent World Encyclopaedia", en la Encyclopédie Française (incluido en el Anexo I). En 1938, el mismo aparecía también en un volumen titulado *World Brain*, donde se incluían otros ensayos vinculados a la gestión de la información y a la educación.

Wells, autor de ideas socialistas, se mostraba preocupado por la situación de la humanidad que, generación tras generación, afrontaba problemas similares sin ser capaz de aprender y beneficiarse de las experiencias pasadas. En su opinión, la falta de organización del conocimiento humano constituía la causa fundamental de esta situación. Su idea de *World Brain* consiste en la creación de una enciclopedia mundial que solventara estos problemas. Wells (1937) lo describe del siguiente modo:

> "This in itself is a fact of tremendous significance. It foreshadows a real intellectual unification of our race. The whole human memory can be, and probably in a short time will be, made accessible to every individual. And what is also of very great importance in this uncertain world where destruction becomes continually more frequent and unpredictable, is this, that photography affords now every facility for multiplying duplicates of this - which we may call? - this new all-human cerebrum."

El proyecto de enciclopedia mundial abordaría todos los campos del conocimiento, proporcionando datos de todo tipo, al tiempo que serviría para el desarrollo de reglas sociales de convivencia. Incluiría también un sistema de referencia completo de fuentes de información y una historia general del mundo. La enciclopedia estaría

compuesta por un conjunto de textos, extractos y notas que, seleccionados por expertos en cada campo del conocimiento, ilustrarían los distintos temas de forma crítica y convenientemente estructurada. Se trataría de un modelo que recuerda de alguna manera al Diccionario de autoridades, primer diccionario confeccionado por la Real Academia Española, en el que los términos se ilustraban con citas de autores que ejemplifican o corroboran la definición dada. La enciclopedia mundial estaría en un continuo proceso de revisión, elaboración y modificación por expertos reconocidos.

Entre las cuestiones que suscitaba el proyecto se planteaba, por ejemplo, quién podría ser el editor de semejante obra, en qué lenguaje se llevaría a cabo y cómo se realizaría la publicación. Wells se mostraba reacio a someter el proyecto a los intereses financieros de una editorial comercial. Con todo, quizá la característica más importante del proyecto es su concepción descentralizada, la cual lo diferencia claramente del *Mundaneum* de Otlet. Si bien la obra tendría una estructura y organización centralizada, debería poder ser distribuida físicamente de manera que fuera realmente accesible para todo el mundo. La enciclopedia debería poder duplicarse, copiarse fácilmente. Ello permitiría crear un verdadero "cerebro mundial" (*world brain*).

Mientras que el *Mundaneum* de Otlet se basa en procedimientos tradicionales de almacenamiento de documentos, Wells plantea el empleo de la tecnología, el *microfilm* concretamente, para alcanzar el objetivo de divulgar el conocimiento. Señala Wells (1937):

> "There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind. And not simply an index; the direct reproduction of the thing itself can be summoned to any properly prepared spot."

2.1.3. Vannevar Bush y el Memex

Tras la segunda Guerra Mundial, Vannevar Bush (1890-1974) publica en el número de julio de 1945 de *The Atlantic Monthly* el artículo "As we may think", en el que plantea un sistema de hipertexto que permitiría almacenar documentos y vincularlos. Bush idea para ello una máquina a la que denomina *Memex*, que proviene de la abreviación de "memory extender".

Bush era ingeniero y profesor en el Massachusetts Institute of Technology (MIT). Su labor más destacada fue la coordinación, como director de U.S. Office of Scientific Research and Development, de las actividades científicas de los más de 6.000 científicos que trabajaron de forma conjunta durante la segunda Guerra Mundial para los Estados Unidos (entre otros, destacan los siguientes trabajos e invenciones: el desciframiento de códigos secretos, la bomba atómica o el radar). El artículo "As we may think" surge como fruto de la experiencia de coordinar las contribuciones de cientos de científicos y de la constatación de la imposibilidad de gestionar todo el conocimiento existente con los medios bibliográficos empleados hasta el momento. Bush (1945) constata el gran volumen de conocimiento disponible:

> "There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends.
>
> [...] The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships."

No se trata únicamente de una cuestión tecnológica sino de los principios empleados en la organización del conocimiento, una organización basada en jerarquías en la que todo documento ha de clasificarse de acuerdo con unas determinadas categorías. Esto imposibilitaba el acceso efectivo a las producciones científicas, dado que nuestro funcionamiento cerebral, nuestro pensamiento,

funciona de acuerdo con otros principios, de forma asociativa. En palabras de Bush (1945):

> "Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path. [...] The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature."

La visión de Bush se materializa en el diseño de una máquina, que combina las últimas tecnologías existentes en la fecha, fundamentalmente el *microfilm*, igual que planteaba Wells. Se trata del *Memex*, cuya descripción se detalla en el artículo y aquí se ilustra con algunas citas del mismo:

> "Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.
>
> [...] This is the essential feature of the memex. The process of tying two items together is the important thing. When the user is building a trail, he names it, inserts the name in his code book, and taps it out on his keyboard. Before him are the two items to be joined, projected onto adjacent viewing positions.
>
> [...] Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and

there amplified."

Los distintos documentos y párrafos de los textos podrían vincularse mediante el establecimiento de "caminos" (*trails*) que funcionan de manera análoga a los hiperenlaces en la Web. Los documentos, junto con las conexiones establecidas por los usuarios, podrían pasarse a otros usuarios, que podrían acceder a ellos a través de su *Memex* personal. De hecho, Bush plantea la posibilidad de que se creen contenidos microfilmados con este fin, por ejemplo, enciclopedias.

El *Memex* nunca llegó a construirse en la realidad. Sin embargo, la idea ha servido de inspiración y referencia para futuros desarrollos, tales como los ordenadores personales, el hipertexto, Internet o la World Wide Web (Simpson et al., 1996). En ocasiones la influencia no ha sido directa. No obstante, tarde o temprano, los creadores han reconocido la importancia de las visiones de Bush para anticipar un sistema de información basado en el hipertexto.

Sus ideas sobre el pensamiento asociativo son fruto de las corrientes de pensamiento predominantes en su tiempo (Houston y Harmon, 2007). Sin embargo, la idea del descubrimiento por asociación remite a conceptos muy de actualidad cuando nos referimos a la Web, la "serendipia" (*serendipity*, en inglés) o, un concepto que veremos con posterioridad, el de los enlaces transversales.

2.1.4. El hipertexto y otros proyectos

El término "hipertexto" fue creado por Ted Nelson, quien reconociendo la influencia del *Memex* propuesto por Bush (Naughton, 1999; Gillies y Cailliau, 2000), propuso un sistema de información global conocido como *Xanadu* (Nelson, 2000). Nelson (1987: 0/2, citado por Houston y Harmon, 2007) define, en su obra *Literary Machines*, el hipertexto como "non-sequential writing with reader-controlled links".

Otro antecedente destacado es Douglas Engelbart, quien diseñó en 1968 su *On-Line System*, trabajando en el Augmentation Research Center, situado en el área de

la Bahía de San Francisco. Engelbart leyó el artículo de Bush en 1945 y lo cita en un trabajo 16 años después (Engelbart, 1961), reconociendo que probablemente la lectura temprana de este texto le influyó en su trabajo posterior, aunque fuera de una manera subliminal.

Por último, cabe mencionar a Bill Atkinson, quien desarrolló en los 80 un sistema llamado *HyperCard* para interconectar información.

2.1.5. Internet

Para Dodge y Kitchin (2001), la historia reciente de las tecnologías de la información y comunicación alcanza un punto de inflexión en el lanzamiento del Sputnik 1 en 1957 y en el alunizaje de Luna 2 en 1958. Con el objeto de no perder la carrera espacial y superar los logros soviéticos, el departamento de defensa de los Estados Unidos creo ARPA (*Advanced Research Projects Agency*). La Agencia requirió la ayuda de visionarios y tecnólogos que les permitieran dibujar el camino a seguir, por ejemplo, en el campo de la informática. Para Castells (2001), Internet nace como resultado de una combinación de tres elementos de diversa naturaleza: el primero, lo que se ha dado en llamar como "gran ciencia" (*big science*); el segundo, la investigación militar; y, por último, la cultura libertaria. Estos elementos se encuentran en su origen, pero también proporcionan algunas claves para entender su desarrollo posterior. En su origen, el proyecto de lo que finalmente dará lugar a Internet se diseña como un sistema para conectar distintos sistemas incompatibles entre sí y localizados en varios lugares de los Estados Unidos.

J.C.R. Licklider (1915-1990) se hizo cargo en 1962 de la Information Processing Techniques Office (IPTO) dentro de ARPA. Un año más tarde lanzaba el proyecto ARPANET, red que permitiría conectar distintos ordenadores en tiempo real empleando una red distribuida. En 1968, Licklider escribió el artículo *The Computer as a Communication Device* en el que anticipaba el empleo de redes de ordenadores para facilitar, con independencia de la localización geográfica, el

establecimiento y colaboración de comunidades con intereses compartidos. El primero de los nodos de ARPANET se estableció en la University of Califonia, Los Angeles (UCLA), a finales de 1969, y el segundo en el Stanford Research Institute, creándose la primera conexión. Antes de final de 1969, se sumaron dos nodos más a la red, la University of California, Santa Barbara, y la University of Utah. En marzo de 1977 ya existían 111 nodos en toda la red. En 1981, había 213. Son años en los que se llevan a cabo importantes avances, como, por ejemplo, el diseño del protocolo de red TCP/IP en 1973 por Vinton G. Cerf y Robert E. Kahn.

La red creada se empleaba fundamentalmente para el servicio de correo electrónico, desarrollado por Ray Tomlinson en 1970, y para el intercambio de paquetes de datos. Entre los hitos que jalonan el desarrollo de Internet, a lo largo de los años 70 y 80, destaca la importancia de los *Bulletin Boards*, que permitían un intercambio de archivos más fluido al tiempo que abrían la puerta para un empleo de Internet no necesariamente vinculado al ámbito académico o a usuarios con elevados conocimientos técnicos.

Un tema recurrente a la hora de abordar los orígenes de Internet es su vinculación con fines militares. La situación de Guerra Fría existente en los años 60 contribuyó a que grandes cantidades de dinero se invirtieran en el desarrollo de nuevas tecnologías con objetivos militares a través de ARPA. En este contexto es donde surge la idea de crear una red distribuida, no centralizada, formada por nodos interconectados, de modo que la destrucción de uno o varios de ellos no causara el colapso completo de la red. De este modo, se podrían garantizar las comunicaciones en un escenario de ataque, probablemente, de tipo nuclear. Es aquí donde entra en escena Licklider y el proyecto ARPANET. Castells (2001) apunta, sin embargo, que el desarrollo de ARPANET no se produjo con fines militares sino principalmente científicos, por parte de aquellos investigadores que trabajaban en ARPA y en el entorno del grupo. Castells considera que el desarrollo de Arpanet no se produce de manera casual como un producto secundario de otros proyectos principales, sino que, desde un principio, fue diseñado de manera consciente por un grupo de científicos especialistas en informática, en su mayoría procedentes del Massachusetts Institute of Technology. En palabras de Castells

(2001: 19), el desarrollo de la futura Internet consistía en: "a scientific dream to change the world through computer communication". Podemos concluir que si bien ARPANET no acabó desarrollándose como un proyecto orientado a fines militares, sin los abundantes recursos disponibles para estos fines no se hubiera podido llevar a cabo. A la larga, es probable que la superioridad tecnológica de los Estados Unidos en este campo acabara influyendo en la Perestroika de Gorbachov y en la caída de la Unión Soviética (Castells y Kiselyova, 1995).

2.1.6. Web

Pese a que la creación de Internet se remonta a finales de los sesenta, su universalización y popularización no se produce hasta el desarrollo de la World Wide Web (WWW). La Web (en adelante) surge como un proyecto del programador inglés Tim Berners-Lee, que trabajaba en el CERN, el centro de investigación europeo de física de partículas con sede en Ginebra.

El sistema de hipertexto World Wide Web, WWW o Web fue lanzado en 1991 por Tim Berners-Lee como una "intranet" para los investigadores del CERN con el objetivo de facilitar el que se pudieran publicar y compartir datos y resultados de investigación. Esto se realizaba a través de un sistema en línea que permitiría el acceso a la información a partir de unos protocolos compartidos (Berners-Lee, 1989/1990; Berners-Lee y Cailliau, 1990). La Web era el resultado de la combinación de dos tecnologías originadas en los años 60; por un lado, Internet y por otro el hipertexto, a partir de las aportaciones de Ted Nelson (1967). La Web está basada en una serie de protocolos y elementos (HTTP, HTML y URI, luego URL) que permitieron que los distintos ordenadores, con independencia de los sistemas operativos y el *software* que emplearan, pudieran interactuar en Internet, interconectarse. Berners-Lee (1999: 26) describe su desarrollo:

> "Entonces pude escribir rápidamente el código para el Protocolo de Transferencia de Hipertexto [Hypertext Transfer Protocol] (HTTP), el lenguaje que los ordenadores usarían para comunicarse por Internet, y el Identificador de Recursos Universal

> [Universal Resource Identifier] (URI), el esquema para direcciones de documentos.
>
> Hacia mediados de noviembre tuve un programa cliente -un navegador/editor de señalar y cliquear-, que llamé World Wide Web. En diciembre estaba funcionando con el Lenguaje Markup de Hipertexto [Hypertext Markup Language] (HTML) que había escrito, que describe como formatear páginas que contengan vínculos de hipertexto. El navegador decodificaría los URIs y me permitiría leer, escribir o editar páginas web en HTML. Podría navegar por el Web usando HTTP, aunque podría guardar documentos sólo en el sistema de ordenador local, no en Internet."

La Web funciona con poco más que HTTP, HTML y los URIs (o URLs). Esta sencillez hace que entonces y aún en la actualidad su funcionamiento sea algo difícil de entender. Muchas personas siguen preguntándose dónde está la Web, en qué sistemas se encuentra alojada.

> "Lo que costaba entender a la gente acerca del diseño era que no había nada más allá de los URIs, HTTP y HTML. No había un ordenador central "controlando" el Web, ni un Web único en el que funcionasen esos protocolos, ni siquiera una organización en ninguna parte que "manejase" el Web. El Web no era una "cosa" física que existiese en determinado "lugar". Era un "espacio" en el que la información podía existir." (Berners-Lee, 1999: 34)

Berners-Lee (1999: 34) explica el funcionamiento de la Web a través de un símil con un sistema de economía de mercado.

> "Yo le decía a la gente que el Web era como una economía de mercado. En una economía de mercado, cualquier puede negociar con cualquiera, y no tienen que ir a una plaza de mercado para hacerlo. Lo que necesitan, sin embargo, son unas cuantas reglas con las que todo el mundo tiene que estar de acuerdo, como la moneda que se vaya a utilizar para el negocio y las reglas del comercio justo. Las reglas equivalentes al comercio justo en el Web son las reglas acerca de lo que significa un URI como dirección, y el lenguaje que usan los ordenadores -HTTP- cuyas reglas definen cosas, como quién habla primero, y según qué turno hablan.

> Cuando dos ordenadores se ponen de acuerdo, pueden hablar, y luego tienen que encontrar una manera común de representar sus datos para poder compartirlos. Si usan el mismo software para documentos o gráficos, pueden compartirlos directamente. Si no, pueden traducirlos ambos a HTML."

Esta tecnología estuvo disponible para el resto del mundo en 1993 (Cailliau, 1995). Uno de las hitos en la popularización de la Web fue el desarrollo de los navegadores. En 1995 por fin Microsoft reconoció, con considerable retraso y falta de visión de futuro, la importancia de esta nueva red e incluyó con su Windows 95 el navegador Explorer. La difusión de la Web se hace ya un fenómeno imparable. En apenas 10 años, un proyecto para el intercambio de información científica se convirtió en la mayor red de información jamás conocida por el ser humano (Lyman y Varian, 2000; Weare y Lin, 2000).

Berners-Lee (1999) reconoce las contribuciones visionarias de Bush, Nelson y Engelbart, aunque no las conociera cuando comenzó a desarrollar la idea de la World Wide Web. En realidad, Berners-Lee se incardina más bien en una tradición, no tanto por su conocimiento y adscripción expresa a ella, sino más bien por la toma de conciencia de una serie de necesidades comunes a muchos de los anteriores proyectos, la falta de comunicación entre los científicos en un mundo cada vez más complejo y con mayor volumen de producción científica y de conocimientos. Él observa este problema en el CERN, como Vannevar Bush ya lo hiciera tras coordinar los esfuerzos de los científicos en Estados Unidos durante la Segunda Guerra Mundial.

Su libro *Tejiendo la Red*, publicado en 1999, arroja muchas de las claves de la Web, desde su naturaleza como red en la que prima una perspectiva relacional del conocimiento, hasta los objetivos y las perspectivas de futuro de la misma, pasando por consideraciones de carácter más técnico. En las siguientes líneas, su pensamiento entronca con la visión asociativa del conocimiento que defendía Vannevar Bush (Berners-Lee, 1999: 1):

> "La visión del Web que tuve fue la de cualquier cosa potencialmente conectada a cualquier cosa. Es una visión que nos proporciona una nueva libertad y nos permite crecer más rápidamente de lo que nunca pudimos crecer cuando estábamos encadenados por los sistemas de clasificación jerárquica a los que nos aferramos. Deja la totalidad de modos de trabajar anteriores como sólo una herramienta entre muchas. Deja los miedos que teníamos al futuro convertidos en uno entre muchos. Y acerca más los funcionamientos de la sociedad a los funcionamientos de nuestra mente."

Berners-Lee concibe el mundo como un sistema en el que el significado de cada elemento sólo puede inferirse a partir de sus relaciones con los demás elementos. Esta es la concepción que, de una manera radical, traslada al diseño de la Web, estableciendo también una analogía con el cerebro humano (Berners-Lee, 1999: 12):

> "Desde un punto de vista extremo, el mundo puede considerarse como sólo un conjunto de conexiones, nada más. Consideramos un diccionario como un depósito de significados, pero define palabras sólo en términos de otras palabras. Me gustaba la idea de que un fragmento de información fuera realmente definido sólo por aquello con lo que está relacionado. Realmente hay muy poco más en el significado. La estructura lo es todo."

El nacimiento de la Web está vinculado a la filosofía del *software* libre. Berners-Lee no patentó en ningún momento su invento permitiendo que se desarrollara y contribuyendo de manera decisiva al sueño de una humanidad interconectada. Su trabajo en pos del interés público se tradujo en la posterior creación del World Wide Web Consortium (W3C).

Para Dodge y Kitchin (2001: 3), la Web consiste en "multimedia data (mostly text and static graphics but also sound, animation, movie clips and virtual spaces) which are stored as hypermedia documents (documents that contain links to other pages of information)". El crecimiento de la Web nunca ha parado hasta el punto de que

desconocemos cuál puede ser su dimensión (Chakrabati et al., 1999).

2.1.7. Web 2.0

La Web 2.0 (O´Reilly, 2005) constituye la última estación en la que nos vamos a detener en este breve recorrido por los orígenes de Internet y la Web. Con especial intensidad, en los últimos cinco años se han desarrollado una serie de servicios en la Web con un componente fundamentalmente social que han permitido una participación masiva de los ciudadanos a través de blogs, redes sociales, wikis, etc. Fundamentalmente, la Web 2.0 ha permitido que la Web se parezca más a lo que originalmente pretendía Berners-Lee, una Web en la que fácilmente se pueda acceder y crear contenido por parte de cualquier usuario. Las aplicaciones que se identifican con la Web 2.0 (blogs, redes sociales, wikis, páginas para compartir vídeos, fotografías o enlaces, etc.) han sabido aprovechar los recursos disponibles, por ejemplo, un mayor ancho de banda o la aplicación de avances informáticos, para enriquecer la experiencia del usuario (Jenkins, 2006). El ciudadano en red ha abandonado en muchos casos el simple rol de consumidor de información para convertirse también en un creador de contenidos.

2.1.7.1. Origen del término

La creación del término Web 2.0 se atribuye a Dale Dougherty de O´Reilly Media. Su objetivo era transmitir la idea de una Web en una nueva fase de desarrollo, encaminada a un nuevo auge tras el fracaso de los modelos de negocio *puntocom* surgidos pocos años atrás. La invención del término responde a motivos comerciales, un buen nombre con el que publicitar las conferencias que los organizadores realizaban anualmente sobre las últimas tendencias en Internet y la Web. De este modo, O'Reilly Media, Battelle y MediaLive lanzaron el primer encuentro sobre la Web 2.0 en octubre del 2004.

Tim O´Reilly (2005) publicó a continuación el artículo "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software". Se trata de un texto seminal en el que desarrollaba cuáles eran las características e implicaciones técnicas y sociales de la Web 2.0: modelos de negocio, factores asociados a lo que se conoce como *software* social (la participación, el poder de los grupos, etc.), especificaciones técnicas. Sin embargo, no se trataba de un cambio tecnológico, sino fundamentalmente de carácter social, ya que la mayoría de tecnologías empleadas ya existían desde años atrás. No es correcto hablar, por tanto, de ruptura frente a un modelo anterior sino de una evolución.

2.1.7.2. El concepto

O´Reilly, en su artículo, intenta establecer una serie de principios que permitan identificar a las empresas y servicios 2.0. Para O´Reilly (2005) los principios sobre los que descansa el diseño de la Web 2.0 son: el aprovechamiento de la "larga cola", la utilización de los datos como ventaja competitiva principal, la obtención de valor a través de los usuarios, el uso automático de externalidades de red, el empleo de la inteligencia colectiva, los cambios en los sistemas de protección de la propiedad intelectual, el desarrollo de servicios web en continuo proceso de evolución (en estado de *beta* continuo) y la primacía de dinámicas de cooperación frente al control.

En cualquier caso, existen voces que cuestionan el concepto Web 2.0, criticando su falta de claridad y definición. Tim Berners-Lee (citado por Anderson, 2007: 5) es uno de los críticos:

> "Web 1.0 was all about connecting people. It was an interactive space, and I think Web 2.0 is of course a piece of jargon, nobody even knows what it means. If Web 2.0 for you is blogs and wikis, then that is people to people. But that was what the Web was supposed to be all along. And in fact, you know, this 'Web 2.0', it means using the standards which have been produced by all these people working on Web 1.0."

Se trata en todo caso de una cuestión en esencia nominativa, ya que los objetivos que abandera la Web 2.0, en realidad ya se encuentran en el propio concepto originario de la Web. Así lo expresa Tim Berners-Lee (1999: 115) refiriéndose a su origen:

> "El Web es más una creación social que técnica. Yo lo diseñé por su efecto social -para ayudar a que la gente trabajase junta- y no como un juguete técnico. El objetivo último de la Web es apoyar y mejorar nuestra entretejida existencia en el mundo. Nos agrupamos en familias, asociaciones y empresas. Tenemos confianza en cosas que están a kilómetros y no la tenemos en cosas que están a la vuelta de la esquina. Lo que creemos, aprobamos, aceptamos y de lo que dependemos es representable y, cada vez más, está representado en el Web. Tenemos que asegurar que la sociedad que construimos con el Web es la que pretendemos construir."

Los términos que se emplean para describir este fenómeno: colaboración, participación, distribución, efectos de red, etc., se parecen mucho a los empleados por Berners-Lee (1999: 149) al señalar que:

> "Cuando presenté el Web en 1989, la fuerza motora que tenía en mente era la comunicación por medio del conocimiento compartido [...]. Al construir un Web de hipertexto, un grupo de personas de cualquier tamaño podría expresarse fácilmente, adquirir y transmitir rápidamente conocimientos, superar los malentendidos y reducir la duplicación de esfuerzo."

Dada su popularidad y la constatación de las importantes transformaciones que ha generado en el empleo social de la Web, consideramos útil el empleo del concepto "Web 2.0". Al margen de las características ya apuntadas, Ribes (2007) aporta esta definición del término:

> "[T]odas aquellas utilidades y servicios de Internet que se sustentan en una base de

> datos, la cual puede ser modificada por los usuarios del servicio, ya sea en su contenido (añadiendo, cambiando o borrando información o asociando metadatos a la información existente), bien en la forma de presentarlos, o en contenido y forma simultáneamente."

Son muchos autores los que coinciden en considerar la Web 2.0 principalmente como una actitud y no como un cambio tecnológico (Davis, 2005; Dans, 2007). Ello es precisamente lo que hace que su repercusión sea tan significativa. Para Fumero y Roca (2007: 10) la Web 2.0 es la "promesa de una visión realizada: la Red –la Internet, con mayúscula o minúscula, que se confunde popularmente con la propia Web– convertida en un espacio social, con cabida para todos los agentes sociales, capaz de dar soporte a y formar parte de una verdadera sociedad de la información, la comunicación y/o el conocimiento". Esta visión es la que entronca de lleno con el relato que venimos contando.

2.1.7.3. Las características

Como señala O´Reilly (2005), la Web 2.0 parte de la concepción de la Web como una plataforma en la que los productos se convierten en servicios *online* que son ubicuos al permitir el acceso a los mismos desde cualquier conexión a Internet, así como "portables" en el sentido de que son accesibles desde distintos tipos de dispositivos. En esta línea, presenta el ejemplo de dos empresas que apostaron por modelos de negocio muy distintos: Netscape y Google. Netscape fue la primera empresa de Internet en el mundo que cotizó en bolsa. Su apuesta fue por una idea de "web como plataforma", promoviendo para ello un "webtop" que sustituyera al "desktop" habitual de los ordenadores personales. Sin embargo, el valor añadido acabó desplazándose de los navegadores web, que se convirtieron en una *commodity* más, a los servicios ofrecidos en la propia Web. Google, por su parte, inició su andadura como una aplicación web nativa, que nunca se ha ofrecido como producto, sino como un servicio. Frente a las actualizaciones, hay una mejora

continua del servicio; frente a las licencias o ventas, simplemente uso. Google realmente es una empresa basada en la gestión de una inmensa base de datos, no sólo ni principalmente una colección de herramientas de *software* (Batelle, 2006).

Entre algunas de las claves del éxito de esta filosofía podemos enumerar las siguientes:

- La generación de una base de datos y su explotación se convierte en la principal ventaja competitiva de la Web 2.0. Las aplicaciones 2.0 explotan los "efectos de red" consiguiendo mayores beneficios para los usuarios conforme mayor número de gente las utiliza. Los datos son la ventaja competitiva clave. La lucha por el control de los mismos entre los propios usuarios y las empresas es una cuestión aún por resolver.

- La facilidad para reutilizar piezas de código de programación creadas por otros y la predisposición a compartir funcionalidades e ideas en aplicaciones de código libre o abierto incentiva el desarrollo de nuevas aplicaciones.

- Fruto de la unión del poder de las bases de datos y de una concepción del *software* más libre aparecen los *mashups*. Los *mashups* son servicios que surgen a partir de la combinación de aplicaciones gracias al empleo de bases de datos de terceros que son accesibles a través de interfaces de programación (conocidas como API, tal y como se verá posteriormente). Esta forma de proceder evita redundancias en la red, al no tener que duplicar algo que ya existe, por lo que los recursos se pueden emplear en mejorar lo ya existente. Todo el mundo se beneficia de ello, ya que las aplicaciones que sirven sus bases de datos a otras aplicaciones se enriquecen mediante su empleo.

- El *software* se entiende como un servicio, no como un producto. Las actualizaciones y mejoras de los servicios en la Web ocurre de espaldas al usuario que no ha de preocuparse por su mantenimiento. La necesidad de una continua actualización ha hecho que los lenguajes de programación dinámicos desempeñen un papel fundamental, como es el caso de los

lenguajes de *scripting,* como Perl, Python, PHP o Ruby.

- El usuario se considera como un co-desarrollador del servicio pudiendo introducir mejoras en el código, enviar sugerencias, quejas, notificar fallos o simplemente utilizándolo, permitiendo a los desarrolladores aprender de sus forma de uso.

- El *software* no está limitado a un solo dispositivo. La Web se hace ubicua.

- Las nuevas herramientas en la red permiten aprovechar la "inteligencia colectiva" al agregar y explotar la información de miles o millones de usuarios que realizan sus aportaciones individuales. No se trata necesariamente de proyectos de colaboración, sino de aprovechar los esfuerzos individuales para evitar redundancias, filtrar contenidos relevantes, extraer nuevo conocimiento y mejorar el existente, etc.

- Uno de los efectos fundamentales de la Web 2.0 es el fenómeno de la "larga cola" (*long tail*, en inglés) (Anderson, 2006) que refleja el modo en que se distribuye la atención entre los sitios web, fundamentalmente en el ámbito del comercio. Un pequeño número de sitios recibe una gran atención mientras que la mayoría de los sitios recibe muy poca. La superación de las limitaciones de los modelos económicos basados en una presencia física de los productos que se pretenden vender ha hecho que en Internet la larga cola adquiera más y más importancia, ya que es posible atender a nichos de mercado muy específicos sin incrementar los costes y por tanto obteniendo beneficios donde antes resultaba imposible conseguirlo.

El futuro de la Web se sigue construyendo día a día. Actualmente ya se habla de Web 3.0 ó de Web semántica. A la Web semántica nos referiremos en páginas posteriores, pues ya se encontraba entre los proyectos de Berners-Lee (1999). La historia aún está por escribir.

2.2. Internet y la Web: no es lo mismo

Internet y la Web no son lo mismo, aunque frecuentemente se empleen de forma equivalente. Internet es un conjunto descentralizado de redes de comunicación distribuidas que se interconectan entre sí empleando la familia de protocolos TCP/IP. Su origen se remonta a la red ARPANET en Estados Unidos a finales de los años 60 y principios de los 70. Por su parte, la Web (World Wide Web o WWW) es una red de documentos que funciona en Internet, basada en un conjunto de protocolos, como es el HTTP. Su origen se sitúa a finales de los años 80 y principios de los 90, a partir de los trabajos llevados a cabo por Tim Berners-Lee en el CERN. El crecimiento y popularización de la Web ha sido tan importante a lo largo de los últimos 20 años que los términos Internet y Web se confunden de forma habitual. Actualmente, muchas de las formas de comunicación en Internet, como son el correo electrónico (basado en el protocolo SMTP), la transmisión de archivos (FTP y P2P), las conversaciones en línea (IRC), la mensajería instantánea, etc., son accesibles desde una interfaz web, lo que contribuye a un mayor solapamiento entre ambos conceptos para los usuarios. Holmberg (2009: 1) sintetiza la diferente naturaleza de ambos elementos en una frase: "The Internet is a global network of computers connected by cables and the World Wide Web is a global network of content partly connected together by hyperlinks".

Pese a tratarse de dos elementos distintos, la confusión entre Internet y la Web se puede apreciar en textos de referencia tan reconocidos como el *Oxford Dictionary of Sociology* (Scott y Marshall, 2009), cuya entrada "WWW" remite al término "Internet". La definición de este último concepto es la siguiente (p. 368): "A global network of computers (also known as the World-Wide Web) which allows instantaneous access to an expanding number of individual Web sites offering information about practically anything and everything -including the contents of daily newspapers, the price of goods in local shopping malls, library holdings, commodity prices,sport news and gossip, eroticism, and so-called chat-rooms (by means of which people can communicate with each other on-line about their interests, hobbies, and opinions)". La definición continúa señalando, por ejemplo, que Internet se puede considerar el desarrollo tecnológico más importante del siglo XX,

comparable a la invención de la imprenta o, incluso, a la electricidad.

Si bien en la mayoría de los casos, en su uso habitual, la asimilación de ambos conceptos no genera ningún tipo de problemas, en el ámbito académico es obligado precisar correctamente su significado. En este trabajo nos centramos fundamentalmente en el estudio de la Web.

Para concluir, cabe apuntar que, si bien Internet tiene su origen en la costa oeste de los Estados Unidos, al igual que el movimiento del Software Libre o las grandes multinacionales que hoy en día dominan la red, tanto la World Wide Web, como el desarrollo del sistema operativo Linux, pueden considerarse dos de las grandes aportaciones europeas a este proceso vivido a lo largo de los últimos 40 años.

2.3. El hiperenlace y la sociedad hiperenlazada

En este apartado se aborda un elemento clave en la existencia de la Web: el hiperenlace, que permite establecer conexiones automáticas entre los distintos elementos que componen la red. Los hiperenlaces conforman la estructura de los textos, dando lugar al hipertexto. El conjunto de hipertextos conforman la Web. A este espacio de lo virtual se le conoce también como el ciberespacio. Más allá del propio ámbito de lo virtual, nos encontramos en un tiempo en el que podríamos hablar de una sociedad hiperenlazada (Turow y Tsui, 2008).

2.3.1. Hiperenlace

Los hiperenlaces son un elemento básico para comprender el funcionamiento de la Web y las posibilidades de investigación que ofrece; en concreto, en relación con la aplicación de técnicas webmétricas. Un hiperenlace es básicamente una referencia que una página web establece a una sección de esa misma página o a otra página web completamente distinta. Los enlaces conforman la estructura de la Web

(Berners-Lee, 1999), ya que sin ellos resultaría imposible acceder a los contenidos alojados en páginas, a no ser que conociéramos la dirección donde se aloja el recurso (su URL). Los enlaces constituyen la estructura que vincula los documentos existentes en la Web y permite proporcionar información útil para obtener conocimiento. Una aplicación conocida del uso de los hiperenlaces es el sistema *PageRank* de Google (Brin y Page, 1998), ya mencionado y que abordaremos con más detenimiento en el Apartado 3.3.3.4. Sobre el hecho de enlazar, Tim Berners-Lee (1999: 115) señala que:

> "Los respaldos o avales como modo de transmitir juicios de calidad funcionan fácilmente en el Web, porque pueden hacerse por medio de vínculos de hipertexto. Sin embargo, por muy importante que sea esta facilidad, es incluso más importante entender que un vínculo no tiene por qué suponer un respaldo. El discurso libre en hipertexto supone el "derecho a vincular", que es la unidad básica constructiva de todo el Web."

La investigación webmétrica se basa en un estudio cuantitativo de la Web, siendo los hiperenlaces una de sus principales fuentes de información.

Dodge y Kitchin (2001) subrayan que la Web proporciona un medio potente para explorar documentos vinculados entre sí, permitiéndoles navegar, saltar, moverse entre documentos relacionados con independencia del lugar en que se encuentren o la distancia que los separen. Es más, el propio hecho de hablar de distancias podría carecer de sentido, si bien ya veremos en el Capítulo 3 la persistencia de patrones geográficos en la Web. Los hiperenlaces tienen el potencial de sortear barreras culturales, lingüísticas, económicas y políticas, entre otras, si bien la investigación demuestra que esto no siempre es así en la práctica.

2.3.2. Hipertexto

Nentwich (2003) describe algunas características del "hipertexto". Se trata de un texto no secuencial o no lineal, frente a los textos tradicionales, que se pueden describir como lineales. La no linealidad se refiere al hecho de que el hipertexto no tiene porqué estar conectado en una única forma, siguiendo una única senda, sino que los hiperenlaces permiten construir diferentes textos, dependiendo de las conexiones elegidas. Tredinnick (2007: 181) insiste en la idea de no linealidad al definir el hipertexto como "non-linear text with inter-textual linking through embedded cross-references".

El hipertexto también se puede definir como una matriz de textos posibles (Lévy, 2008). Mientras el texto tradicional es lineal, el hipertexto está construido por redes, permitiendo al usuario, lector o navegante, elegir su propia senda y de ese modo construir su propio texto. Abre la posibilidad de utilizar no solamente el texto, sino también otros formatos multimedia. Como apunta Montoya Juárez (2008: 278), el hipertexto llevado a su máxima expresión en Internet, "plantea una metamorfosis o un proceso continuo de construcción y transformación del discurso en que los actores negocian, la integración multimedia, una organización fractal de las conexiones, que pueden producirse unas bajo otras iterativamente, y un permanente descentramiento, puesto que la red no tiene un centro fijo y su mapa se modifica de manera permanente y cualquier nodo o conexión puede ser un centro provisorio funcionando siempre en red con otros (Landow, 1992; Lévy, 2008)".

De acuerdo con Nentwich (2003: 266), los enlaces son "text-markers – i.e. highlighted, underlined or otherwise marked text or small pictures such as "icons" or "buttons" – which the reader /user can click on with the PC mouse (in case s/he uses a graphical user interface) or by selecting it with the keyboard (in a pure text environment)." La idea de hipertexto encuentra sus raíces en varias tradiciones propias del mundo de la imprenta. Por un lado, la tradición científica y por otra diversas formas de la tradición literaria; por ejemplo, diccionarios, bestiarios o enciclopedias (Tredinnick, 2007). Obras como los diccionarios encierran en su propia naturaleza una idea de no linealidad del texto, sino de un entramado

conjunto de conexiones en el que unos términos se definen por su relación con otros.

Para Berners-Lee (1999: 16), el hipertexto debería ser capaz de integrar cualquier tipo de elemento: "El hipertexto funcionaría al máximo rendimiento si pudiera acceder a cualquier cosa concebible. Cada nodo, documento -o como se llamara-, sería fundamentalmente equivalente de alguna manera. Cada uno tendría una dirección de referencia. Todos existirían juntos en el mismo espacio: el espacio de la información."

De acuerdo con Tredinnick (2007), tanto la idea del hipertexto como la de la Web surgen como intentos de superar las limitaciones derivadas de organizar la información mediante el empleo de clasificaciones formales y rígidas.

2.3.3. El ciberespacio

El término "cyberspace" fue creado por el escritor William Gibson a principios de los ochenta. Algunos sitúan su primera aparición en la historia breve *Burning Chrome* de 1982, mientras que otros lo sitúan en la novela de 1981, *Neuromancer*. Gibson (1989) presenta el ciberespacio como una alucinación percibida a través de un ordenador que conecta al usuario con una matriz de una "complejidad inimaginable" situándolo en un ámbito hiperreal en la mente humana. El concepto tuvo una rápida popularidad y se extendió durante los años 80 y 90, siendo empleado en distintas formas para referirse al fenómeno emergente de la comunicación mediada por ordenador y a las tecnologías de realidad virtual. Formado por la palabra griega *kyber*, que significa navegar, su significado literal es "espacio navegable", lo cual resulta muy descriptivo.

Dodge y Kitchin (2001) señalan que no existe un ciberespacio homogéneo sino un conjunto de ciberespacios con características propias. Se trata de espacios que tienden a hacerse cada vez más híbridos y superpuestos, constituyendo extensiones de la realidad física. Se trata de un nuevo ámbito de realidad en el que

progresivamente se han ido desarrollando actividades humanas y al que, consecuentemente, las ciencias sociales han prestado atención como objeto de estudio. Se trata de nuevos espacios, de nuevas geografías que es necesario explorar y explicar. Al margen del espacio de la realidad virtual, podemos encontrar Internet en su sentido más amplio, espacios de realidad aumentada, y otros espacios de interacción más restringidos como puedan ser las intranets. John Perry Barlow (1996) es uno de los primeros que utiliza el término "ciberespacio" para referirse a Internet en su célebre "Declaración de independencia del Ciberespacio". Barlow pronunció este discurso el 8 de febrero de 1996 ante el foro de líderes políticos y empresariales reunido en Davos. La Declaración, no exenta de grandilocuencia, empieza con estas palabras:

> "Gobiernos del Mundo Industrial, vosotros, cansados gigantes de carne y acero, vengo del Ciberespacio, el nuevo hogar de la Mente. En nombre del futuro, os pido en el pasado que nos dejéis en paz. No sois bienvenidos entre nosotros. No ejercéis ninguna soberanía sobre el lugar donde nos reunimos."

El discurso desarrolla la idea de un Internet/Ciberespacio como un espacio de libertad que trasciende los dominios de la política y de los intereses económicos. Para Barlow (1996), el Ciberespacio no se puede crear "como si fuera un proyecto público de construcción", sino que es "un acto natural que crece de nuestras acciones colectivas". Por supuesto, no se trata de una visión exenta de críticas (Morrison, 2009).

El empleo del término ha tenido fortuna desigual en distintos países y actualmente ha perdido vigencia. El concepto, que se había utilizado tanto en el ámbito académico como en el imaginario colectivo, ha perdido utilidad ya que, al igual que ha ocurrido con la Web e Internet, el espacio y el ciberespacio se confunden en la conciencia de una única realidad. Como en otros casos, las novedades tecnológicas dan lugar a una nueva realidad distinta de la ya existente. Esto se mantiene así hasta que la práctica social habita dicho espacio convirtiéndolo en una extensión más de la sociedad y lo humano.

2.3.4. El significado y las implicaciones de la Web e Internet

La Web es una realidad que encierra múltiples dimensiones, significados e implicaciones de todo tipo. El objetivo de este apartado es poner de relieve algunas de las cuestiones más significativas por su especial incidencia social y económica. El conjunto de cuestiones que se presentan en los apartados siguientes pretende señalar algunos temas que pueden ser de especial interés para la investigación en Internet y la Web, así como para ilustrar el contexto de la investigación empírica llevada a cabo en la tesis doctoral.

2.3.4.1. La sociedad de la información

El desarrollo de la Web ha contribuido a la popularización de conceptos, como "Sociedad de la Información" o "Sociedad del Conocimiento", que se han utilizado en las últimas décadas a partir del desarrollo de las tecnologías de la información y de su impacto cada vez más global (Capurro y Hjørland, 2003). Entre los factores claves del desarrollo económico, la información ha ido adquiriendo cada vez una mayor relevancia. Una característica clave de la misma es su naturaleza digital o digitalizable, lo cual permite que se pueda reproducir una y otra vez sin sufrir ningún tipo de merma o desgaste por su uso. Por el contrario, se argumenta que justamente ocurre lo contrario: la difusión de la información genera un efecto positivo sobre sí misma, ya que permite aumentarla y mejorarla. Ejemplos en este sentido pueden ser iniciativas como el *software* libre o proyectos como la Wikipedia.

De acuerdo con Webster (1995), el concepto de sociedad de la información se puede analizar de acuerdo con cinco criterios:

1. Tecnológico, vinculado a la mera aplicación de tecnologías de la información en la sociedad;

2. Económico (Machlup, 1962; Boulding, 1966; Porat, 1977; Arrow, 1979);
3. Profesional (Bell, 1973; Porat, 1977);
4. Espacial, vinculado a la aparición de redes de información (Castells, 1989); y
5. Cultural, referido a la presencia de los medios en la sociedad.

Señalan Capurro y Hjørland (2003) que el cambio de la "sociedad de la información" a la "sociedad del conocimiento" pone el énfasis en el contenido, y no en la mera adopción de nuevas tecnologías, como principal elemento que facilita el desarrollo económico y social. Desde un punto de vista de la gestión del conocimiento, el término "información" hace referencia a datos que, tomados de forma individualizada, son significativos. Éstos, una vez que se integran en un contexto, se convierten en conocimiento (Probst, Raub y Romhard, 1999). Así, el concepto de "información" se sitúa en un lugar intermedio entre el de "dato" y el de "conocimiento". Hay toda una literatura de gran impacto en nuestros días relativa a la gestión del conocimiento, aunque no existe un consenso acerca de si éste como tal puede ser gestionado como si de un activo más se tratara o bien únicamente puede ser facilitado (von Krogh, Ichijo y Nonaka, 2000). La Web, en este sentido, tiene un papel muy importante, ya que en ella encontramos enormes cantidades de información en forma de documentos de muy diverso tipo, pero también de las relaciones existentes entre ellos a través de los hiperenlaces que los vinculan.

La idea de una economía cada vez más basada en la información y el conocimiento se encuentra en la base que motiva los trabajos empíricos que se incluyen en esta tesis doctoral. Por ejemplo, Cornella (2000) llega a señalar que las empresas son información. Manuel Castells (1996-1998; 2009) es uno de los sociólogos que con mayor interés ha teorizado y analizado la sociedad de la información.

La Web en sus 20 años de vida ha proliferado en todas las esferas de la vida humana, en cualquier tipo de interacción social, cultural, política, económica y

científica (Castells, 1996-1998; 2001). Manuel Castells (2001: 1) comienza su libro *The Internet Galaxy*, con las siguientes palabras:

> "The Internet is the fabric of our lives. If information technology is the present-day equivalent of electricity in the industrial era, in our age the Internet could be likened to both the electrical grid and the electrical engine because of its ability to distribute the power of information throughout the entire realm of human activity. Furthermore, as new technologies of energy generation and distribution made possible the factory and the large corporation as the organizational foundations of industrial society, the Internet is the technological basis for the organizational form of the Information Age: the network."

Actualmente, los medios de producción y distribución de información se han puesto a disposición de todos los ciudadanos con acceso a Internet, algo de tiempo y unos mínimos conocimientos técnicos. Son muchos los expertos que señalan que esta transformación en la naturaleza de los medios de producción de información constituye un cambio tan decisivo como el de la invención de la imprenta por Gutenberg. Es el caso de Manuel Castells (2001), que nos sitúa en la "Galaxia Internet", de forma paralela a la "Galaxia Gutenberg" acuñada por Marshall McLuhan. Nunca antes en la historia tantas personas habían opinado libremente de forma pública y al alcance de todo el mundo. Así, Castells (2001: 2) define Internet como "a communication medium that allows, for the first time, the communication of many to many, in chosen time, on a global scale".

2.3.4.2. Software libre y la ética del hacker

El movimiento de *software* libre que surgiera a principios de los años 80 de la mano de Richard Stallman, creador del sistema UNIX, se ha convertido en una de las fuerzas más revolucionarias dentro del mundo del *software* y fuera de él, impregnando con su espíritu buena parte de lo que ocurre en la Web. La lucha del *software* libre contra el *software* privativo, de código cerrado, imposible de copiar,

compartir y mejorar por los usuarios, se ha desarrollado alrededor de dos conceptos principales. Por un lado, el rechazo de las patentes de *software*, en tanto que enemigas de la innovación y del progreso humano; y, por otro, el desarrollo del trabajo entre iguales en comunidades.

El *software* libre ha permitido una descomposición de la cadena de valor en este sector, modificando modelos de negocio que ahora tienden más a proporcionar servicios de adaptación de *software* a las empresas, formación y mantenimiento. Antes, unas pocas empresas dominaban el mercado de las aplicaciones informáticas para empresas. La extensión del *software* libre ha hecho que existan programas de calidad para atender todas las funciones de la empresa sin necesidad de realizar grandes desembolsos en licencias. Ello hará que empresas que antes no se lo podían permitir puedan disponer de estas utilidades, aumentando previsiblemente su productividad y proporcionando ventajas significativas a la pequeña y mediana empresa. La adopción del *software* libre supone "adoptar nuevos modelos mentales y nuevas formas de conceptualizar la creación de valor" (Tapscott y Williams, 2007: 143).

Friedman (2005), en su popular libro *La tierra es plana,* señala entre los factores que han contribuido al desarrollo de un planeta más globalizado y sin barreras, el *software* libre y el acceso libre a la información. Sobre la nueva ética que subyace a estos cambios, Himanen (2003) desarrolló hace unos años el concepto de la *ética del hacker*. Esta cultura libertaria se encuentra en la base de la arquitectura de Internet y de la Web, lo cual lo convierte en una de sus máximas fortalezas, permitiendo un desarrollo continuo en el que los usuarios se convierten en productores de tecnología y en agentes transformadores de la red en su conjunto. Castells (2001: 28) señala: "It is a proven lesson from the history of technology that users are key producer of the technology, by adapting it to their uses and values, and ultimately transforming the technology itself". Este hecho en el caso de Internet se acentúa extraordinariamente ya que los cambios se producen en tiempo real a partir de las interacciones de los usuarios. De alguna manera se trata de: "a process of learning by producing" (Castells, 2001: 28).

2.3.4.3. La producción entre iguales y la cultura del remix

La participación se convierte en un imperativo ético. Si no se participa, no se contribuye a la comunidad. La Web, y de forma más evidente la Web 2.0, es un organismo vivo que necesita que sus usuarios contribuyan activamente para generar mayores beneficios para todos. Conceptos como los de "inteligencia colectiva" o "ciudadanía digital" no son más que utopías sin la asunción de Internet y la Web como medio fundamental para la autonomía personal. La participación requiere ser de alguna manera un activista. La pasividad no contribuye a la comunidad. Las aplicaciones Web 2.0 se han convertido en herramientas que permiten desarrollar y poner en práctica la inteligencia colectiva a partir de las aportaciones individuales, no coordinadas, de los usuarios. La Web 2.0 ha permitido desarrollar las herramientas para que las personas expertas en un determinado tema contribuyan a través de acciones particulares a un bien común. Para Ribes (2007), las actividades de la "inteligencia colectiva" en Internet pueden dividirse en tres grandes grupos: la producción de contenidos, la optimización de recursos y el control ejercido sobre contenidos e individuos. La producción de contenidos se realiza de forma individual por miles, incluso millones, de usuarios y los resultados globales se obtienen como agregación de las aportaciones de cada uno de ellos.

Henry Jenkins (2006), en su obra *Convergence Culture*, ha reflexionado sobre la cultura de la participación en Internet. Cuando los individuos han dispuesto de las herramientas apropiadas de edición y publicación, las grandes corporaciones se han visto sorprendidas al comprobar que están dispuestos a desarrollar una significativa labor creativa que antes no afloraba de una manera tan potente por la falta de medios tecnológicos apropiados (Weinberger, 2007). El acceso a la información, a creaciones musicales, vídeos, fotografías, etc. ha dado lugar a una cultura de la "remezcla", del "remix", de lo derivado, que ha impulsado el desarrollo de las licencias de "algunos derechos reservados", promovidas por Creative Commons, por ejemplo. Nos encontramos ante una arquitectura potencialmente abierta que permite combinar y reutilizar mis datos junto con los de otros usuarios, con la ayuda de interfaces flexibles y dinámicos que los grandes servicios *online* permiten personalizar a nuestro gusto. Existen también posturas críticas como las

de Andrew Keen (2007) que, en *The Cult of the Amateur*, sostiene una visión pesimista al afirmar que la participación en masa en Internet lo que realmente ha provocado es ruido, obstaculizando el surgimiento de verdadero talento.

La producción entre iguales en comunidades abiertas a través de Internet se ha revelado como una poderosa arma en determinados ámbitos, como las comunidades de programadores y de científicos, así como en la propia dinámica del proyecto enciclopédico de la Wikipedia. Este tipo de colaboración plantea prometedoras posibilidades para el desarrollo de determinadas funciones de las empresas (Tapscott y Williams, 2007). No todo son, sin embargo, ventajas para ellas. La producción entre iguales implica menos control y necesidad de asumir y adaptarse a las dinámicas de participación propias de las comunidades en red. Ello obliga a rediseñar los modelos de negocio para poder seguir generando beneficios, crear nuevos incentivos para los trabajadores dentro de la empresa, aportar nuevo valor añadido, etc. Otra forma que tienen las empresas para optimizar recursos es acudiendo a Internet para innovar (a través de servicios como *InnoCentive*) o bien empleando prácticas de *crowdsourcing*, que surge en determinados campos como alternativa al *outsourcing*. Consiste en proponer problemas y dar inventivos a quien o quienes propongan soluciones, utilizando para ello principalmente los medios que proporciona Internet. *Crowd* hace referencia a multitud; mientras que *sourcing*, a la obtención de materia prima. En el Apartado 2.3.6.2 abordaremos más en profundidad las implicaciones para la empresa al referirnos a lo que McAfee (2006) ha denominado la Empresa 2.0.

2.3.4.4. Folcsonomía

La Web ha facilitado el surgimiento de nuevas formas de organizar el conocimiento. La *folcsonomía* es el orden resultante de la aplicación de etiquetas a los contenidos digitales de forma libre e independiente por parte de los individuos. Se contrapone a la idea de taxonomía, como orden previo y apriorístico dentro del cual hay que ajustar las obras y contenidos. El etiquetado de contenidos digitales (como ocurre

en sitios como Flickr o Youtube) nos aproxima a la idea de la *Web semántica* (Berners-Lee y Hendler, 2001; Berners-Lee, Hendler y Lassila, 2001), esa Web en la que las máquinas serían capaz de interactuar entre ellas al poder procesar el significado de la información que manejan.

Las *folcsonomías*, en tanto que no existe un único orden posible sino tantos como los resultantes de la agregación de las etiquetas de contenido, constituyen uno de los mejores ejemplos de aplicación de la "inteligencia colectiva", donde las decisiones individuales del individuo forman parte de un todo en el que "significan" más que consideradas por sí solas aisladamente. Con todo, no se trata de un sistema exento de limitaciones o problemas.

2.3.4.5. Confianza y reputación

Las monedas de cambio en Internet son la confianza y la reputación que los usuarios adquieren y otorgan a otros. Las relaciones se basan en la confianza, al tiempo que se establecen "jerarquías" en función de la reputación. Los iguales son los que se califican entre sí, contando con diversos medios para ello (valoraciones directas, audiencias, enlaces que guían a un sitio, recomendaciones, votaciones, etc.). La reputación funciona como una garantía de relevancia.

Como hemos señalado, en un entorno en red, la reputación constituye un filtro fundamental de la información en sistemas *peer-to-peer*, redes sociales o en creación de conocimiento y desarrollo de economías ligadas a la atención. Son los pares, los participantes en el sistema, los que otorgan la reputación a un individuo o entidad mediante la expresión de opiniones, el establecimiento de enlaces, la inclusión en agregadores de fuentes, la valoración mediante sistemas de *rating*, la inclusión en una lista de "favoritos", etc. La implicaciones de obtener una determinada reputación son muy importantes, tanto en una perspectiva en red como en la vida *offline*, conllevando en ocasiones la posibilidad de cargar precios más altos por determinados productos o servicios o directamente alcanzar mayores

niveles de influencia. También puede ocurrir lo contrario en los casos en que un actor pierda su reputación (un caso clásico es el de la empresa *Kriptonite*, cuyos candados de alta seguridad para bicicletas podían abrirse con un simple bolígrafo Bic, como mostraba un vídeo difundido en *Youtube*). Las "economías de la reputación" son una consecuencia directa del desarrollo de una "inteligencia colectiva" o del concepto de "wisdom of crowds" (Surowiecki, 2005).

2.3.4.6. Propiedad intelectual

El cuestionamiento de las leyes de propiedad intelectual tradicionalmente vigentes es algo profundamente arraigado en el nuevo entorno digital. La problemática que origina la aplicación de leyes nacidas en la economía industrial a una economía o entorno social basados en los flujos de información puede entenderse de forma más global si se pone en el contexto de la "cultura del remix", que mencionamos con anterioridad. Las industrias del ocio y del *software* son unas de las principales afectadas en este proceso de digitalización de los contenidos. Sin embargo, no estamos principalmente ante un problema de "piratería", sino ante un cambio en los modelos de producción y distribución de un contenido que se puede transmitir en lenguaje binario. Hay dos factores básicos en este proceso: la práctica eliminación de los costes de reproducción de los contenidos y la posibilidad de supresión de la intermediación en la relación creador-consumidor, si bien ambos roles a veces se superponen y confunden.

La crisis de los derechos de autor, tal y como los hemos conocido hasta ahora, se ilustra muy claramente en el caso de las industrias de la música, el cine, la televisión o el negocio editorial. Una alternativa adaptada a los movimientos de Internet basados en la cooperación, en la remezcla, en el compartir contenidos, es la planteada por Creative Commons y su conocido "algunos derechos reservados" ("*some rights reserved*"), frente al tradicional y restrictivo "todos los derechos reservados" ("*all rights reserved*"). Lessig (2004), con su libro *La liberación de la cultura*, es el principal adalid de esta revisión de los derechos de autor relativos a

las obras de creación. Estas licencias están inspiradas en las licencias GNU-GPL (*GNU General Public License*), base del *software* libre.

2.3.5. Otras visiones desde la literatura, la filosofía y la cultura

La literatura, principalmente desde la ciencia-ficción, ha constituido siempre una fuente de inspiración para futuras invenciones. La idea de Internet y la Web resulta tan fantástica a la par que aparentemente sencilla, que no han sido excesivas las referencias directas que se le puedan atribuir. Algunos, por ejemplo, han recurrido repetidas veces a Borges y su *Biblioteca de Babel*, el *Aleph* o *Tlön, Uqbar, Orbis Tertius*; otros han rememorado la idea de Julio Verne de una red telegráfica mundial.

Los principales autores que desde distintos campos prefiguraron el sueño de una red de información mundial han sido expuestos en páginas anteriores. Sin embargo, en este punto, queremos poner en solfa el revelador paralelismo entre literatura y tecnología. Por ejemplo, el artículo de Vannevar Bush, "As we may think" (1945), es contemporáneo del libro de Borges, "Ficciones" (1944), donde encontramos relatos como la Biblioteca de Babel, ya abordado, y Tlön, Uqbar, Orbis Tertius. Esta última obra parece imaginar el proyecto enciclopédico de la Wikipedia y su funcionamiento (Llosa Sanz, 2006). Borges (1944/2005b: 434) describe Tlön como una enciclopedia: "Ahora tenía en las manos un vasto fragmento metódico de la historia total de un planeta desconocido, con sus arquitecturas y sus barajas, con el pavor de sus mitologías y el rumor de sus lenguas, con sus emperadores y sus mares, con sus minerales y sus pájaros y sus peces, con su álgebra y su fuego, con su controversia teológica y metafísica. Todo ello articulado, coherente, sin visible propósito doctrinal o tono paródico". El escritor argentino (1944/2005b: 434-435) se pregunta también por la autoría de tamaña obra: "¿Quiénes inventaron a Tlön? El plural es inevitable, porque la hipótesis de un solo inventor -de un infinito Leibniz obrando en la tiniebla y la modestia- ha sido descartada unánimemente. [...] Ese plan es tan vasto que la contribución de cada escritor es infinitesimal". Y sobre el orden de semejante creación se afirma lo siguiente (1944/2005b: 435): "Al principio

se creyó que Tlön era un mero caos, una irresponsable licencia de la imaginación; ahora se sabe que es un cosmos y las íntimas leyes que lo rigen han sido formuladas, siquiera en modo provisional."

Por su parte, en los 60, mientras Nelson madura su idea de hipertexto, Julio Cortázar publica *Rayuela* (1963) y Umberto Eco su *Opera Aperta* (1962/1984). El libro Rayuela presenta dos modos de lectura, uno lineal y otro siguiendo el orden propuesto en el "Tablero de dirección" de las primeras páginas del libro. Esta técnica, que se asemeja al *collage,* también sugiere la idea del hipertexto. Indican las instrucciones de lectura de Rayuela (Cortázar, 1963/1997:111): "A su manera este libro es muchos libros, pero sobre todo es dos libros. El lector queda invitado a elegir una de las dos posibilidades siguientes: [...]". El hipertexto hace que sean tecnológicamente posibles las conexiones que en otros textos impresos estaban implícitas o sujetas a la disponibilidad de acceso a otras fuentes.

Al margen de la literatura, es posible acercarse a Internet y la Web desde una perspectiva filosófica. Tredinnick (2007) subraya su conexión con el Post-estructuralismo, realizando una revisión de las aportaciones de autores como Foucault, Kristeva, Barthes, Derrida o Genette. Desde el punto de vista de esta corriente de pensamiento, los textos encuentran su significado en las relaciones y diferencias con otros textos, de modo que el significado no existe como tal en el texto en sí sino que se genera a partir de la intertextualidad. Así se expresan, por ejemplo, Foucault (1972) o Kristeva (1980). El concepto de intertextualidad entronca perfectamente con el concepto de hipertexto. De alguna manera, el hipertexto viene a materializar las conexiones existentes entre los distintos textos a través de los hiperenlaces. Nelson y Berners-Lee entienden el hipertexto como un modo de romper las barreras del propio texto a través de las conexiones que se realizan con otros. Como Berners-Lee (1999) apunta, todo al fin y al cabo es estructura. La ruptura de los límites del texto a través de la tecnología viene a subvertir de algún modo la tradición de los textos impresos que surge con fuerza a partir de la invención de la imprenta, devolviendo a la generación de los discursos su carácter colectivo.

Montoya Juárez (2008) hace una completa revisión de las relaciones entre lo virtual, especialmente Internet y el hipertexto, y la literatura. En concreto, se detiene en la construcción del concepto de simulacro y en las visiones posmodernas vinculadas a este nuevo ámbito. En su opinión, el ordenador es la máquina que mejor representa la posmodernidad. La interconexión de los ordenadores formando Internet permite ir aún más allá en esta imagen. A lo largo del siglo XX, se ha construido un conjunto de utopías e antiutopías que actualmente cobran especial sentido en el marco de Internet y la Web. Para Montoya Juárez (2008), hay dos interpretaciones principales del impacto de estas tecnologías. La primera engarza con la tradición antiutópica que encabezan autores como Orwell, en la que el ordenador encarnaría ese Gran Hermano que simboliza el control y el dominio totalitario. La segunda es de tipo anarquista, en ocasiones con tintes utópicos, metáfora de una sociedad autogestionada, de un conocimiento cooperativo (Zizek, 1996), cuyo gobierno se ejerce desde abajo, conduciendo a lo que Vattimo denominaba una "sociedad más transparente" (Vattimo, 1990). En este sentido se incluyen textos como la "Declaración de independencia del Ciberespacio", de John Perry Barlow (1996) o los puntos de vista de Tim Berners-Lee (1999), que ve la Web como un medio de cooperación y desarrollo de la humanidad.

2.3.6. Internet y la empresa

2.3.6.1. e-Business

Tres procesos confluyen en el último cuarto del siglo XX, dando lugar a una nueva estructura social basada fundamentalmente en las redes (Castells, 2001; 2009): la necesidad económica de unos sistemas de gestión flexibles y de una globalización del capital, la producción y el comercio; la demandas sociales de libertad individual y de comunicación abierta; y los extraordinarios avances de la informática y de las telecomunicaciones, gracias al gran desarrollo de la microelectrónica.

La variedad de posibilidades de uso de la Web, y de Internet en su conjunto, unida

a una reducción generalizada de los costes de conexión y de los equipos necesarios para ello, han redundado en que un número creciente de personas se conecte en red, empleando distintos tipos de dispositivos, desde diversos lugares y durante más tiempo. El poblamiento de Internet ha supuesto que la gran mayoría de actividades sociales, económicas, políticas, educativas, de toda índole, que hasta hace poco se desarrollaban únicamente en el mundo físico, ahora encuentren reflejo en Internet o hayan desarrollado fenómenos que sólo existen *online*. Entre ellos, se podría mencionar el desarrollo de una potente red de blogs a nivel mundial o la profusión de las redes sociales virtuales (por ejemplo, Facebook o Tuenti). De acuerdo con los datos del informe *La Sociedad en Red. Informe anual 2008*, elaborado por el ONTSI (Observatorio Nacional de las Telecomunicaciones y la Sociedad de la Información; 2009), según las estadísticas más recientes de Internet World Stats, en marzo de 2009 existían 1.596 millones de usuarios de Internet en el mundo. Por su parte, en España, más de ocho millones de hogares, el 51%, estaban conectados a Internet en 2008. Ese mismo año el número de personas de diez y más años de edad que habían usado Internet en alguna ocasión, alcanzaba los 23,5 millones.

Estas cifras tan elevadas se han visto acompañadas de una significativa presencia de las empresas en la Web. Si nos centramos en el ámbito español, dicho informe (ONTSI, 2009) señala algunos datos que hablan de la importancia creciente de Internet para la economía:

- El nivel de acceso a Internet, la telefonía móvil y la banda ancha en las pymes y grandes empresas alcanza porcentajes por encima del 90%.

- Entre las empresas con Internet, que es la gran mayoría, la disponibilidad de página web se sitúa en el 57,5%. Este porcentaje alcanza el 72% entre las empresas medianas y supera ya el 50% entre las pequeñas. Las funciones más frecuentes de la página web son para presentación de la empresa (88,2%) y para el acceso a catálogos de productos (56,9%).

- En relación con el uso de la Web por parte de las pymes y grandes empresas, los usos principales son buscar información (97,2%) y acceder a

servicios bancarios y financieros (86%). Algunos de los usos con un mayor crecimiento son los servicios de posventa o preventa y la observación del comportamiento del mercado.

Del auge de Internet y de sus implicaciones comerciales dan buena cuenta las cifras de inversión publicitaria (ONTSI, 2009). En el contexto mundial, PricewaterhouseCoopers prevé una tasa compuesta de crecimiento anual para el mercado de la publicidad en Internet del 19,5%, hasta alcanzar los 120,4 billones de dólares (período 2008-2012). España se encuentra entre los países europeos que más inversión publicitaria aún realiza en medios convencionales; sin embargo, el fuerte crecimiento de la inversión *online* entre 2005 y 2007 (de un 2% a un 6% del total de publicidad en medios convencionales) ha reducido significativamente las diferencias con los países de referencia en Europa.

Una mayor publicidad *online* irá de la mano de una mayor presencia de las empresas en la Web, lo cual apunta a que en los próximos años se producirá una evolución significativa en los modelos de negocio hacia un mayor énfasis en la información y en el comercio a través de la red, esto es, hacia el *e-Business*. Castells (2001: 66) define *e-Business* como "any business activity whose performance of the key operations of management, financing, innovation, production, distribution, sales, employee relations, and customer relations takes place predominantly by/on the Internet or other networks of computer networks, regardless of the kind of connection between the virtual and the physical dimensions of the firm". El *e-Business* es también la empresa en red o la empresa red como conjunto de interrelaciones, internas y externas, con los distintos agentes económicos. Según Castells (2001), la empresa red no está restringida a la industria tecnológica, sino que se expande rápidamente por otros sectores de actividad. Si bien, en su opinión, la empresa red es previa a la difusión de Internet a través de la Web, considera que este medio le aporta unas características genuinas a la misma:

- Escalabilidad (*Scalability*): la red puede incluir tantos componentes como

sean necesarios, en un ámbito local o global, para cada operación o transacción que se lleve a cabo. El ser local o global no es un obstáculo técnico para la organización, que puede expandirse o retraerse conforme lo exija la estrategia de la empresa; todo ello sin incurrir en un excesivo coste derivado de mantener capacidad de producción no utilizada.

- Interactividad (*Interactivity*): ya sea en tiempo real o en el intervalo de tiempo que se elija entre los actores que participan en la actividad de la organización (proveedores, clientes, directivos, empleados, etc.). El sistema de comunicación se distribuye y se expande en múltiples direcciones generándose un flujo continuo de información que permite ajustar continuamente la actividad de la empresa.

- Gestión de la flexibilidad (*Management of flexibility*): permite mantener el control sobre el negocio durante el continuo proceso de expansión y adaptación. La flexibilidad también debe permitir el ajuste en función de cada proyecto que se lleve a cabo.

- *Branding*: se trata de un elemento esencial como signo reconocido del valor de un negocio, de la capacidad de creación de valor de una organización. El *branding* en la era de Internet adquiere una dimensión distinta ya que en muchas ocasiones los proyectos son resultado de la cooperación entre múltiples partes. Se trata pues de un reto a la vez que una oportunidad. Un ejemplo de actualidad es el de los proyectos de Internet basados en *mashups*, donde varios servicios se combinan creando uno nuevo. Así ocurre con los mapas de *Google maps*, que adquieren una gran difusión a través de su integración en servicios prestados por terceros.

- Personalización (*Customization*): Internet permite ofrecer un producto personalizado a cada consumidor, así como atender nichos de mercado que antes estaban desatendidos, o generar demandas que no existían (Anderson, 2006; Tapscott y Williams, 2007).

Dos de las cuestiones más destacadas cuando hablamos de una economía en red son: por un lado, el cambio en los modelos de negocio, en la línea de lo expuesto anteriormente; y, por otro, la gestión de los recursos intangibles de la empresa, cada vez más vinculados a la información que son capaces de procesar y de generar, especialmente en la red. El problema de los intangibles tiene un largo recorrido en áreas como la Contabilidad, que durante años se ha esforzado en cuestiones como su identificación y valoración. En esta línea podemos recordar la crisis de las empresas tecnológicas *puntocom*, que estalló en el año 2000. La presencia en Internet de las empresas es un activo intangible más y un modo de medirlo es a través del número de enlaces que reciben dichas páginas.

2.3.6.2. La empresa 2.0

El concepto de Empresa 2.0 (*Enterprise 2.0*, en el original inglés) fue acuñado por el profesor Andrew McAfee (2006) para referirse al empleo de las emergentes plataformas de *software* social dentro de las propias empresas o entre empresas y sus clientes y otros terceros. Su inclusión responde a su conexión directa con la Web 2.0 y su incidencia, expuesta en páginas anteriores. La idea de Empresa 2.0 que McAfee propone está estrechamente vinculada con una empresa en red basada en el conocimiento. De acuerdo con sus propias palabras (McAfee, 2006: 28): "Enterprise 2.0 technologies have the potential to usher in a new era by making both the practices of knowledge work and its outputs more visible."

Las tecnologías a las que se refiere McAfee comprenden entre otras las siguientes (entre paréntesis se citan algunas empresas que, según Matuszak (2007), las emplean):

- Blogs (General Motors, Hitachi, Intel, Novell). En el ámbito español, podemos mencionar la red de blogs del BBVA (http://bbvablogs.com).
- Wikis (Dresdner Kleinwort, Microsoft, Nokia, SAP).

- RSS (Amazon, Cisco, The Wall Street Journal). Actualmente prácticamente todos lo periódicos con presencia en la red han incorporado esta tecnología, permitiendo a los lectores suscribirse a hilos de información que sean de su interés. Por poner un ejemplo, en el ámbito académico este sistema es utilizado por la base de datos ISI Web of Knowledge para ofrecer a los usuarios información sobre los nuevos artículos que citan una determinada referencia bibliográfica.

- Etiquetas (*tags*) (Honeywell, IBM, Sony-BMG).

- Redes sociales (Cisco, Dresdner Kleinwort, Microsoft, Nike).

- *Mashups* (Amazon, Google, IBM, Siemens, Société Générale).

- Mercados de predicción (Google, HP, Microsoft, Yahoo).

- Metaversos, como *Second Life* (IBM, Pontiac, Sun Microsystems, Dell, Reuters, Cisco Systems)

En su artículo, McAfee (2006) contrapone dos paradigmas tecnológicos distintos, ambos caracterizados por un conjunto de componentes básicos. Frente al denominado WIMP (*Windows*, *Icons*, *Menus*, *Pointers*), formado por el sistema de ventanas, iconos, menús y "señaladores", que es el más generalizado entre los usuarios hoy en día, surge el paradigma SLATES (*Search*, *Links*, *Authoring*, *Tags*, *Extensions*, *Signals*), compuesto de búsqueda, enlaces, criterio de autoridad, etiquetas, extensiones y señales, lo cuales constituyen los componentes tecnológicos de la Empresa 2.0.

Los seis componentes de la Empresa 2.0 presentan las siguientes características:

- Búsqueda (*Search*): es fundamental que los usuarios tengan las herramientas de búsqueda apropiadas para conseguir la información que precisen. Se imponen por su facilidad de uso y efectividad la búsqueda mediante palabras claves, frente a los habituales sistemas de ayuda de

navegación de las *intranets*. Hoy en día está al alcance de cualquier empresa incorporar herramientas de búsqueda en las *intranets* tan potentes como las empleadas por el buscador de referencia, Google. Algunos de los futuros desarrollos tecnológicos se encaminan a "humanizar" aún más las búsquedas de información, en línea con la idea de Web Semántica (Apartado 2.4.2.10).

- Enlaces (*Links*): permiten estructurar el contenido de la red, proporcionando una forma de valorar los contenidos en función del número de veces que son enlazados. Se supone que los mejores contenidos son los más enlazados, o dicho de otra manera, los más enlazados son los que más recomendaciones de lectura reciben por otras páginas web. Se trata de la idea en la que se basa el sistema *PageRank* de Google (Apartado 3.3.3.4). Dentro de las *intranets* de las empresas, este sistema presenta algunas limitaciones derivadas fundamentalmente del reducido número de personas que tienen la capacidad de generar enlaces. Es necesario, por lo tanto, que el mayor número de usuarios posible participe en la construcción de la *intranet* empresarial, lo cual conllevará una mejor estructuración de los contenidos y promoverá la creación de conocimiento.

- Autoría (*Authoring*): en la naturaleza humana está el deseo y la necesidad de contar y escuchar historias, generando una memoria colectiva recogida históricamente en la literatura popular de tradición oral. En este sentido la Web 2.0 supone una vuelta a los "orígenes". Las sistemas de gestión de contenidos como los blogs o los wikis proporcionan medios de producción de contenidos a los usuarios y éstos han respondido con una masiva difusión y uso de estas tecnologías. El desarrollo de la Wikipedia, los millones de blogs existentes o la gran popularidad de las redes sociales son buenos ejemplos de ello. Estos instrumentos pueden permitir que los recursos humanos de la empresa compartan su conocimiento, en ocasiones tácito, de manera natural aportando, por ejemplo, sus experiencias, comentarios, puntos de vista, opiniones, enlaces de interés, recomendaciones, etc. Todo ello puede aflorar en la empresa de forma más

sencilla, fácil y humana, gracias a las tecnologías 2.0.

- Etiquetas (*Tags*): consisten en palabras descriptivas de contenidos digitales que permiten categorizar el contenido. Muchas aplicaciones, especialmente aquellas que contienen gran cantidad de información, permiten a sus usuarios añadir palabras claves para facilitar su búsqueda. De esta práctica surge el concepto de folcsonomía (Apartado 2.3.4.4), que es la categorización que surge de la descripción más o menos libre de contenidos por parte de los usuarios, frente al tradicional concepto de taxonomía. La principal ventaja del sistema es que permite reflejar las estructuras de los contenidos y las relaciones entre ellos que la gente utiliza de manera efectiva.

- Extensiones (*Extensions*): se refieren al empleo de algoritmos que permiten proporcionar al usuario información que puede ser relevante para sus intereses y en función de sus preferencias. Esta idea ha sido explotada con mucho éxito por empresas como Amazon, que incluye en sus ofertas de productos mensajes como "Customers who bought this item also bought".

- Señales (*Signals*): es una idea que se basa en el empleo de tecnologías de sindicación como las mencionadas anteriormente RSS (o Atom). Son herramientas que avisan al usuario cuando hay nuevos contenidos de su interés, a los que previamente se ha suscrito.

Al margen de los componentes tecnológicos, las ideas básicas que mueven la Empresa 2.0 son:

- Su facilidad de uso, permitiendo que los usuarios, sin conocimientos técnicos, puedan participar en la misma creando y compartiendo información sin limitaciones significativas.

- Las herramientas informáticas de escritorio pierden relevancia frente al navegador, que se convierte en la puerta hacia todo tipo de servicios

ofrecidos en red.

- No se impone a los empleados un modo cerrado y determinado de gestión del conocimiento. No existen ideas preconcebidas sobre cómo se debe trabajar o sobre cómo se deben estructurar los contenidos que se producen. Tanto los wikis como los blogs funcionan como plataformas en blanco sobre las cuales los usuarios empiezan a generar contenidos. De igual modo, una folcsonomía no existe hasta que los usuarios no empiezan a etiquetar.

El papel de los gerentes en la implantación de las tecnologías de la Empresa 2.0 es decisivo, ya que la difusión del uso de las mismas no es algo que ocurra de forma automática, sino que necesita de unos incentivos y de un compromiso activo que debe partir de la dirección.

2.4. Internet y la Web en la investigación

La Web, como objeto de estudio, ha recibido la atención de los investigadores desde su creación; sin embargo, no siempre se ha contado con herramientas adecuadas ni con marcos teóricos que permitieran explicar los nuevos fenómenos que se iban desarrollando. Desde distintas áreas de conocimiento se han efectuado acercamientos a este nuevo espacio. En algunos casos podemos decir que fue fruto de una evolución lógica, por ejemplo, cuando nos referimos a áreas de conocimiento como Ciencias de la Información o a la aplicación de la teoría de grafos o el estudio de redes sociales.

En páginas anteriores nos hemos aproximado a la Web y a Internet desde el punto de vista de su impacto social y económico, así como de su influencia en el desarrollo de nuevas formas de entender el mundo. En este apartado nos centramos en una perspectiva de investigación científica. En el Apartado 2.4.1, nos referiremos brevemente a Internet y la Web como medios para la investigación científica. Las ciencias naturales y sociales han encontrado en las redes un gran potencial para desarrollar proyectos que anteriormente se enfrentaban a serias

limitaciones. El empleo de redes proporciona ventajas de muy diverso tipo, desde un mayor aprovechamiento de la capacidad de computación de ordenadores que están distribuidos en red hasta el desarrollo de investigación con grupos e investigadores localizados en lugares muy distantes. Se habla de e-Ciencia (*e-Science*) y de e-Investigación (*e-Research*), enfocados fundamentalmente a las ciencias naturales y experimentales. Más recientemente se ha empezado también a hablar de e-Ciencia Social (*e-Social Science*) y de e-Humanidades (*e-Humanities*).

Por otra parte, en el Apartado 2.4.2, vamos a tratar la Web en tanto que objeto de estudio. Se trata de una aproximación más técnica que persigue poner de relevancia alguna de las características de la Web. Veremos, por ejemplo, la importancia de distinguir entre conceptos como sitio web o página web, entre otros. Conocer las características de la Web permitirá entender mejor la naturaleza del contexto en que se desarrolla nuestra investigación, así como comprender sus limitaciones y sus posibilidades de desarrollo. Por último, se presentarán algunas de las fronteras en las que se mueve la Web en nuestro tiempo, al objeto de conocer la previsible evolución y, por tanto, anticipar, si es que fuera posible, los nuevos retos de investigación que deberán afrontarse en un futuro próximo.

Una buena muestra del interés suscitado por este tipo de estudios es la aparición en la última década de institutos de investigación en Internet en las universidades más prestigiosas del mundo. Por ejemplo, con una perspectiva más volcada en las ciencias sociales, podemos citar el *Oxford Internet Institute* en el Reino Unido y el *Berkman Center for Internet and Society* de la Harvard University en los Estados Unidos.

2.4.1. La Web como medio para la e-Ciencia

El surgimiento de Internet y de la Web está muy vinculado a actividades científicas, especialmente en una dimensión comunicativa que pretendía la difusión del conocimiento y la cooperación entre científicos. Castells (2001: 60) subraya

claramente esta perspectiva científica en relación con la creación de Internet:

> "At the top of the cultural construct that led to the creation of Internet is the techno-meritocratic culture of scientific and technological excellence, emerging essentially from big science and the academic world. This techno-meritocracy was enlisted on a mission of world domination (or counter-domination) by the power of knowledge, but kept its autonomy, and relied on a community of peers as the source of its self-defined legitimacy."

Ya hemos observado que el impacto de Internet y la Web es muy significativo en términos sociales, afectando prácticamente a todos los ámbitos de desarrollo del ser humano. Esto se debe fundamentalmente a que la capacidad de comunicación consciente a través del lenguaje es lo que hace al ser humano único (Castells, 2001). Dado que Internet transforma el hecho comunicativo, se genera también un cambio radical en el modo de interactuar de los individuos. En el campo de la ciencia, la transformación que se ha producido no ha sido menor que en el campo social. En todo caso, todavía podemos afirmar que nos hayamos en una primera fase a la luz del potencial de desarrollo tecnológico al que previsiblemente nos enfrentaremos en las próximas décadas.

Las implicaciones en la ciencia de todos estos avances se reconocen ya en los años 60, cuando Alvin Weinberg (1961) se refiere en la revista *Science* a la superación de la ciencia a pequeña escala. Desde entonces han surgido varios conceptos que intentan describir un nuevo modo de hacer ciencia que se desarrolla aprovechando fundamentalmente la infraestructura de Internet. Entre ellos se pueden citar: Ciberciencia (*Cyberscience*), Ciberinfraestructura (*Cyberinfrastructure*), e-Ciencia (*e-Science*), e-Investigación (*e-Research*), e-Ciencia Social (*e-Social Science*).

La definición de Ciberciencia que ofrece Nentwich (2003: 22) es amplia: "... all scholarly and scientific research activities in the virtual space generated by the networked computers and by the advanced information and communication technologies in general". El origen del término parece situarse en un artículo de

Paul Wouters (1996). Sin embargo, tras unos años de popularidad, parece que empieza a entrar en desuso (Jankowski, 2009).

La e-Ciencia, según Jankowski (2009), hace referencia fundamentalmente a las ciencias naturales y biológicas y al procesamiento de grandes volúmenes de información mediante computación grid. La computación grid es una tecnología innovadora que permite utilizar de forma coordinada todo tipo de recursos no sujetos a un control centralizado. La arquitectura grid permite coordinar distintos elementos como son el acceso a infraestructuras científicas remotas, a recursos computacionales o a información almacenada en bases de datos especializadas. La coordinación de los recursos interconectados se realiza fundamentalmente mediante el empleo de Internet. La propia Web constituye en sí misma una inmensa base de datos que puede emplearse con fines de investigación. Para Beaulieu y Wouters (2009), e-Ciencia se define por la combinación de tres diferentes desarrollos: la puesta en común de recursos computacionales, el acceso distribuido a conjuntos de datos masivos y el uso de plataformas digitales para la colaboración y comunicación. De acuerdo con sus propias palabras (Beaulieu y Wouters, 2009: 55): "The core idea of e-science is that knowledge production will be enhanced by the combination of pooled human expertise, data and sources, and computational and visualization tools". El *National e-Science Centre* define e-Ciencia como: (NeSC, http://www.nesc.ac.uk/nesc/define.html, consultado 8 febrero 2010):

> "What is meant by e-Science? In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual."

Este concepto alcanza su máxima difusión y popularidad a través del informe Atkins (2003) "Revolutionazing Science and Engineering Through Cyberinsfraestructure". El informe se refiere a una infraestructura distribuida de ordenadores, información y

tecnologías de comunicación (Jankowski, 2009: 5-6). Los conceptos, e-Ciencia y Ciberinfraestructura hacen referencia fundamentalmente al ámbito de las ciencias experimentales, naturales y biológicas, por ejemplo, en disciplinas como la astronomía, la física de partículas, la meteorología o la investigación del ADN.

Sin embargo, estas iniciativas no se han visto restringidas únicamente a este ámbito científico. Así, por ejemplo, en 2006, el *American Council of Learned Societies* (ACLS, 2006) emitió un informe sobre ciberinfraestructuras en las humanidades y las ciencias sociales. Otros intentos han supuesto la introducción del análisis de redes sociales como una herramienta para el estudio de las comunidades científicas (SNAC, 2005) o en la incorporación de la investigación en Internet y estudios digitales dentro de los programas universitarios (Nissembaum y Price, 2004).

Beaulieu y Wouters (2009) indican que a través del empleo del término e-Investigación (*e-Research*) se debe abrir la puerta a diversas formas de investigación que se centren en el uso de los nuevos medios y de las redes digitales y no necesariamente en empleo intensivo de los recursos de computación. Esto permite que las ciencias sociales y las humanidades puedan encontrar un mejor acomodo en este nuevo impulso investigador, que podría incluso considerarse un cambio de paradigma científico (Kuhn, 1970/1962). El término e-Investigación se puede entender como el sucesor de la noción de Ciberciencia (Jankowski, 2009). El término no se centra principalmente en ordenadores para el procesamiento de un gran volumen de información, sino en la incorporación de una amplia variedad de nuevos medios y redes electrónicas en el proceso investigador. Esta lista puede ser continuada haciendo referencia a la aplicación de las herramientas de la Web 2.0 a la ciencia, en especial, innovaciones como el "cloud computing". Jankowski (2009: 7) ofrece la siguiente definición de e-Investigación: "a form of scholarship conducted in a network environment utilizing Internet-based tools and involving collaboration among scholars separated by distance, often on a global scale". Algunas características fundamentales de este nuevo entorno son (Jankowski, 2009):

- Un incremento en el grado de informatización del proceso investigador, generalmente implicando trabajo en red.

- Una confianza en estructuras de organización virtuales basadas en redes para llevar a cabo la labor investigadora, implicando además un mayor grado de colaboración entre investigadores en el ámbito internacional.

- El desarrollo de herramientas basadas en Internet, lo cual facilita muchas fases del proceso investigador, desde la recogida de datos, su procesamiento y análisis o la publicación de resultados tanto a través de medios informales nativos de la web, como pueden ser los blogs, como de medios formales, con un mayor desarrollo de las revistas digitales.

- El desarrollo de instrumentos de visualización de la información, como respuesta a la necesidad de procesar grandes volúmenes de información, así como la exploración de las posibilidades multimedia para comunicar dicha información.

En todo caso, queremos subrayar de manera muy especial que la aparición de Internet y la Web constituyen una verdadera revolución en el estudio, no solo de cuestiones propias de las ciencias naturales y experimentales, sino muy significativamente de las ciencias sociales. Hoy en día, prácticamente cualquier comportamiento humano se lleva a cabo o tiene su reflejo en algún tipo de red. La Web se ha convertido, de este modo, en un gran lienzo en el que cada persona o institución va dejando su rastro conforme visita páginas, busca información, realiza transacciones económicas, etc. En muchos casos ni siquiera es preciso estar conectado a un ordenador. Internet es más ubicuo que nunca. Basta con emplear un teléfono, un lector de libros digitales, un reproductor de música o un GPS para generar información que puede ser procesada por terceros a miles de kilómetros de nosotros y en tiempo real. Internet y la Web se han convertido en una inabarcable base de datos donde cualquier posibilidad de investigación está abierta. Ello no exime, sin embargo, de resolver importantes dilemas éticos, constituyéndose esta

materia en un ámbito de especial interés dentro de este tipo de investigación (Jankowski y Van Selm, 2007).

La e-Ciencia Social (*e-Social-science*), de acuerdo con el *National Centre for e-Social Science* (NceSS), se refiere a "collaboration between computer scientist and social scientist to design and develop middleware in order to address social scientists' substantive research problems in new ways that recognize more fully the complexity of economic and social activities" (NceSS, e-Social science newsletter Issue 1, Summer 2000: 1; citado por Wessels y Craglia, 2009). En los últimos años, se ha producido un gran desarrollo en los métodos de investigación *online*. En este sentido, Christine Hine (2005: 6) señala:

> "The air of anxiety and innovation around online methods, and the sheer burgeoning of literature on the topic might be taken, in the terms of the sociology of scientific knowledge, as pointing to the birth of a new research network around a freshly identified focus, in this case CMC [Computer Mediated Communication]. As a new research area gathers momentum, one would expect a proliferation of publications proposing methodological approaches to the new problem. We might therefore simply be witnessing, if not a period of Kuhnian revolution (Kuhn 1970 [1962]), then at least the emergence of a new problem area around which researchers can orient their interests (Mulkay, Gilbert and Woolgar 1975). The 'pioneering' ethos may render online research in general, and online methods in particular, a fruitful field to enter: rewards can be high for being among the first to enter uncharted territory. There is no doubt that the perceived newness of the new technologies is a powerful resource in stimulating the development of the field. However sceptical we might be of the hype sorrounding the new technologies, their self-evident newness and the radical potential which is proposed for them provides a powerful resource to which researchers can hitch their own research agendas."

Un característica clave de la ciencia en red, en relación con cualquiera de los términos anteriormente señalados es la interdisciplinaridad. Nentwich (2003: 447) indica que Internet ejerce un impacto positivo en este sentido: "interdisciplinary work may become more likely since it is both easier to get in contact with people

interested in the same subject area but looking at the issues from another disciplinary angle and to access the academic knowledge of other fields". Wildman (1998) señala al respecto que la interdisciplinariedad, como resultado de las interacciones en red, supone también un cambio en la concepción del conocimiento, que deja de entenderse como algo principalmente sustantivo, frente a una concepción más relacional fruto de una negociación entre diferentes perspectivas. Wildman (1998: 628; citado por Nentwich, 2003) afirma: "Here meaning is less facts and figures locked within their respective discipline boxes and more nodes in networks of real time web interaction. Consequently meaning is not objective, universal and fixed rather it is intrajective, provisional and partial."

El mundo es demasiado complejo para intentar explicarlo desde un único punto de vista. La Web e Internet son campos abocados a la interdisciplinariedad, ya que son muchos los elementos en común que las diversas áreas de investigación deben afrontar, entre ellos, la existencia de bases de datos masivas, la visualización de la información o la simulación de procesos. Baker (2009) se pregunta acerca de qué tienen en común diferentes áreas como el marketing, la informática, la medicina o la física, para llegar a la conclusión de que básicamente lo que comparten es que los datos que manejan pueden reducirse a ceros y unos. Baker (2009: 58) se refiere a esta información en los siguientes términos:

> "It all travels through the same networks and vies for space in the same computers. This means that the mathematical tools used to analyze this data can cross disciplines and industries, from the barnyard to the aisles of Saks, almost effortlessly. This has a nearly miraculous multiplier effect – the brains working in one industry can power breakthroughs in many others Researchers long isolated in different fields, different departments on campus, different industries are now solving the same problems."

De alguna manera, todos estos investigadores están trabajando actualmente en un único laboratorio global interconectado.

2.4.2. La Web como objeto de estudio: características principales

La definición del ámbito de estudio es un episodio fundamental en el proceso de investigación (Thelwall, Vaughan y Björneborn, 2005). Nos encontramos en un terreno difuso, dinámico y en continua transformación, lo cual dificulta mucho la labor del investigador que, muchas veces, intenta en vano aprehender u observar una realidad de la que únicamente puede acceder a una pequeña parte en un instante concreto. Tim Berners-Lee (1999) no se refiere a la Web desde un punto de vista cerrado, como si se tratara de un concepto ya fijado de una vez por todas, sino que por el contrario reconoce su naturaleza abierta, compleja y cambiante. Estas características influyen decisivamente en la labor investigadora en la Web. Castells (2001: 6) afirma en relación con la investigación en Internet que:

> "It does not exhaust the sources of available information because research cannot be completed when the object of the research (the Internet) develops and changes much faster than the subject (this researcher-or, for that matter, any researcher)."

La definición de Web, así como la de sus principales componentes, no constituye únicamente la descripción de un espacio social donde se desarrollan fenómenos, sino que implica inevitablemente la especificación técnica de los elementos que delimitan el alcance objetivo de las investigaciones científicas que se realizan en la Web y sobre la Web. Algunas definiciones básicamente describen la Web como todo aquello a lo que se puede acceder a través de un navegador. Este punto de vista, sin embargo, puede presentar algunos problemas dada la versatilidad de los navegadores actuales, que se han convertido en la interfaz de referencia en el ordenador prácticamente para cualquier actividad. Son cada vez más los servicios de Internet que son accesibles desde la Web (correo electrónico, IRC, FTP, etc.). Es por ello que la Web e Internet han llegado a superponerse en muchos casos, siendo indistinguibles para una gran mayoría. Otro ejemplo de definición es el de Thelwall (2004: 17-18), que se basa en los mecanismos de acceso a una página web:

- "Requests made using the official 'port number' of the Web, 80, or

- requests made using the official computer request language of the Web, the HyperText Transfer Protocol (HTTP, as seen at the start of many URLs), or
- requests made using any mechanism available to a modern Web browser, including common non-web protocols such as FTP (File Transfer Protocol)."

2.4.2.1. Niveles de análisis en la Web: definición del objeto de estudio

Nuestra labor aquí consiste en identificar claramente los niveles de análisis de la Web a los que nos podemos referir a la hora de diseñar una investigación. Como señalan Thelwall, Vaughan y Björneborn (2005: 90), los investigadores deben precisar qué definición de la Web y qué nivel de análisis están abordando, dado que de esta elección se pueden derivar importantes consecuencias.

A continuación vamos a definir una serie de conceptos que se emplearán como referencia a lo largo de nuestra investigación. Para su definición se ha empleado una fuente *online* de reconocido prestigio en este ámbito tecnológico, la Webopedia (http://www.webopedia.com). En primer lugar nos referiremos al propio concepto de *Web*, denominada de forma completa *World Wide We*b o, por sus siglas, WWW (http://www.webopedia.com/TERM/W/World_Wide_Web.html, consultada el 28 de enero de 2010). La WWW es un sistema de servidores de Internet que soportan documentos con un tipo de formato especial basado en el lenguaje HTML (*HyperText Markup Language*). El lenguaje HTML permite la creación de hipertextos mediante el establecimiento de hiperenlaces a otros documentos u a otras partes dentro de un mismo documento. También admite contenidos multimedia, tales como gráficos, vídeo o sonido. Para poder acceder a la Web es necesario disponer de un navegador que permita interpretar el código y navegar por las distintas páginas. Como ya se apuntó anteriormente, la Web es algo distinto a Internet. Se puede decir que Internet es una red de ordenadores interconectados físicamente, mientras que la Web es un conjunto de documentos conectados por hiperenlaces. La Web funciona sobre la plataforma de Internet. Sobre Internet también funcionan otros

protocolos.

Otro concepto vinculado es el de *servidor web* (*Web server*; http://www.webopedia.com/TERM/W/Web_server.html, consultada el 28 de enero de 2010), que hace referencia a un ordenador que sirve páginas web a partir de consultas realizadas por los usuarios a través de los navegadores. Cada servidor web dispone de una dirección IP (ver más abajo) y generalmente también de un nombre de dominio. En el momento de introducir una dirección web (técnicamente, URL) en el navegador, se envía una petición al servidor para recuperar la página específica que es requerida. Cualquier ordenador puede convertirse en un servidor web instalando el *software* adecuado, por ejemplo, uno de los más populares es Apache.

IP son las siglas de *Internet Protocol* (http://www.webopedia.com/TERM/I/IP.html, consultada el 28 de enero de 2010). Se trata de un protocolo que especifica el formato de los paquetes de información que se transmiten. Por lo general, el protocolo IP se usa en conjunción con el protocolo TCP (*Transmission Control Protocol*), que establece una conexión virtual entre el destino de una información y su fuente. El protocolo *TCP* se emplea para conectar dos ordenadores de una red, permitiendo el envío de un flujo de datos y garantizando que los datos serán entregados en su destino sin errores y en el mismo orden en que se transmitieron.

Un *dominio* (*Domain name*; http://www.webopedia.com/TERM/d/domain.html, consultada el 28 de enero de 2010) funciona como un sistema para identificar una dirección IP a través de un código alfanumérico que permite acceder de una manera más sencilla a páginas y sitios web deseados. La Web está basada en códigos IP numéricos y no en nombres de dominio, lo cual obliga a que cada servidor web cuente con un *servidor DNS* (*Domain Name System*) que permita traducir los nombres de dominio en direcciones IP. Los nombres de dominio facilitan el uso de la Web por parte de los humanos, así como permite solucionar problemas tales como el traslado de contenidos de un servidor web a otro, con direcciones IP distintas, sin que el usuario perciba ningún cambio. Por ejemplo, la dirección IP de Google es http://74.125.45.100, mientras que gracias al sistema de dominios sólo

tenemos que teclear http://google.com.

Dos conceptos fundamentales, dada su amplia utilización en la investigación webmétrica, son el de página y sitio web. Una *página web* (W*eb page*) es cada uno de los documentos incluidos en la Web. Cada una de ellas está identificada por una URL específica. *URL* (http://www.webopedia.com/TERM/U/URL.html, consultada el 28 de enero de 2010) son las siglas de *Uniform Resource Locators*, que significa "localizador uniforme de recursos". Una URL es la dirección de un documento o cualquier otro recurso incluido en la Web. Mientras que un nombre de dominio es www.ugr.es, una URL sería http://www.ugr.es/pages/perfiles/estudiantes. Por su parte un *sitio web* (W*eb site*; http://www.webopedia.com/TERM/W/Web_site.html, consultada el 28 de enero de 2010) hace referencia a un sitio en la web compuesto, en principio, por un conjunto de páginas web, siendo la página de inicio (*home page*) la puerta de entrada al resto de contenidos. Por lo general un sitio pertenece y es gestionado por un particular u organización. Como se verá en los trabajos empíricos incluidos en el Capítulo 4, el sitio web es tomado como unidad de referencia a la hora de abordar la presencia de una empresa en la Web, medida a través del número de enlaces que recibe.

Un *hiperenlace* (*hyperlink*, también llamado enlace, vínculo, o hipervínculo; http://www.webopedia.com/TERM/h/hyperlink.html, consultada el 28 de enero de 2010) es un elemento de un documento electrónico que hace referencia a otro documento o a un apartado específico del mismo documento. Se trata de un componente fundamental de la arquitectura de la Web y de cualquier sistema de hipertexto. Su empleo se ha extendido también a otros ámbitos. Por lo general, emplearemos el término "enlace".

Las siglas *HTTP* hacen referencia a *HyperText Transfer Protocol* (http://www.webopedia.com/TERM/H/HTTP.html, consultada el 28 de enero de 2010). Indican al navegador, mediante la inclusión de la segmento *http://* al inicio de una URL, el protocolo de Internet que ha de emplear para interpretar la información solicitada. HTTP es el protocolo de transferencia de hipertexto que se

emplea en cada transacción en la Web. *HTML* (*HyperText Markup Language*) es el lenguaje de programación original de la Web. Presenta algunas limitaciones que han sido superadas mediante el empleo de otras tecnologías como ActiveX, Java, JavaScript y *cookies*.

Si tomamos como ilustración la siguiente URL, http://www.ugr.es, correspondiente a la página de inicio de la Universidad de Granada, podemos identificar los siguientes componentes:

- *http://*, indica que se ha de emplear el protocolo HTTP para interpretar la página.
- *www*, se refiere al nombre del servidor.
- *ugr*, se refiere al nombre principal del sitio web.
- *.es*, hace referencia en este caso a un dominio de nivel superior geográfico o de código de país. Proporciona información sobre el origen del dominio, si bien la gestión del mismo depende de cada país y en ocasiones no proporciona información geográfica relevante. Otro tipo de dominios de nivel superior son los genéricos, tales como .com, .edu, .gov, .net, o .org, que en principio deberían proporcionar información sobre el contenido de la página, por ejemplo, si es de tipo comercial, educativo, etc. La imposibilidad de ejercer un control efectivo sobre el uso de estos dominios de nivel superior hace que en muchos casos su empleo no aporte información de utilidad. En este sentido, algunos países emplean un segundo nivel en sus dominios de nivel superior. Es el caso del Reino Unido, donde los dominios incorporan componentes para indicar su uso, por ejemplo *ac* para sitios académicos (por ejemplo, www.ox.ac.uk para la University of Oxford) o *co* para sitios comerciales (www.amazon.co.uk para la empresa Amazon).

De cara al diseño de una investigación, la Web se puede estructurar en distintos niveles o capas, los cuáles van a definir el alcance de la investigación que se

realice. El estudio de un mismo fenómeno arrojará resultados diversos en función del nivel al que se realice el examen. Si investigamos la estructura de la Web, es preciso elegir el nivel al cuál se va a realizar la investigación con el fin de poder determinar qué tipo de enlaces son los que se tienen en cuenta. Algunos de los niveles sobre los que puede llevarse a cabo el análisis son los siguientes: páginas web, sitios web o dominios de nivel superior.

La determinación de la unidad de medida relevante que se va a tener en consideración a la hora de llevar a cabo el análisis es una decisión muy importante. Un mismo documento, por ejemplo, puede presentarse dividido en varias páginas o en una sola. Con el fin de abordar este problema, Thelwall (2002a) propone como solución los *Alternative Document Models* (ADM), que permiten agregar páginas en niveles superiores, tratándolas como si fueran una unidad.

2.4.2.2. ***Las características de la Web***

La Web constituye un inmenso conjunto de información, de naturaleza muy diversa y no organizada sistemáticamente, que se genera por las aportaciones singulares de individuos y organizaciones, creando un entorno relacional de gran complejidad y dinamismo. Sus efectos pueden alcanzar potencialmente a todos los ámbitos de la actividad humana; si bien, son especialmente significativas sus implicaciones en las relaciones sociales y económicas, propias de una sociedad cada vez más intensiva en conocimiento. Björneborn e Ingwersen (2001) señalan que la naturaleza de la Web es distribuida, diversa y dinámica. En tanto que distribuida, la Web se construye a través de un proceso no centralizado ni supervisado en el que cualquier agente genera contenidos que trascienden potencialmente al ámbito global de la red. La Web es dinámica ya que se encuentra en continuo cambio y transformación, creciendo progresivamente en tamaño.

Björneborn (2004) define la Web como "a new type of information system without central control, without centrally coordinated acquisition and indexing of contents".

Algunos autores han comparado la Web con un ecosistema del conocimiento. Así, en palabras de Huberman et al. (1998: 97; citado por Björneborn, 2004: 2): "... the sheer reach and structural complexity of the Web makes it an ecology of knowledge, with relationships, information 'food' chains, and dynamic interactions that could soon become as rich, if not richer, than many natural ecosystems".

A lo largo de las siguientes páginas y en línea de los resultados obtenidos en nuestra investigación empírica, podemos anticipar que la falta de un control centralizado de la información, la ausencia de mecanismos de coordinación o la no existencia de filtros que controlen la calidad de los contenidos, no implica que la Web constituya un espacio caótico y sin ningún tipo de estructura. Los análisis llevados a cabo demuestran que, a partir de la agregación de aportaciones individuales e independientes, se pueden observar patrones generales significativos. Por ejemplo, al estudiar la estructura global de la Web a partir de los hiperenlaces, se observa la existencia de *clusters,* que funcionan como nodos centrales en el sistema, o la existencia de enlaces, que conectan nodos lejanos entre sí, ejerciendo una importante labor en la generación y transferencia del conocimiento. La visualización de estructuras a partir de datos de la Web en su conjunto permite extraer información para entender mejor y actuar consecuentemente en este medio. De este modo, se pueden observar fenómenos que en el mundo físico resultarían imposibles de visualizar. De alguna manera, la Web constituye un lienzo donde toda actividad humana va dejando su rastro; de modo que, con las herramientas apropiadas, podemos analizar elementos que antes resultaban inaccesibles u obligaban a realizar mediaciones que podían condicionar los resultados observados. Señalaba Jorge Luis Borges (1944/2005c: 490) en su cuento *Funes el Memorioso*: "Pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos". La investigación en la Web, así como en otros campos, busca precisamente obviar lo accesorio para entender lo fundamental.

La Web constituye un reto para la ciencia en muchos aspectos. Sin duda uno de los más significativos es el que sitúa a la Web como objeto de investigación en sí misma. En los apartados siguientes pretendemos abordar cuáles son las

características que hacen de la Web un desafío.

2.4.2.3. El tamaño de la Web

La Web se encuentra en continuo cambio. Se trata de un entorno muy dinámico. Esto hace que sea prácticamente imposible obtener una medición exacta, o siquiera aproximada, de su tamaño en un momento preciso. Se han llevado a cabo distintos intentos de medición a lo largo de los últimos veinte años. La mayoría de ellos son anteriores al surgimiento de la Web 2.0 y al consiguiente *boom* social de Internet en el último lustro.

Un trabajo clásico de Lawrence y Giles (1998) estimó, hace más de diez años, que el tamaño de la Web indexable por los motores de búsqueda era de aproximadamente 320 millones de páginas. Sin embargo, sólo un año después, los propios autores (Lawrence y Giles, 1999a) elevaban ese número a 800 millones de páginas. Moore y Murray (2000) estimaron, a fecha julio de 2000, que había al menos 2.100 millones de páginas indexables, con un ritmo de crecimiento de 7 millones de páginas diarias. El surgimiento de la Web 2.0 y el incremento en el número de usuarios conectados, tanto en los países más desarrollados como en los países en vías de desarrollo, indica que en los últimos diez años el ritmo de crecimiento ha debido incrementarse considerablemente. Gulli y Signorini (2005) estiman, de forma coherente con el ritmo de crecimiento arrojado por Moore y Murray, que, en la fecha de su estudio, existían aproximadamente 11.500 millones de páginas web indexables.

Probablemente una de las mejores fuentes de información sobre este tema son los datos que arrojan los buscadores comerciales. El sitio Search Engine Showdown (http://searchengineshowdown.com) presentaba de forma regular actualizaciones basadas en estadísticas de buscadores relativas, por ejemplo, al tamaño relativo o a solapamientos en las bases de datos. En cualquier caso, a julio de 2008, las últimas estadísticas encontradas relativas al tamaño de la Web indexada por los motores de búsqueda se remontan a finales de 2002. La eficacia de los motores de

búsqueda como indicador del tamaño de la Web se basa especialmente en su capacidad para indexar potencialmente toda la información disponible. Este ambicioso objetivo cuenta con algunas serias limitaciones. Con todo, Google, el principal motor de búsqueda global, tiene como misión fundamental la de organizar la información del mundo. A mediados de 2008, Google (2008) informó que su base de datos había sobrepasado el billón de URLs únicas indexadas. Estos datos contrastan con los 26 millones que componían la base de datos de Google en 1998, cuando apareció el servicio, o con los mil millones en el año 2000. Las cifras ofrecidas son el resultado de eliminar páginas duplicadas o contenidos generados automáticamente. Si únicamente tuviéramos en cuenta las páginas web que se generan automáticamente a petición del usuario, la cifra sería infinita. Podemos imaginar, por ejemplo, un calendario en el que se incluye un enlace al mes siguiente, de manera indefinida. Un motor de búsqueda creado recientemente, Cuil (2010), declara tener indexadas, a 30 de marzo de 2010, más de 127.000 millones de páginas web.

Cabe mencionar otro proyecto de medición de la Web, si bien actualmente se encuentra muy desactualizado. Se trata del proyecto *Web Characterization,* llevado a cabo por la *Office of Research of the Online Computer Library Center, Inc.* (OCLC; en http://www.oclc.org/research/projects/archive/wcp/), que recogía cada año una muestra aleatoria de direcciones IP con el fin de analizar tendencias en relación al tamaño y al contenido de la Web. La última actualización data de abril de 2003.

En todo caso, cabe subrayar que estas mediciones se refieren únicamente a lo que se conoce como Web indexable. En la Web existen muchas zonas oscuras con datos que no son accesibles públicamente o por las herramientas empleadas para indexarla. Esto ha hecho surgir el concepto de Web invisible, profunda u oculta.

2.4.2.4. La Web invisible

El concepto de Web invisible hace referencia a aquellas partes de la Web que no son accesibles para los motores de búsqueda (Sherman y Price, 2001). Otros

términos vinculados son los de Web profunda (*deep Web*) o Web oculta (*hidden Web*), si bien generalmente se emplean en relación con páginas que generan contenido de manera automática, a partir de algún tipo de consulta efectuada por el usuario (Shestakov, 2008). Completando el panorama, Lawrence y Giles (1998) introdujeron también el concepto de Web indexable (*indexable web*), que hace referencia a la parte de la Web que puede ser indexada por los motores de búsqueda, excluyendo entre otros documentos, bases de datos web que requieren la realización de algún tipo de consulta, por ejemplo, un diccionario *online* en el cual es preciso introducir una palabra para acceder a su contenido.

Las páginas creadas de forma automática, como se ha apuntado anteriormente, generan problemas a los buscadores ya que pueden dar lugar a un número infinito de páginas web, que no añaden información relevante a las bases de datos de los motores de búsqueda. Como ejemplo, el número de páginas web que Google puede generar a partir de la introducción de palabras de consulta en su buscador es potencialmente infinito.

Bergman (2001) estimó que la parte oculta de la Web era 500 veces más grande que la Web indexable. En cualquier caso, este tipo de estimaciones se encuentran actualmente muy desactualizadas y de volver a realizarse serían difícilmente practicables debido al fenómeno de las páginas web generadas automáticamente. Actualmente, a todo el contenido oculto, cuya naturaleza de algún modo ya conocíamos, habría que sumar fenómenos tan importantes como el de las redes sociales, cuya información alcanza volúmenes muy elevados y no es pública.

Holmberg (2009) apunta la existencia de cuatro tipos de invisibilidad en la Web:

1. La Web opaca, que contiene páginas y archivos que, si bien podrían ser indexados por los motores de búsqueda, no lo son;

2. La Web privada, que incluye contenidos que de forma deliberada se excluyen para que no sean indexados por los motores de búsqueda;

3. La Web propietaria, que está compuesta de contenidos cuyo acceso está restringido, por ejemplo, mediante el empleo de contraseñas; y,

4. La Web verdaderamente invisible, que incluye páginas que por razones técnicas no pueden ser indexadas por los motores de búsqueda.

El hecho de que los autores de determinadas páginas decidan la exclusión de las mismas de los índices de los buscadores, por ejemplo, mediante el empleo de los ficheros *robot.txt*, pone de relieve problemas de índole ética, ya que técnicamente se podría acceder a dicha información, sin respetar la decisión de sus autores.

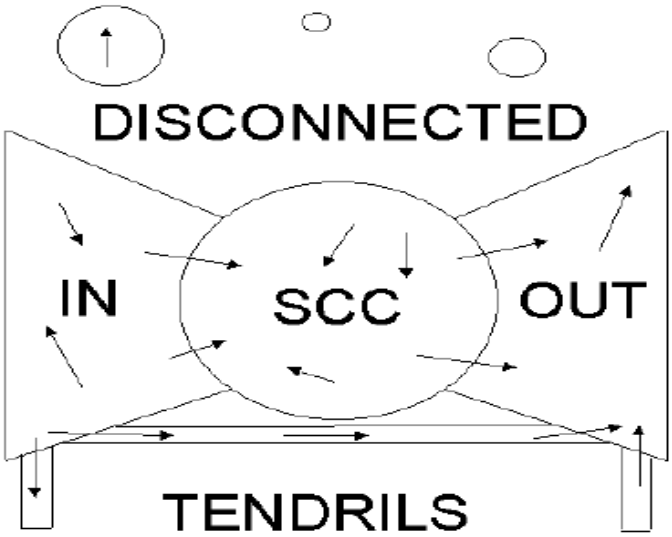
2.4.2.5. Estructura de la Web

Centrándonos en la cuestión de cómo es la estructura de la Web, algunas investigaciones han tratado de dibujar la red de hiperenlaces que conectan los distintos nodos (páginas web, sitios web, etc.) existentes en la red. Son trabajos fundamentalmente descriptivos, entre los que destacan el llevado a cabo por Broder et al. (2000) y el de Björneborn (2004).

Albert, Jeong y Barabási (1999) y Barabási, Albert y Jeong (2000) midieron el diámetro de la Web, encontrando que la distancia media entre dos documentos web seleccionados de forma aleatoria era de 19 enlaces o conexiones. Este experimento se relaciona de alguna manera al llevado a cabo por Milgram (1967), dando lugar a la teoría de los seis grados de separación. Dado que los trabajos comentados se remontan a hace diez años, es probable que el crecimiento de la Web y de las conexiones entre las páginas hayan reducido esa distancia.

Un equipo de investigadores (Broder et al., 2000), trabajando para el buscador AltaVista, emplearon una araña para construir un diagrama de la estructura de enlaces de la Web. El diagrama mostraba la forma de una pajarita (*bow tie model*) como el que aparece en la Figura 2.1 (tomado de Thelwall, 2008c: 613).

Figura 2.1. Modelo de pajarita de la Web, por Broder et al. (2000)



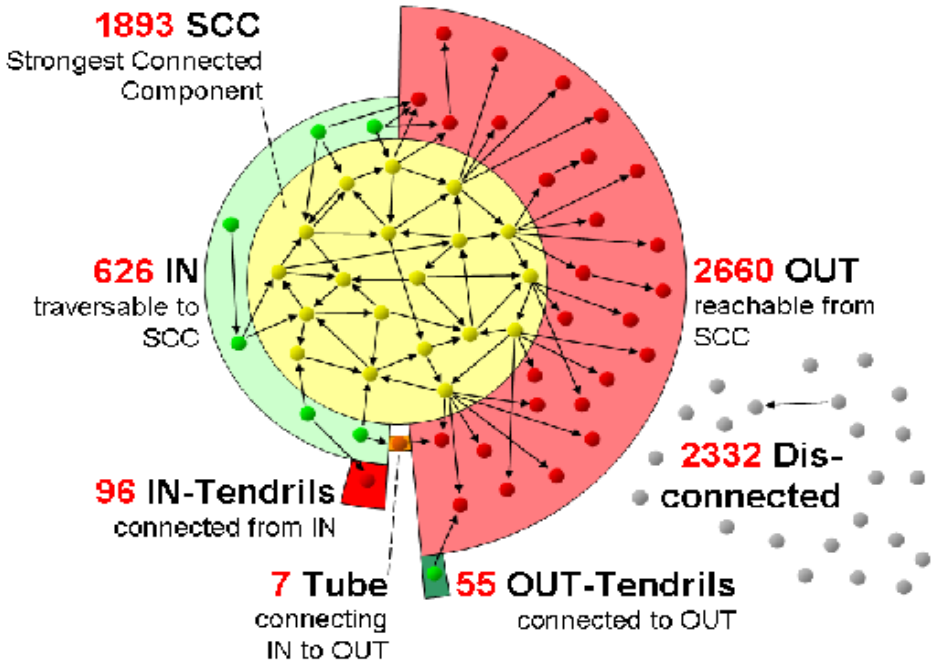
Dicho modelo con forma de pajarita muestra una parte central SCC, "*Strongly Connected Component*" donde aparecen aproximadamente el 28% de páginas de la Web, que son accesibles a través de uno o más enlaces. Esta parte constituye el corazón de la Web siendo fácil de navegar. En la zona OUT se concentran páginas web (21%) a las que se puede acceder desde la zona SCC pero que no contienen ningún enlace exterior. Por el contrario la parte IN incluye otro 21% de páginas web que enlazan a páginas del SCC pero que no son enlazadas por aquellas incluidas en el núcleo, por lo que de alguna manera permanecen aisladas. Existen también un grupo de páginas que no enlazan ni son enlazadas. Representan un 8% del total y aparecen representadas por el título DISCONNECTED. Por último hay algunas páginas (22%, con el nombre de TENDRILS) con un comportamiento más heterogéneo a la hora de enlazar, por ejemplo, conectando el componente IN con el componente OUT.

Las cuatro partes tienen aproximadamente el mismo tamaño, lo cual indica que existen grandes partes de la Web que no son fácilmente accesibles por los motores de búsqueda, ya que o bien son páginas que se encuentran aisladas o bien únicamente establecen enlaces, pero no los reciben. El núcleo de la Web, el SCC,

se encuentra fuertemente interconectado, lo cual hace más viable las conclusiones del trabajo de Barabási, Albert y Jeong (2000) previamente comentado. Sin embargo, no parece que las mismas sean válidas para esas otras partes de la Web que se encuentran más aisladas o con difícil acceso. Este modelo ha sido empleado, entre otros, por Thelwall y Wilkinson (2003) y por Baeza-Yates, Castillo y López (2005).

Por otra parte, Björneborn (2004), en su tesis doctoral, retoma el modelo de pajarita y lo transforma en un modelo con forma de corona, con el objeto de enfatizar la centralidad del componente SCC y la conexión cercana entre SCC e IN y SCC y OUT. Su investigación se centra en el ámbito académico de la Web en el Reino Unido. La Figura 2.2 muestra el modelo de corona.

Figura 2.2. Modelo de corona de la Web, por Björneborn (2004: 80)



2.4.2.6. La naturaleza dinámica de la Web

La Web está en continuo cambio. Igual que Heráclito de Éfeso afirmaba que uno nunca se baña dos veces en el mismo río, nosotros podemos decir que nunca visitamos la misma Web dos veces. Si bien los cambios a niveles concretos pueden no ser tan evidentes o tan rápidos, la velocidad y magnitud de los mismos en el conjunto del sistema hace que sea inviable para ningún sistema indexar la Web de manera íntegra. Se trata en vano de atrapar la naturaleza efímera de la Web. Cohabita esta característica, sin embargo, con el potencial que tiene la Web para almacenar contenidos como si se tratara de una gran biblioteca o archivo en el que los problemas de espacio derivados del almacenamiento de los documentos hubieran dejado de ser uno de los problemas más significativos. La dualidad entre lo efímero y lo permanente ha hecho surgir, por ejemplo, la reivindicación del derecho a olvidar en relación a la información personal que incluimos en la red. Las dualidades y contrastes no acaban aquí: en el apartado anterior hemos visto que frente al aparente caos, surge una estructura invisible que constituye el orden oculto de la Web.

Desde su creación, la Web ha experimentado un incremento continuado en número de páginas, dominios, servidores, etc. Se trata de un crecimiento que se produce como resultado de un intenso proceso de creación de nuevos documentos, que contrapesa la desaparición de muchos otros. Si bien la perdurabilidad de los contenidos en la Web podría ser ilimitada, sobre todo gracias al abaratamiento de los costes del *hardware*, en ocasiones los documentos se eliminan. El crecimiento en el número de hiperenlaces a lo largo de los años es también especialmente significativo, ya que son el elemento fundamental empleado para analizar la estructura de la Web.

Al margen del crecimiento de la Web, las páginas y sitios existentes están sujetos a continuas modificaciones. Por ello es tan importante señalar, cuando se hace referencia a algún contenido de la Web, el momento en el que se ha accedido al mismo, ya que es posible que no encontremos dicha versión posteriormente o incluso que el documento haya desaparecido. Hay sistemas de gestión como el de

los wikis, en el que se basa la Wikipedia, que permite conseguirlo y acceder a versiones anteriores de documentos. Sin embargo, esta posibilidad es una excepción ya que la mayoría de páginas web no funcionan así. Es también relevante en este sentido la iniciativa del Internet Archive, llamada Wayback Machine (http://www.archive.org), la cual almacena copias periódicas de las páginas web, manteniendo una especie de memoria arqueológica. Revela esto un rasgo fundamental del tipo de investigación que se puede efectuar en la Web: sólo se pueden realizar análisis desde la perspectiva del presente, salvo que de algún modo accedamos a bases de datos históricas de algún tipo de servicio. En todo caso, si el estudio se basa en la cuantificación de enlaces y en cómo los enlaces interconectan las páginas y sitios web, la información únicamente se puede conseguir en el momento actual, nunca referente a momentos anteriores.

Otros cambios que se pueden producir en la Web hacen referencia a la modificación en la localización de los documentos; por ejemplo, una página web puede ser exportada a un nuevo dominio o a una nueva URL sin dejar un rastro tras ella que permita localizarla. No se trata de un tema baladí, ya que la localización del conocimiento es un factor clave en la Web. Imaginemos por un momento que la estructura de enlaces que tejen la red desapareciera. Nos quedaríamos prácticamente aislados, habiendo perdido los vínculos hacia los contenidos relevantes. Una de las características técnicas que ha extendido el desarrollo de la Web 2.0 es el empleo de enlaces permanentes (*permalinks*) para los contenidos. Por ejemplo, en un blog, cada uno de los nuevos artículos que se publican tienen su URL única y permanente, lo cual facilita tanto la localización del conocimiento por parte de otros usuarios que pueden enlazar dicho contenido como la indexación de esa información por las arañas de los motores de búsqueda.

En última instancia, el dinamismo de la Web está estrechamente vinculado con los cambios en las necesidades y el comportamiento de los usuarios en Internet. Se puede observar fácilmente si prestamos atención a la evolución en el empleo de herramientas de la Web social a lo largo de los últimos años. Chat, blogs, redes sociales, etc. han vivido momentos de éxito. Diferentes estudios académicos han abordado también estas cuestiones vinculadas al comportamiento de los usuarios.

Por ejemplo, Jansen, Spink y Pedersen (2005) estudiaron el comportamiento de búsqueda de los usuarios en los motores de búsqueda.

Diversos trabajos de investigación han incidido en la naturaleza dinámica de la Web (por ejemplo, Bar-Ilan y Peritz, 1999; 2009; Tan, Foo y Hui, 2001; Koehler, 2002; 2004; Ortega, Aguillo y Prieto, 2006).

Estas características condicionan qué información puede obtenerse de la Web y cómo. También determinan qué puede saberse sobre la Web. El dinamismo que hemos puesto de manifiesto nos debe hacer conscientes de que cualquier aproximación a esta realidad es únicamente el análisis de una imagen, más o menos completa, en un momento determinado y en una situación particular.

2.4.2.7. La Web como base de datos distribuida y no estructurada

Björneborn e Ingwersen (2001: 69) subrayan que la realización, en determinada medida, del sueño de la libertad de información a través de una comunicación no sujeta a control y sin coste o a muy bajo coste a través de la Web, ha provocado también un entorno en el que la calidad y fiabilidad de la información se ha puesto en cuestión. En un entorno tan cambiante es complicado hacer previsiones o diagnósticos que perduren en el tiempo. Sin embargo, la diversidad de la Web puede constituir una fértil fuente de información para llevar a cabo hallazgos de variada índole.

Björneborn e Ingwersen (2001) describen la Web como una base de datos distribuida que crece de forma exponencial. Esta idea de la Web como gran base de datos con unas características únicas está en la base del desarrollo de técnicas para la obtención de conocimiento a partir de la explotación de los grandes volúmenes de información existentes. La aplicación de las técnicas de minería de datos a la Web permite aprovechar y sacar el máximo partido a formas de análisis que se han venido aplicando hasta ahora a grandes bases de datos, con el objetivo de identificar patrones de comportamiento que permitieran desvelar información

relevante para la gestión de la empresa. La Web se considera de este modo como una inmensa base de datos no estructurada y de naturaleza heterogénea. Para algunos investigadores, la minería de datos se considera parte integrante de un mecanismo más amplio que es el *Knowledge Discovery Database* (KDD, en adelante) (Björneborn y Ingwersen, 2001). El KDD surge de la combinación del poder de computación de los ordenadores y de la experiencia y la intuición humanas. Entre otros elementos, se conforma como una combinación de técnicas que incluyen: la recuperación de la información, la estadística, el reconocimiento de patrones y la visualización de la información. Mientras que la minería de datos se encarga del reconocimiento de patrones en grandes bases de datos, el KDD comprende el proceso global de generación de conocimiento relevante. Vickery (1997) apunta alguna de las áreas donde el KDD se ha aplicado con mayor éxito (por ejemplo, al análisis del comportamiento del consumidor, el diagnóstico del cáncer, la identificación de estructuras químicas, el análisis de poblaciones, el control de la calidad, etc.). En el ámbito de empresa, la minería de datos en la Web plantea soluciones al problema de saturación de información que afrontan las empresas, lo cual dificulta su capacidad para identificar aquellos datos que pueden ser útiles para su toma de decisiones (Choi y Varian, 2009).

El empleo de "metadatos", términos que describen los recursos incorporados en la Web, permitiendo que los ordenadores puedan procesar de manera inteligente dichos elementos, puede representar un nuevo paso para el avance de la minería de datos. Björneborn e Ingwersen (2001) señalaban el escaso uso de los mismos. En los últimos siete años, el desarrollo de la Web 2.0 y la popularización de herramientas de etiquetado han facilitado la inclusión en la red de una gran cantidad de recursos con metadatos. La difusión de los mismos avanza, en cierta medida, hacia el objetivo de la Web Semántica propuesta por Tim Berners-Lee (1999).

2.4.2.8. Los mundos pequeños en la Web

La Web ha sido también estudiada mediante la teoría de grafos (*Graph Theory*). Se trata de una aproximación que se conoce como "*topológica"* porque trata la Web como un grafo matemático, ignorando las relaciones espaciales y los contenidos de las páginas. Un grafo es una representación matemática de una red que consiste en vértices o nodos conectados por enlaces. Los grafos pueden ser dirigidos o no. Esto depende de si la dirección de la conexión es relevante o no para la investigación.

La teoría del "mundo pequeño" o de los "mundos pequeños" se basa en el análisis de redes sociales acerca de las cortas distancias que, a través de una cadena de conocidos, separan a dos personas seleccionadas aleatoriamente. En un estudio ya clásico, Milgram (1967) pidió a una muestra aleatoria de personas en Omaha (Nebraska) que hicieran llegar una carta a una persona desconocida para ellos situada en la costa este de Estados Unidos. La carta debía ser enviada a algún conocido que continuaría la cadena hasta alcanzar a la persona señalada. Para aquellas cartas que alcanzaron su objetivo, hubo una media de cinco intermediarios, dando lugar a los, popularmente conocidos, "seis grados de separación" que existen entre dos personas en el mundo a través de una cadena de conocidos. Kochen (Pool y Kochen, 1978/1979; Kochen, 1989) profundizó en la teoría del "mundo pequeño". En ocasiones nos referimos a "mundos pequeños", en plural, debido a que se ha constatado, al menos en la Web, la existencia de diversos conjuntos de nodos bien comunicados, pero al mismo tiempo alejados de otros grupos de nodos, formando diversos mundos pequeños.

Los mundos pequeños representan un equilibrio entre orden y azar. Los gráficos que los representan combinan características de gráficos trazados al azar y gráficos ordenados, en los que las conexiones entre nodos siguen determinados patrones. En los gráficos de mundos pequeños existen conexiones transversales que permiten unir partes diversas y distantes de la red, reduciendo considerablemente las distancias existentes. Este tipo de enlaces han sido puestos en relación con el concepto de relaciones débiles expuesto por Granovetter (1973). Las relaciones

fuertes se producen entre los amigos o contactos que pertenecen a nuestro mismo grupo social y, por tanto, al compartir el mismo entorno, no proporcionan nueva información al sistema. Esta estructura permite la existencia de *clusters* muy enlazados, pero también la existencia de enlaces entre *clusters* establecidos como por azar. El fenómeno de los mundos pequeños se ha detectado en una amplia variedad de redes; por ejemplo, sociales, neuronales, biológicas o incluso en la estructura de las redes eléctricas (Watts y Strogatz, 1998; Albert, Jeong y Barabási, 1999; Watts, 1999; 2003; Newman, 2000; Barabási, 2002; Buchanan, 2002; Strogatz, 2003). Diversos trabajos han detectado también la existencia de mundos pequeños en Internet (Adamic, 1999; Albert, Jeong y Barabási, 1999; Jin y Bestravros, 2002), así como en la propia evolución de la Web (Watts y Strogatz, 1998; Albert, Jeong y Barabási, 1999).

Tim Berners-Lee, en su libro *Tejiendo la red* (1999), reflexiona sobre la estructura de Internet y la Web, apuntando que en un comienzo Internet mostraba una estructura técnica y social descentralizada en la que cada nodo era igual que otro, sin ningún tipo de orden o estructura preestablecida. Sin embargo, con la creación y desarrollo de la Web esta situación cambia ya que, en sus propias palabras (1999: 189), "sin jerarquía había demasiados grados de separación para evitar que las cosas se reinventaran". Así pues, parece que esa combinación entre una estructura ordenada y al azar hace que se mantenga un equilibrio que, en su opinión, evita caer en un igualitarismo absoluto, que situaría la información relevante a demasiados grados de separación para acceder a ella de manera eficiente, pero también en el predominio de unos pocos nodos.

Björneborn y Ingwersen (2001) apuntan que en redes con características de mundo pequeño, los nodos están agrupados en gráficos regulares, existiendo únicamente un pequeño número de enlaces que conectan partes alejadas de la red. Para Björneborn e Ingwersen (2001) una de las consecuencias de este fenómeno es la posibilidad de descubrir información relevante, no esperada, en la Web. Una idea central en este sentido es la de los enlaces transversales, que son aquellos enlaces que funcionan como atajos entre páginas web distantes. Una de las principales implicaciones es la posibilidad de acceder a conocimientos potencialmente valiosos

y genuinos que se derivan de conexiones no aparentes o no previstas. La idea de los enlaces transversales nos remite a las relaciones débiles en el marco social (Granovetter, 1973). Nos remite también a la idea que el *Memex* de Vannevar Bush pretendía poner en la práctica. Garfield, en el campo de la Bibliometría, ha utilizado el término de "serendipia sistemática" (*systematic serendipity*) en varias ocasiones, aplicado al proceso de descubrir relaciones científicas previamente desconocidas a través de la exploración de citaciones en bases de datos (Björneborn e Ingwersen, 2001). El término "serendipia", muy popular a la hora de referirnos a la Web, aunque no incluido en el Diccionario de la Real Academia Española, deriva del inglés *serendipity.* Este término es un neologismo acuñado por Horace Walpole, en 1754, cuyo origen se encuentra en un cuento persa del siglo XVIII llamado "Los tres príncipes de Serendip", en el que los protagonistas, unos príncipes de la isla Serendip (nombre árabe de Ceilán, la actual Sri Lanka) resolvían sus problemas gracias a increíbles casualidades. Si bien en sus orígenes fue un término muy popular, cayó en desuso y recientemente se ha vuelto a poner de actualidad referido al proceso de descubrimiento o hallazgo de conocimiento en circunstancias inesperadas o afortunadas. La idea de sistematizar la "serendipia" en el ámbito de la Web se puede adaptar perfectamente a los procesos de minería de datos, en concreto a la detección y análisis de los enlaces transversales.

Al margen de las propiedades estructurales que presenta la Web relativas a los mundos pequeños, se pueden identificar también rasgos propios de las redes de escala libre así como leyes de potencias en relación con la distribución de conexiones en la red o con el número de enlaces que reciben las páginas web. A ellos nos referimos en el apartado siguiente.

2.4.2.9. Las leyes de potencias y redes de escala libre en la Web

Muchos fenómenos de la Web pueden representarse mediante leyes de potencias. Las redes que siguen este tipo de distribución se llaman con frecuencia redes de escala libre (Barabási, Albert y Jeong, 2000). En la Web unos pocos sitios son muy

populares y reciben millones de enlaces, mientras que una inmensa mayoría de sitios reciben muy pocos enlaces. Son varios los trabajos que han detectado la existencia de fenómenos en la Web que siguen leyes de potencias (Albert, Jeong y Barabási, 1999; Barabási y Albert, 1999; Faloutsos, Faloutsos y Faloutsos, 1999; Barabási, Albert y Jeong, 2000; Broder et al., 2000; Baeza-Yates, Castillo y López, 2005). En nuestro caso, principalmente nos interesa constatar este hecho en el caso del número de enlaces que reciben los sitios Web.

Una de las causas que explica la existencia de este tipo de distribución es el modo en que navegamos por la Web y la forma en la que los motores de búsqueda funcionan basándose en la estructura de hiperenlaces para elaborar *rankings* de resultados (Brin y Page, 1998). Esto ocasiona que las páginas más enlazadas sean las que aparecen en primer lugar en dichos *rankings* y, por tanto, sean las más visitadas y las que cuenten con mayores posibilidades de volver a ser enlazadas. Se trata de un fenómeno que se retroalimenta y que se conoce como *preferential attachment*. El resultado es un efecto acumulativo que mantiene o incrementa la visibilidad de las páginas que ya son visibles. Este funcionamiento de los motores de búsqueda hace que la aplicación de algoritmos de búsqueda automáticos, sin intervención humana, no basten para generar resultados no sesgados, pudiendo cuestionar la calidad de los resultados que se ofrecen.

El *preferential attachment* puede explicar la existencia de distribuciones basadas en leyes de potencias. Algunas de las distribuciones de este tipo detectadas en la Web son, por ejemplo: el número de enlaces recibidos por un sitio web (Albert, Jeong y Barabasi, 1999; Adamic y Huberman, 2001), el número de enlaces establecidos por un sitio web (Adamic y Huberman, 2001), el número de páginas que componen un sitio web (Huberman y Adamic, 1999; Adamic y Huberman, 2001), el número de visitas a un sitio web (Huberman et al., 1998; Pitkow, 1998; Adamic y Huberman, 2001), y el número de páginas visitadas dentro de un sitio web (Huberman et al., 1998; Pitkow, 1998). Un caso de especial interés, desde el punto de vista empresarial, es el de la larga cola (Anderson, 2006), que pone de relieve la existencia de grandes oportunidades de negocio sirviendo a los nichos de mercado desatendidos que no resultaban rentables en una economía tradicional. El comercio

electrónico hace que sea rentable atender a dichos segmentos de mercado sin incurrir en costes añadidos.

Dill et al. (2001) concluyen que, debido a las leyes de potencias que se observan en la Web, cualquier parte de la misma puede ser observada como si se tratara de un sistema en miniatura que reproduce las propiedades del sistema en su conjunto. Ello hace que la Web presente una naturaleza fractal, propiedad que también sugiriera Berners-Lee con anterioridad (1999).

2.4.2.10. La Web semántica

La Web semántica (Berners-Lee y Hendler, 2001; Berners-Lee, Hendler y Lassila, 2001) permitirá que tanto los seres humanos como las máquinas puedan procesar la información y comunicarse de una manera más eficiente. Para Berners-Lee, uno de los más importantes incentivos para desarrollar la Web era la posibilidad de seguir la pista a "the complex web of relationships between people, programs, machines and ideas" (Berners-Lee, 1997). Actualmente, la mayoría de información existente en la Web está pensada y diseñada para que la puedan entender los seres humanos. Sin embargo, el volumen de datos existente es cada vez mayor y es preciso emplear ordenadores y otras máquinas para su procesamiento. Para que el tratamiento de la información sea realmente efectivo es preciso que los sistemas informáticos conozcan qué información es la que están trabajando, por ejemplo, cuál es el contenido y el significado de una fotografía. Existen dos formas de aproximarse al problema: por un lado, empleando inteligencia artificial y métodos de aprendizaje por parte de las máquinas, de modo que puedan llegar a "entender" los documentos creados por los humanos; por otro lado, añadiendo información complementaria a los contenidos que se crean, por ejemplo, en forma de etiquetas (Berners-Lee y Hendler, 2001). Algunos de los elementos actualmente desarrollados que nos acercan a la Web semántica son el XML, RDF y las ontologías (Ding et al., 2002). En la práctica, podemos citar algunos casos destacados en el campo de la información financiera; por ejemplo, en relación con el desarrollo del XBRL, lenguaje basado en XML para la divulgación de información

financiera (Bonsón, Cortijo y Escobar, 2009) o el empleo de ontologías para el intercambio de información entre los supervisores bancarios europeos (Bonsón-Ponte, Escobar-Rodríguez y Flores-Muñoz, 2009).

En 1999, Tim Berners-Lee escribía (1999: 145): "Tengo un sueño acerca del Web... y ese sueño tiene dos partes". La primera se refería a una Web convertida en un medio potente de colaboración entre las personas; la segunda a que las máquinas fueran capaces de leer todos los documentos existentes en la Web, es decir, que éstos fueran también significativos para las computadoras sin necesidad de la intervención humana directa. En palabras de Berners-Lee (1999: 145): "En la segunda parte del sueño, la colaboración se extiende a los ordenadores. Las máquinas se vuelven capaces de analizar todos los datos que hay en el Web: el contenido, los vínculos y las transacciones entre personas y ordenadores".

3. LA WEBMETRÍA

«Los ambientes son invisibles. Sus reglas fundamentales, su estructura penetrante y sus patrones generales eluden la percepción fácil.»

Marshall McLuhan

El medio es el masaje. Un inventario de efectos *(1995)*

La Webmetría tiene su origen en la aplicación de conceptos y técnicas bibliométricas al análisis de la Web. La adaptación y empleo de los conceptos y metodologías de la Bibliometría, disciplina centrada en buena medida en la investigación de la actividad científica, al estudio de la Web puede considerarse, salvando importantes diferencias, como una evolución lógica, en tanto que los elementos que componen los hipertextos en la Web encierran evidentes similitudes y paralelismos con los elementos propios de otros tipos de documentos, tales como los artículos científicos. Por su parte, las Ciencias de la Información en su conjunto y la Informetría, en tanto que tienen como objeto de análisis la información, conducen de manera lógica a abordar los ingentes volúmenes de datos disponibles en la Web.

Uno de los elementos fundamentales para el análisis webmétrico es el hiperenlace, figura análoga a la de las citaciones bibliográficas. Al igual que éstas, los enlaces son una fuente valiosa de información acerca del uso y la visibilidad de artículos científicos y de las conexiones entre autores y documentos. Pronto las analogías se

hacen más complejas ya que la Web acoge prácticamente cualquier fenómeno humano, por lo que el campo de estudio se hace más extenso, ampliándose fácilmente a temas políticos, de gobierno electrónico y corporativo, empresariales, etc. Si bien a efectos de introducir la Webmetría en el área de la investigación empresarial resulta útil subrayar las analogías entre los elementos bibliográficos y los elementos propios de la Web, es preciso reconocer que las diferencias entre ambos son también significativas, especialmente si tenemos en cuenta que ambos contextos difieren sobre todo en su grado de formalización y en la existencia o no de controles de calidad previos a la publicación de contenidos (Han y Chang, 2002).

3.1. Origen, definición y perspectivas de futuro de la Webmetría

Desde finales de los años 90 se han propuesto diversos términos para dar nombre a esta nueva disciplina. Entre ellos, podemos señalar: *netometrics* (Bossy, 1995), *internetometrics* (Almind y Ingwersen, 1996), *webometry* (Abraham, 1996) o *Web bibliometry* (Chakrabarti et al., 2002). El término inglés *Webometrics* aparece por primera vez en 1997 de la mano de Almind e Ingwersen (1997). Prácticamente al mismo tiempo surge la denominación *Cybermetrics*, dando nombre a una revista científica. Determinados términos, como *Web metrics* son empleados también en otras áreas estrechamente vinculadas, tales como ciencias de la computación (*computer science*; por ejemplo por Dhyani, Keong y Bhowmick, 2002).

Tanto los términos *Webometrics* (en español, Webmetría) como *Cybermetrics* (en español, Cibermetría) parecen haberse consolidado a lo largo de estos años en la literatura. Björneborn (2004; citado en Björneborn e Ingwersen, 2004: 1217) intentó delimitar y fijar una definición de ambos conceptos. Por un lado, define Webmetría como "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches." Por otro, la Cibermetría se define de manera análoga aunque referida a un ámbito de estudio más amplio, Internet en su

conjunto. En inglés, el término *Webometrics* toma también la forma *Webmetrics*. El equivalente en español es *Webmetría* para ambos casos, no siendo necesario hacer ningún tipo de distinción al respecto.

De las definiciones propuestas por Björneborn destacan dos elementos clave. Por un lado, la relación directa con la disciplina bibliométrica, lo cual permitirá entender tanto las metodologías como las áreas de investigación a las que se ha aplicado mayoritariamente. Por otro, la distinción entre Web (vinculada a la Webmetría) e Internet (vinculada a Cibermetría). Ya hemos visto que Web e Internet son realidades diferentes, si bien cada vez tienden más a superponerse e identificarse. Para Thelwall, Vaughan y Björneborn (2005: 81), la Webmetría es "the quantitative study of Web-related phenomena" y comprende investigaciones que van más allá de las Ciencias de la Información, vinculándose a campos como los estudios de comunicación, la física o la informática. En un trabajo posterior, Thelwall (2008c: 611), define Webmetría como "the quantitative analysis of web phenomena, drawing upon informetric methods, and typically addressing problems related to bibliometrics". Incidiendo en los vínculos con la Bibliometría, apunta que la Web dispone de sus propios índices de impacto a través de los motores de búsqueda (por ejemplo, Google y Yahoo). Entre las áreas de actuación de la disciplina, Thelwall (2008c) incluye: el análisis de enlaces, el análisis de citaciones en la Web, la evaluación de buscadores o la realización de estudios descriptivos de la Web. Más recientemente, se ha sumado también el análisis del fenómeno de la Web 2.0. Dada la variedad de contenidos existentes, algunas de estas áreas se concretan en el estudio de la ciencia, las empresas, los partidos políticos, etc.

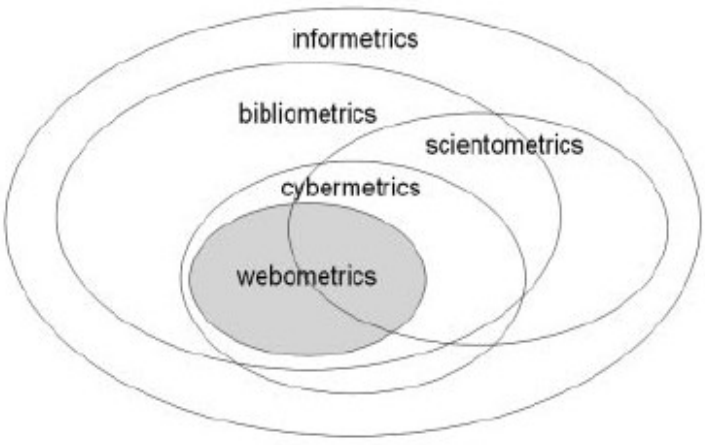
Bar-Ilan (2008: 19), por su parte, define la Webmetría a través de su relación con los siguientes ámbitos: índices de impacto web, sitios web de revistas, sitios web de universidades y recuento de enlaces. Como se puede observar, la vinculación de la disciplina a la Bibliometría sigue siendo muy poderosa, a pesar de que cada vez con mayor profusión se han llevado a cabo trabajos en áreas distintas a la del estudio de la producción científica. De acuerdo con *ISI (Thomson)´s Essential Science Indicators*, la Webmetría constituye un campo de investigación actual con un buen número de artículos clave (*core papers*) (Essential Science Indicators,

2005).

Holmberg (2009) sitúa las raíces de la Webmetría tanto en la Bibliometría como en la Cienciometría. La primera estudia los aspectos cuantitativos de la producción, distribución y uso de la información registrada (Tague-Sutcliffe, 1992), mientras que la Cienciometría hace referencia a un estudio cuantitativo de la ciencia. Actualmente los distintos campos de actuación, como se puede observar en la Figura 3.1, se superponen en gran medida. Atendiendo al significado etimológico del término, Bibliometría es la "medición de los libros", mientras que Webmetría haría referencia a la "medición de la Web", lo cual describe de manera muy clara el objeto de la disciplina.

La Figura 3.1 sitúa estas disciplinas dentro del área más general de las Ciencias de la Información. Pese a su origen, la investigación llevada a cabo en los últimos años ha subrayado su carácter profundamente abierto e interdisciplinar, proporcionando métodos de investigación que pueden ser de provecho en otras áreas de conocimiento, por ejemplo, las ciencias sociales en su conjunto o la informática.

Figura 3.1. La Webmetría y la Cibermetría en el contexto de las Ciencias de la Información (Björneborn y Ingwersen, 2004: 1217)



La Bibliometría ha prestado especial atención a los sistemas de citación y de elaboración de *rankings* de impacto en el campo de la investigación científica (Garfield, 1979). Las evidentes analogías entre artículos científicos y páginas o sitios web, así como entre las citaciones o citas y los enlaces, como nexo de unión entre documentos, condujo a que, a mediados de los años noventa, investigadores del campo de la Bibliometría aplicaran sus técnicas de análisis cuantitativo a la Web, con resultados positivos. El propio Berners-Lee (1999: 35) sitúa la conexión entre el hipertexto y las prácticas de la producción científica en el origen de la misma invención de la Web: "La comunidad de investigadores había usado vínculos entre documentos en papel desde siempre: las tablas de contenidos, los índices, las bibliografías y las notas son vínculos de hipertexto". Estas similitudes quedan de manifiesto en trabajos que de forma pionera han aplicado conceptos específicos de la Bibliometría a la Web, por ejemplo:

- Rousseau (1997) emplea el término de "sitación" para hacer referencia a una página web que es enlazada por otra, de manera equivalente al término tradicional de "citación";
- Ingwersen (1998) diseña el índice de impacto web (*web impact factor*) de forma análoga al índice de impacto de revistas científicas;
- Larson (1996) lleva a cabo un análisis de co-enlaces a semejanza de los análisis de co-citación (*co-citation analysis*); o
- Thelwall y Wilkinson (2004) adaptan el concepto de enlace bibliográfico (*bibliographic coupling*).

A pesar de las analogías, existen destacadas diferencias, especialmente las derivadas del hecho de que la Web no es un contexto en el que el comportamiento esté sistematizado, por ejemplo, a la hora de enlazar, algo que sí ocurre en las producciones científicas, donde las citaciones responden a una serie de motivaciones y se codifican de una manera específica (van Raan, 2001). La falta de

uniformidad y de formalización hace que la variedad de información disponible en la Web sea más rica al tiempo que más complicada de analizar y tratar.

Dado que la Webmetría es todavía una disciplina reciente, mucha de la investigación tiene que ver con cuestiones metodológicas y con una perspectiva exploratoria, al objeto de evaluar las posibilidades de estudio y su aplicación en la evaluación de teorías existentes en diversos campos así como en el desarrollo de nuevas teorías que expliquen hechos propios de la Web. Por las razones expuestas, se puede comprender el hecho de que la mayor parte de la investigación realizada se haya efectuado en relación con la producción científica. El empleo de las técnicas y metodologías webmétricas a otras áreas es todavía un campo por explotar, si bien las contribuciones son cada vez más numerosas (por ejemplo, en campo de la empresa o la política).

Thelwall (2008c) apunta algunas de las ventajas que la Webmetría presenta frente a la Bibliometría tradicional. Entre ellas señalamos algunas que son de interés para los estudios de empresa de cara a afrontar nuevos proyectos de investigación: por un lado, en la Web podemos disponer de información actualizada de manera continua (únicamente sujeta a la viabilidad de acceder a ella a través de motores de búsqueda, arañas u otros medios); y, por otro, el acceso a la información es en principio gratuito. Adicionalmente, la Web como espacio de interacción social y económico constituye un incentivo en sí mismo para la investigación, proporcionando la Webmetría una perspectiva consistente para abordar investigaciones en este campo.

El desarrollo de la Webmetría se vincula también con la aplicación de técnicas de minería de datos a Internet y a la Web en particular. La Minería de Datos (Benoît, 2002) plantea soluciones para poder enfrentarse a la ingente cantidad de información no estructurada disponible en la Web. La aplicación de técnicas de minería de datos permite aprovechar y sacar el máximo partido a formas de análisis que se han venido aplicando hasta ahora en bases de datos con el objetivo de identificar patrones significativos en la información.

En el Apartado 2.4.2.7 se ha presentado la Web, entendida como una base de

datos que puede ser explotada mediante la minería de datos o, de forma más global, el *Knowledge Discovery Database* (KDD, en adelante). Desde el punto de vista de la minería de datos aplicada a la Web, se distinguen tres líneas principales, según los elementos objeto de análisis (Madria et al., 1999):

- Minería del contenido de la Web (*Web content mining*), que se centra en el contenido de las páginas web;
- Minería de la estructura de la Web (*Web structure mining*), que analiza la estructura de enlaces de la Web; y
- Minería del uso de la Web (*Web usage mining*), que estudia los patrones de comportamiento de los usuarios, así como el uso que hacen de las páginas web.

Björneborn y Ingwersen (2001) también identifican estos tres ámbitos de estudio en relación con el descubrimiento de conocimiento en la Web. Los trabajos sobre la estructura de la Web se refieren fundamentalmente al análisis de hiperenlaces. Sobre ellos trataremos en profundidad a lo largo del capítulo. La investigación sobre el contenido de la Web puede proporcionar información de utilidad sobre tendencias, corrientes de pensamiento y opinión, así como respaldar y validar los trabajos basados en la estructura de la misma. Los estudios sobre el empleo de la Web por los usuarios permiten comprender mejor su comportamiento en la red y sus necesidades de información. Buena parte de los trabajos en esta área se han centrado en estudiar cómo los usuarios emplean los buscadores para satisfacer sus necesidades de información (Zhang, Jansen y Spink, 2009).

Actualmente, la Webmetría, así como otras formas de minería de datos, afronta el reto de analizar la Web 2.0 y la inmensa cantidad de información que estas herramientas, tales como blogs o redes sociales, han contribuido a generar. Thelwall, Vaughan y Björneborn (2005) consideran que la Webmetría debe ser capaz de desarrollar herramientas y técnicas que sean capaces de identificar, observar y cuantificar los cambios sociales que tienen lugar en Internet. Una línea

de trabajo, por ejemplo, es el *Sentiment Analysis*, que pretende desarrollar sistemas para detectar de manera automática las opiniones vertidas por los usuarios sobre un asunto, identificando si son positivas o negativas, por ejemplo. El empleo de las bases de datos de servicios 2.0 representa una oportunidad de investigación muy atractiva, especialmente para realizar microanálisis de determinados fenómenos en la Web. De cara al futuro de la investigación webmétrica, un elemento fundamental es la evolución de las funciones de consulta que los motores de búsqueda proporcionan para obtener información sobre enlaces, dado que, como veremos, son la única fuente viable para abordar estudios de la Web en su conjunto.

Por último, cabe señalar que el estudio de la Web se ha abordado desde distintas perspectivas, que van más allá de la Webmetría. Han surgido diversas "disciplinas" que confluyen en el objeto de estudio, la Web como tal, proporcionando visiones complementarias (para evitar confusión se incluyen los términos en inglés):

- *Cyber geography* y *Cyber cartography* (por ejemplo: Girardin, 1995; 1996; Dodge y Kitchin, 2001; 2002).
- *Web ecology* (por ejemplo: Pitkow, 1997; Chi et al., 1998; Huberman, 2001).
- *Web mining* (por ejemplo: Etzioni, 1996; Cooley, Mobasher y Srivastava, 1997; Kosala y Blockeel, 2000).
- *Web graph analysis* (por ejemplo: Broder et al., 2000).
- *Web dynamics* (por ejemplo: Levene y Poulovassilis, 2001).
- *Web intelligence* (por ejemplo: Yao et al., 2001).

3.2. Terminología

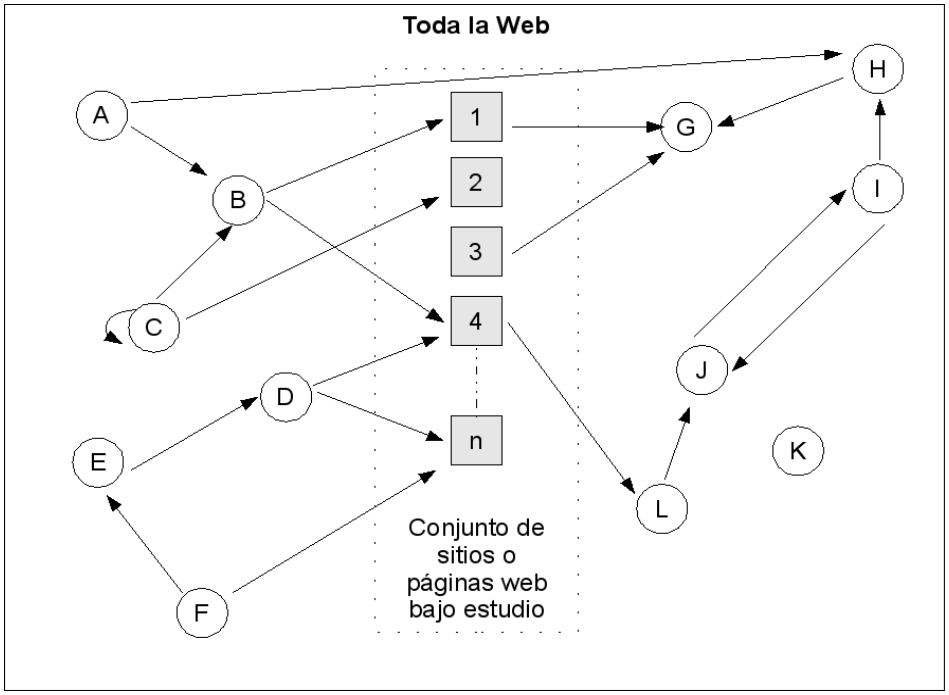
Han sido varios los intentos de fijar una terminología webmétrica que permita evitar el empleo de palabras con más de un significado o el uso de varios términos para referirse a un mismo concepto. Con todo, la cuestión aún no se ha podido resolver

satisfactoriamente y son muchas las variaciones que se presentan en la literatura.

El principal intento de fijar una terminología es el de Björneborn e Ingwersen (2004). Posteriormente, Thelwall y Wilkinson (2008), partiendo de las aportaciones anteriores, elaboran un marco de referencia de léxico unificado con el objeto de permitir a los usuarios describir y comparar sistemáticamente un amplio conjunto de métodos alternativos empleados en la investigación webmétrica. La Figura 3.2 muestra un esquema donde se representan distintos elementos web (sitios, páginas, etc.) y los posibles tipos de enlaces existentes entre ellos. El rectángulo de puntos en el centro, que incluye una serie de nodos representados por cuadrados grises, hace referencia a un conjunto de sitios o páginas web bajo estudio. Permite ilustrar la investigación basada en co-enlaces que se expondrá posteriormente.

Los conceptos expuestos a continuación se basan en la propuesta de Björneborn e Ingwersen (2004) y, de forma complementaria, en Thelwall y Wilkinson (2008). Son conceptos que tienen sus orígenes en la teoría de grafos, en el análisis de redes sociales y en la Bibliometría.

Figura 3.2. Esquema de la Web y de los tipos de conexiones existentes



A continuación se describen los principales conceptos webmétricos (donde sea relevante se incluye el término inglés más usual):

- **Enlace entrante, enlace recibido, sita o sitación (*Inlink*).** *B tiene un enlace entrante de A. B recibe un enlace de A. B recibe una sita de A.*

Faba Pérez, Guerrero Bote y Moya Anegón (2004a) emplean la terminología de Rousseau (1997), llamando a los enlaces recibidos *sita* o *sitación*, por analogía con el concepto de *cita* o *citación*, propio del análisis bibliométrico. En el trabajo nos inclinamos por el empleo de la palabra "enlace" por ser de uso común y propio de la Web.

- **Enlace saliente (*Outlink*).** *E tiene un enlace saliente a D. E enlaza a D. E sita a D.*

- **Auto-enlace (*Self-link*).** C *tiene un autoenlace. C se autoenlaza.*

Se pueden distinguir distintos tipos: auto-enlace a nivel de página web (*page self-links*; son enlaces que van de una sección a otra dentro de una misma página), auto-enlace a nivel de sitio web (*site self-links*) o enlaces internos (enlaces con funciones de navegación que van de una página a otra dentro de un mismo sitio web).

- **Enlaces recíprocos (*Reciprocal links*).** *J e I tienen enlaces recíprocos. J e I se sitan recíprocamente.*

Este tipo de vinculación no es posible en otras publicaciones impresas. Se trata de un tipo de conexión en el que existen *inlinks* y *outlinks* mutuos entre dos nodos. La reciprocidad no tiene porqué ser simétrica ya que es posible que existan más

enlaces en una dirección que en otra. En ocasiones, es posible que los enlaces recíprocos sean fruto de un acuerdo entre los dos creadores del sitio web con el fin de obtener mayor relevancia en los motores de búsqueda.

- **Enlace transversal (*Transversal link*).** *A tiene un enlace transversal saliente a H. A sita transversalmente a H.*

Se trata de un enlace que une dos áreas distantes de la Web que no se encuentran bien interconectadas. La mayoría de enlaces en la Web conectan páginas con contenidos similares; sin embargo, en ocasiones, ese patrón se rompe y se conectan espacios muy distintos. Este tipo de enlaces son los transversales (Björneborn, 2001; 2004; Björneborn e Ingwersen, 2001), que funcionan como atajos entre temas en apariencia desconectados entre sí. Los enlaces transversales hacen referencia al fenómeno del "mundo pequeño" ("small-world"), que abordamos en el Apartado 2.4.2.8.

- **Página o sitio web aislado**. *K representa una página o un sitio aislado ya que no recibe y establece ningún enlace. No se encuentra conectado con otros nodos.*

- **Co-enlaces entrantes, co-enlaces recibidos, co-sita, co-sitación (*Co-inlinks*).** *D es un co-enlace entrante de 4 y n. 4 y n son co-sitados por D, dado que B enlaza a ambos al mismo tiempo.*

Parte de la investigación empírica llevada a cabo en esta tesis se basa en co-enlaces entrantes, a los que nos referiremos más brevemente como co-enlaces. Como apuntamos anteriormente, los co-enlaces de este tipo se basan en el concepto de co-citación (Small, 1973). El análisis de co-citación parte de la idea de que si un documento cita a otros dos documentos, estos dos documentos tendrán

con probabilidad un contenido similar. Cuantas más veces sean co-citados mayor supondremos que será la similitud existente. Este razonamiento es en el que se fundamenta el análisis de co-enlaces.

- **Co-enlaces salientes o espacios web relacionados (*Co-outlinks*).** *1 y 3 son co-enlaces salientes de G, dado que ambos enlazan a G simultáneamente. 1 Y 3 tienen la referencia común G.* El término "espacios web relacionados" aparece en Faba Pérez, Guerrero Bote y Moya Anegón (2004a: 74).

El análisis de co-enlaces salientes se inspira en el análisis de documentos relacionados (*bibliographic coupling*). Se parte de la idea de que dos documentos distintos serán similares si enlazan conjuntamente a un tercero (Kessler, 1963; 1965). Ello significa que ambos documentos tienen un interés compartido en el tercer documento y por tanto es lógico pensar que están relacionados. El contenido del sitio enlazado puede proporcionar información acerca de cuáles son los intereses compartidos por ambos documentos.

La similitud que se desprende del análisis de co-enlaces puede tener significados diversos en función del contexto de análisis. Las motivaciones para la creación de los co-enlaces podría proporcionar información valiosa al respecto. Por ejemplo, en el ámbito de empresas, si analizamos un conjunto de entidades pertenecientes a un mismo sector empresarial es lógico pensar que la similitud hace referencia al grado de competencia entre ellas (Vaughan y You, 2006; 2008).

Adicionalmente se incluyen un par de términos definidos por Thelwall y Wilkinson (2008: 94) que son de interés para entender mejor este tipo de investigación:

- **Enlace**: entendido como un par de URLs (A y B) donde A es la URL de la página que contiene el enlace y B es la URL de la página que recibe el enlace.

- **Lista de co-enlaces**: en principio una lista de co-enlaces no es más que una lista de enlaces que cumplen unas determinadas condiciones. El modo en que estos co-enlaces están construidos condicionan el que hablemos de co-enlaces entrantes o salientes.

Normalmente al hablar de co-enlaces, sin hacer ninguna especificación adicional, se hará referencia a co-enlaces entrantes (*co-inlinks*). Toda la investigación empírica incluida en la tesis se basa en co-enlaces entrantes.

Si aceptamos las analogías entre Bibliometría y Webmetría, el término equivalente a *outlink* es "referencia" y a *inlink* es "citación". La diferencia entre *inlink* y *outlink* no es más que una cuestión de la perspectiva desde la que hablamos. Si un enlace es visto desde la el punto de vista del nodo que lo recibe será un *inlink*, mientras que si es observado desde el punto de vista del que lo emite será un *outlink*.

La terminología expuesta anteriormente se limita a una perspectiva exclusivamente webmétrica, que es la que vamos a emplear en nuestros trabajos empíricos. Sin embargo, es posible el análisis de enlaces desde diferentes perspectivas, como las matemáticas (teoría de grafos), la física (redes complejas) o la sociología (análisis de redes sociales). Holmberg (2009) lleva a cabo un análisis más pormenorizado de la terminología de las distintas áreas.

3.3. La obtención de datos para la investigación webmétrica

La obtención de datos constituye una parte fundamental de cualquier investigación. En el caso de las investigaciones basadas en técnicas webmétricas, la obtención de datos sobre enlaces presenta notables particularidades causadas por la naturaleza de la información que hay que recabar, así como limitaciones y sesgos, en función del sistema elegido para capturarla. Una característica que hay que tener en cuenta es la ausencia de bases de datos que recaben y almacenen este tipo de información como sí ocurre, por ejemplo, con variables empresariales de tipo

financiero. Es por ello que se han de emplear unos métodos particulares para la obtención de la información.

Existen tres formas principales para obtener información de enlaces en la Web. En este apartado abordaremos cada una de ellas, haciendo especial mención al empleo de motores de búsqueda, ya que es el sistema empleado en los distintos trabajos desarrollados en esta tesis doctoral. Los buscadores son además el método más viable y en ocasiones el único disponible para llevar a cabo una investigación que abarque a la Web en su totalidad.

A finales de los noventa, cuando la investigación webmétrica aún se encontraba en sus inicios, Almind e Ingwersen (1997) llevaron a cabo una enumeración de fuentes empleadas para la recolección de datos. Si bien se encuentra desfasada debido a los profundos cambios experimentados desde entonces, se incluye la propuesta de los autores con el objeto de introducir la cuestión y mostrar la existencia de una pluralidad de medios:

- Basada en servidores:
    - Obtención de datos basada en el acceso a los archivos almacenados en el servidor. Se trata del método más óptimo para investigar la presencia en la Web de una institución en particular o de un conjunto de ellas. Si bien permite recabar una información exhaustiva, su obtención está condicionada a que los responsables permitan acceder al servidor. Cuando se realiza un estudio a mayor escala este sistema no resulta viable en la práctica.

- Basada en clientes:
    - Índices web. Permiten acceder a un gran número de resultados. Sin embargo, los resultados de los motores de búsqueda empleados muestran resultados poco consistentes.

    - Listas o directorios de páginas o sitios web (como Yahoo).
    - Bases de datos e índices (como Lycos).
- A través de páginas conocidas, que contienen a su vez enlaces a otras páginas.
- Navegando a través de la red.

Bar-Ilan (2008) señala que hay dos métodos principales para la obtención de datos: por un lado, el empleo de una araña (*web crawler*) y, por otro, los motores de búsqueda (*search engines*). En la misma línea se pronuncian Björneborn e Ingwersen (2001). Thelwall, Vaughan y Björneborn (2005) incluye también otro método como es el empleo de ficheros de registro, conocidos en inglés como *Web log files*. Esta perspectiva permite el análisis de la información que los servidores guardan sobre las consultas de información que realizan los usuarios, por lo que es muy indicado para estudiar el comportamiento de los usuarios en la red.

Centrados en la obtención de información sobre hiperenlaces, Stuart (2008) y Holmberg (2009) señalan que son tres las formas principales de obtener este tipo de datos:

- de forma manual, visitando los sitios en estudio (por ejemplo: Kretschmer y Aguillo, 2004);
- mediante el empleo de arañas (por ejemplo: Chen et al., 1998; Terveen y Hill, 1998; Thelwall, 2001c); o,
- mediante datos obtenidos de buscadores (por ejemplo: Rousseau, 1997; Vaughan 2004a).

El empleo de un sistema u otro depende de los objetivos de la investigación. En

líneas generales se puede señalar que, para estudios de partes específicas de la Web o de conjuntos de sitios web delimitados es preferible el empleo de arañas. En el caso de que la investigación se limite a un grupo pequeño de sitios web, pero se base en la estimación del número de enlaces que éstos reciben de sitios fuera del propio grupo de páginas en el estudio, es muy probable que este sistema sea inviable dado que habría que tener en cuenta la información procedente de toda la Web o de partes muy amplias que fueran relevantes para la investigación. Es lo que ocurre, por ejemplo, con los estudios de co-enlaces presentados en esta tesis. Aunque sólo se estudian grupos de aproximadamente 50 sitios web, se requiere conocer el número de enlaces que provienen de terceras páginas que pueden estar distribuidas por toda la Web. Ello implica que se deba emplear un buscador para obtener la información a pesar de las limitaciones que presentan, como se verá en el Apartado 3.3.3.5.

Los motores de búsqueda comerciales son, en determinados casos, la única opción factible para hacer estudios de la Web con una mayor extensión. Mientras que con las arañas el investigador tiene un control más completo sobre la información que recaba, con este sistema depende de la información que faciliten los motores de búsqueda, los cuales emplean criterios para indexar y proporcionar resultados que en la mayoría de los casos son desconocidos por tratarse de secretos comerciales. En las siguientes páginas veremos las ventajas y desventajas de cada uno de los sistemas, si bien es justo reconocer que en la mayoría de los casos no se trata de una cuestión en la que el investigador pueda elegir libremente, sino que se verá forzado a utilizar uno u otro método en función de sus objetivos.

En cualquier caso, se sugiere, cuando sea posible, la triangulación entre los distintos métodos, ya que puede servir para verificar y validar la información recabada. Por ejemplo, el investigador puede visitar algunos de los sitios web proporcionados por los buscadores para comprobar la existencia de los enlaces que se buscan así como la naturaleza de la página en cuestión.

A continuación se aborda cada uno de los métodos de manera más detenida, centrándonos en los motores de búsqueda.

3.3.1. Visita manual de los sitios web

La visita de sitios web de forma manual puede requerir mucho tiempo y recursos por lo que no es recomendable salvo que el número de sitios web en estudio sea muy reducido (Thelwall, Vaughan y Björneborn, 2005). Además, el modo en que los hiperenlaces se presentan en una página web -por ejemplo incluidos en fotografías, no resaltados en el texto o bien directamente ocultos-, hace que este procedimiento se encuentre muy expuesto al error humano. De este modo, aparte de ser un sistema apenas utilizado, sólo es recomendable en el caso de que deseen hacerse comprobaciones adicionales en base a otras metodologías. Kretschmer y Aguillo (2004), por ejemplo, emplearon este sistema para investigar los enlaces entre 17 páginas de inicio de carácter académico.

3.3.2. Arañas Web

Las arañas (en inglés, *web crawlers*, *spiders* o *robots*) son programas especialmente diseñados para recabar información en la Web, ya sea tanto de hiperenlaces como de contenidos de las páginas. Para ello, por lo general, el investigador suministra una serie de páginas de inicio a partir de las cuales la araña inicia el proceso de obtención de datos siguiendo los enlaces contenidos en dicha página hacia otras páginas. Thelwall, Vaughan y Björneborn (2005: 92) definen araña web como "software that can automatically and iteratively download pages and extract and download their links". Este programa permite, por consiguiente, recorrer la red dentro de la cual se enmarca la página de inicio suministrada. El proceso finaliza cuando ya no hay más enlaces que seguir o bien cuando se alcanza el nivel de análisis establecido por el investigador.

Stuart (2008) señala la existencia tanto de arañas con finalidades comerciales como con finalidades académicas. También distingue entre aquellas que están centradas en la recolección de información sobre contenido de las páginas o sobre los hiperenlaces que contienen. Cada araña proporciona unas características distintas

al investigador, desde el tipo de información que se puede recoger hasta la configuración de distintos parámetros de consulta. Hay un gran número de trabajos que han empleado arañas para la obtención de los datos:

- tanto arañas diseñadas con propósitos comerciales (por ejemplo: Terveen y Hill, 1998; Koehler, 1999; Heimeriks, Horlesberger y van den Besselaar, 2003; Priego, 2003); como,

- arañas diseñadas específicamente para investigación (por ejemplo: Thelwall, 2001d; 2002a; 2002b; 2003a; 2003b; Thelwall y Price, 2003; Thelwall, Harries y Wilkinson, 2003; Wilkinson et al., 2003; Thelwall y Wilkinson, 2004).

A continuación se enumeran algunas de las limitaciones que presenta este sistema a la hora de obtener datos de cara a la investigación:

- Existen dificultades para indexar páginas web creadas de forma dinámica, ya que la araña podría contribuir a generar contenido de manera automatizada por el simple hecho de consultar unos enlaces. Sería el caso de un calendario que genera nuevas páginas a partir de un enlace que lleve a la semana o al mes siguiente.

- Existen también problemas para indexar contenido web no programado en HTML, por ejemplo, información incluida en elementos diseñados en Flash o JavaScript. Esto puede causar problemas tales como la no indexación de determinadas zonas relevantes de la Web, debido a que el *software* empleado es incapaz de identificar el hiperenlace que debe seguir.

- Es posible que determinadas partes de la Web se encuentren protegidas mediante contraseñas o bien que determinados sitios se encuentren bloqueados mediante el empleo del protocolo *robots.txt* (Koster, 2009a; 2009b).

- Se pueden presentar dificultades para la indexación de otros tipos de archivos, tales como Microsoft Word, Excel, PowerPoint, PDF u otros

formatos análogos.

- La araña puede no tener la potencia suficiente o puede no estar programada para detectar páginas con contenido duplicado.

- También habría que examinar cuestiones técnicas como las que se derivan del tipo de conexión a Internet (si existe banda ancha o no, por ejemplo) y de la capacidad de procesamiento para tratar el volumen de información requerido. Igualmente, es preciso valorar el coste que puede conllevar el realizar procesos de indexación demasiado extensos en el tiempo o demasiado costosos por su consumo de recursos.

- Otras limitación que cabe mencionar es de tipo ético. Tiene que ver, entre otras cosas, con la denegación de acceso a un servidor motivada por el excesivo número de consultas realizadas por la araña, lo cual puede ser interpretado como un ataque a los sistemas informáticos. También tiene que ver con cuestiones de privacidad derivada del tipo de información que se procesa o de los derechos de *copyright*. Thelwall y Stuart (2006) examinan de forma más detenida estas observaciones, así como proporcionan indicaciones para mantener un comportamiento ético en este sentido.

El empleo de arañas requiere que el investigador supervise el proceso de obtención de los datos, con el objeto bien de evitar estos problemas, en la medida de lo posible, o bien de controlar su efecto sobre los datos recabados. A pesar de las limitaciones, el empleo de este sistema permite un mayor control sobre la cobertura realizada de los sitios en estudio así como sobre los algoritmos empleados para el recuento de enlaces para el posterior análisis.

Como se ha indicado, existen distintos tipos de arañas, unas con propósitos comerciales y otras destinadas específicamente a la investigación. Es importante señalar que el empleo de un programa u otro tiene un impacto en los resultados obtenidos. Tanto Koehler (1999), empleando dos programas distintos, como Arroyo (2004), comparando siete arañas de tipo comercial y una con fines académicos,

concluyen que la elección del programa puede tener repercusiones significativas en el número de páginas obtenidas dentro de un sitio web, debido principalmente al fenómeno de las páginas dinámicas. Se puede concluir, a la luz de estos trabajos, que no existe una medida exacta del hecho que se quiere cuantificar, sino más bien distintas medidas en función de los parámetros que se seleccionan para recabar la información y de los programas empleados. Por ejemplo, estos programas pueden configurarse para recabar información únicamente de determinadas partes de la Web o de páginas con determinadas características, tanto de forma como de contenido, con el objeto de respetar determinadas restricciones éticas o con la vista puesta en detectar páginas que incluyen contenidos duplicados.

Los motores de búsqueda comerciales también emplean arañas para indexar la Web con el fin de proporcionar sus servicios de búsqueda. La diferencia con respecto a las arañas empleadas por particulares es que, en el primer caso, toda la información obtenida no se encuentra disponible ya que no se puede acceder libremente a las inmensas bases de datos de los motores de búsqueda. En este sentido, existen programas que permiten automatizar la recogida de datos en los motores de búsqueda. Esto puede realizarse mediante el empleo de una API (Apartado 3.3.3.6), que permite realizar consultas directamente a las bases de datos del motor de búsqueda, o mediante el empleo de programas que automatizan la obtención de datos a partir de los resultados mostrados por la propia interfaz de los buscadores. Dado que la clasificación de las distintas formas de obtener datos se ha efectuado en función de su origen, este tipo de programas se encuadraría con mayor propiedad en la sección siguiente, dedicada a los motores de búsqueda.

3.3.3. Motores de búsqueda

Puesto que los motores de búsqueda constituyen la fuente de información empleada en los trabajos incluidos en el Capítulo 4, en este apartado examinaremos más detenidamente cuestiones relativas a su importancia, funcionamiento, historia, investigación y las ventajas e inconvenientes de su empleo

con fines webmétricos.

La importancia de los buscadores es tal en nuestra sociedad que organizaciones y empresas se esfuerzan por posicionarse en los puestos más altos de sus *rankings* (Batelle, 2006). Los motores de búsqueda se han convertido en la principal puerta a la información, en la guía de acceso a empresas y mercados, por lo que la visibilidad de los sitios web es fundamental para el éxito de cualquier entidad (Espadas, Calero y Piattini, 2008). Esto ha creado la demanda de nuevos servicios como el de los profesionales destinados a optimizar el posicionamiento en los buscadores (Rimbach, Dannenberg y Bleimann, 2007). Son conocidos como SEO (*Search Engine Optimization*).

De cara a la investigación, el empleo de buscadores está indicado cuando se abordan amplias áreas de la Web (Thelwall, Vaughan y Björneborn, 2005) o la Web en su totalidad. Por ejemplo, sería el caso de aquellos estudios que requieren conocer el número de enlaces que apuntan a un sitio web. Potencialmente estos enlaces podrían proceder de cualquier parte de la Web, por lo que no es factible abordar la obtención de los datos a partir de una araña propia del investigador. Los datos obtenidos de buscadores se han empleado en un gran número de trabajos webmétricos (por ejemplo: Larson, 1996; Rousseau, 1997; Ingwersen, 1998; Thomas y Willet, 2000; Thelwall, 2001d; Thelwall y Smith, 2002; Musgrove et al., 2003; Tang y Thelwall, 2003; Thelwall y Tang, 2003; Thelwall y Harries, 2004; Bar-Ilan, 2004; 2005; Vaughan 2004a; Vaughan y Thelwall, 2003; Vaughan y Wu, 2004; Vaughan y You, 2006; 2008).

Al margen de las búsquedas realizadas a partir de palabras clave, los buscadores ofrecen una serie de operadores especiales para la obtención de información sobre hiperenlaces. Estas funciones en ocasiones se pueden combinar mediante operadores booleanos, permitiendo refinar la obtención de los datos.

3.3.3.1. Concepto

Conviene aclarar cuál es el concepto de "motor de búsqueda" (*search engine*) o "buscador", ya que bajo dicha denominación se incluyen una amplia variedad de instrumentos de búsqueda que, en ocasiones, funcionan de forma completamente distinta, desde Google a Yahoo, pasando por los metabuscadores como Metacrawler, Vivisimo o Copernic, hasta llegar a los buscadores especializados en blogs, noticias, wikis, imágenes, vídeos, etc.

Entre las distintas definiciones podemos citar las siguientes:

- Wikipedia, versión en inglés (http://en.wikipedia.org/wiki/Search_engine; consultada el 27 de junio de 2008): "A Web search engine is a search engine designed to search for information on the World Wide Web. Information may consist of web pages, images and other types of files. Some search engines also mine data available in newsgroups, databases, or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input." Se trata de una definición que se centra en los buscadores basados en arañas.

- Harrod´s Librarians Glossary (Prytherch, 2000: 657): "On the World Wide Web, third party search engines are available to permit searching the 'whole web' and these rely on robots that traverse the Web following links between pages and copying relevant information to create a database which is then indexed to form searchable keywords."

- Bar-Ilan (2004: 234): "A tool to which we can present textual queries of any kind, it retrieves results based on information in its database". De algún modo esta definición puede referirse a cualquier tipo de motor de búsqueda. En nuestro caso nos ocupan aquellos que alimentan su base de datos a partir de la información que obtienen de la Web.

Atendiendo a las distintas formas en las que puede funcionar un buscador, en Search Engine Watch (http://searchenginewatch.com/2168031, consultado el 9 de marzo de 2010) se distingue entre motores basados en arañas (por ejemplo, Google) y directorios de páginas o sitios web clasificados por seres humanos (por ejemplo el caso de Yahoo!, en su origen). Por su parte, Oppenheim et al. (2000) distinguen cuatro categorías de buscadores: robots, directorios, meta buscadores y herramientas de *software* (programas que se instalan en el propio ordenador del usuario). El término motor de búsqueda se ha empleado a menudo de forma genérica para referirse tanto a motores de búsqueda basados en arañas como a directorios clasificados por humanos. El modo en que funcionan ambos modelos es completamente distinto. La velocidad de crecimiento de la Web, así como su tamaño, hace prácticamente imposible mantener un modelo basado en una clasificación manual de páginas web, por lo que los buscadores más importantes se basan actualmente en el diseño de sistemas automáticos que emplean hiperenlaces para indexar la mayor parte posible de la Web. En la actualidad, dentro del marco de la Web 2.0, existe una gran cantidad de información generada por usuarios relativa a páginas web, documentos, contenido multimedia, etc. Hay nuevos buscadores que se basan en la utilización de esta información para proporcionar resultados que, de algún modo, combinan una búsqueda algorítmica con valoraciones cualitativas de los usuarios. Los usuarios generan metadatos que proporcionan al sistema información cualitativa relevante sobre la naturaleza de los recursos. Si bien no propiamente un buscador, podríamos decir que una búsqueda en un servicio para compartir "favoritos", como es, por ejemplo, *delicious.com*, podría responder a esta idea.

3.3.3.2. Historia de los motores de búsqueda

Como apunta Bar-Ilan (2004: 238), "ten years is much more than a lifetime on the Internet". La evolución de los buscadores a lo largo de los veinte años de vida de la Web ha sido muy rápida. En un principio, cuando la Web acababa de nacer, existía una lista completa de servidores web, mantenida por Tim Berners-Lee y alojada en

el servidor del CERN. Sin embargo, pronto, con la expansión y crecimiento de la Web, fueron apareciendo distintas herramientas de búsqueda cuyo objetivo era facilitar el acceso a la información. Entre los primeros motores de búsqueda a comienzos de los años 90 se puede mencionar Archie o Gopher. Posteriormente, hubo una gran proliferación de estas herramientas. Entre otros muchos, en 1994 se lanzó Lycos, que alcanzó un gran éxito comercial. Su historia sirve de algún modo para ilustrar la gran burbuja tecnológica que estalló en el año 2000 y de la que ahora apenas se cumplen diez años. Terra, propiedad de Telefónica, adquirió el buscador en una operación millonaria con el fin de posicionarse como líder en portales entre la comunidad latina de Estados Unidos; sin embargo, la operación acabó siendo un fracaso y la empresa se acabó liquidando años después.

Altavista fue uno de los buscadores más potentes durante los años 90, siendo utilizado de forma mayoritaria para las primeras investigaciones webmétricas. Altavista, al igual que otros buscadores como Hotbot, empleaban la frecuencia de aparición de palabras en los documentos para ordenar los resultados de acuerdo con un criterio de relevancia (Schwartz, 1998). Este sistema, sin embargo, permitía la manipulación por parte de los editores de las páginas web, que podían controlar el número de términos clave que incluían en los textos.

Google hizo su aparición en 1998, alcanzando el éxito en tan solo dos años. Su algoritmo para organizar los resultados de las consultas, el *PageRank*, se basaba en la estructura de enlaces entre las páginas web (Brin y Page, 1998; para una explicación de su funcionamiento consultar el Apartado 3.3.3.4). Este sistema supuso un cambio decisivo en la industria de los buscadores de contenidos que, con el tiempo, modificaron su diseño para emplear la valiosa información implícita en los hiperenlaces.

Yahoo!, que había surgido como un directorio de páginas web creado de forma manual, no lanzó hasta el año 2004 su propio motor de búsqueda en línea con lo que venía haciendo Google. Hasta entonces, en su página proporcionaba servicios de búsqueda a través de Inktomi, Altavista e incluso el propio motor de búsqueda de Google.

Microsoft, por su parte, lanzó en 1998, MSN Search, empleando también los servicios del buscador Inktomi. Hasta 2004, Microsoft utilizó para sus búsquedas resultados de Inktomi, Looksmart y Altavista. A partir de ese año, trabajó con su propia tecnología. En 2005, lanzaron MSN Search, que, en 2006, pasa a llamarse Live Search y, en 2009, Bing.

Microsoft y Yahoo! llegaron en 2009 a una alianza en el campo de las búsquedas de contenidos y la publicidad, con el fin de plantar cara a Google, que acapara más del 75% de las búsquedas patrocinadas, siendo las cuotas de mercado superiores en algunas zonas, por ejemplo, en Europa. El panorama del sector de los motores de búsqueda tiende a concentrarse, pudiéndose hablar en determinados casos de posiciones oligopolistas o incluso de monopolio. Los movimientos en el sector se suelen saldar con cambios en los servicios de búsqueda ofrecidos. En este sentido, la alianza entre Yahoo! y Microsoft probablemente afectará, a medio plazo, a algunas funciones de consulta empleadas en la investigación webmétrica.

3.3.3.3. Tipos de motores de búsqueda

Bar-Ilan (2004) clasifica los motores de búsqueda en tres tipos:

- Basados en arañas (*crawler-based engines*);
- Basados en directorios de páginas o sitios web organizados por personas (*human-powered directories*); y
- Meta-buscadores (*Meta-search engines*).

Los motores de búsqueda basados en arañas que indexan la Web son las herramientas clave para una parte relevante de la investigación webmétrica. Por arañas nos referimos a programas informáticos que, de manera automática, visitan y descargan páginas web para su posterior indexación y clasificación en las bases de datos de los motores de búsqueda. Actualmente el principal ejemplo de este tipo de buscadores es Google, si bien también podemos citar a los otros dos grandes

motores de búsqueda, Yahoo! y Bing de Microsoft, los cuales alcanzaron una alianza en 2009. Los buscadores están basados en tres componentes principales: la araña (*crawler*), el indizador (*indexer*) y el motor de consultas (*query engines*).

Un elemento clave a la hora de entender el funcionamiento de este tipo de motores de búsqueda es el sistema de algoritmos que emplean para ordenar la información de acuerdo con su relevancia (por ejemplo, el *PageRank* de Google). Estos algoritmos se ajustan periódicamente, si bien, constituyen secretos comerciales de las empresas por lo que no es posible conocer con exactitud cuáles son las variables y su peso a la hora de realizar un *ranking* de resultados a partir de la consulta efectuada por un usuario. Bar-Ilan (2004) indica que dichos algoritmos son secretos comerciales porque en ellos reside en buena medida el valor añadido de un motor de búsqueda y su ventaja competitiva frente a otros sistemas de búsqueda. Otra razón es la de evitar la manipulación de los *rankings* (fenómeno conocido como s*pamdexing*).

Los métodos para establecer dichos *rankings* están basados, entre otros criterios, en la relación entre el documento y el término empleado en la consulta, la situación de los términos de consulta en la página web y en algoritmos centrados en el análisis de la estructura de enlaces que vincula las distintas páginas. Este es el sistema que utiliza Google con su *PageRank*, que jerarquiza una página en función del número de enlaces que recibe, así como de la relevancia de las páginas que establecen las conexiones. De alguna manera se trata de un sistema de "votación". Existen también otros sistemas alternativos, por ejemplo, basados en la frecuencia con que una página es visitada (conocido como "*click-through*"). Sin embargo, como señala Bar-Ilan (2004), los parámetros para ordenar los resultados de una búsqueda no responden únicamente a criterios de popularidad o puramente matemáticos, sino que están influidos también por consideraciones de índole comercial. Es preciso recordar que los motores de búsqueda constituyen una suculenta industria que controla una buena parte de la inversión en publicidad, así como la puerta de acceso a servicios comerciales. Introna y Nissembaum (2000) consideran que la Web es un bien público y que los buscadores funcionan como facilitadores de información, desempeñando una función clave como es la de poner

en valor determinados sitios web mientras que hacen básicamente desaparecer otros. Partiendo de esta concepción, alertan de los peligros derivados de que el desempeño de los buscadores se supedite a criterios e intereses comerciales. Para más detalles sobre el funcionamiento de los motores de búsqueda se puede consultar: Brin y Page (1998), Arasu et al. (2001), Rasmussen (2003), o Bar-Ilan (2004). En el Apartado siguiente, 3.3.3.4, nos centraremos, a modo de ejemplo en el funcionamiento de Google.

Los buscadores basados en directorios cuentan con las limitaciones propias de un sistema que no está basado en una indexación automatizada de la Web. Era el caso de Yahoo!, que, en su origen, surge como un directorio de sitios web creado por sus fundadores, Jerry Yang y David Filo. Otro ejemplo es el del Open Directory (http://dmoz.org). Por lo general, la revisión de los contenidos se efectúa sobre todo el sitio web en su conjunto y no sobre el contenido de cada una de sus páginas. Existe una estructura de temas jerarquizada, aunque en ocasiones se establecen referencias cruzadas ya que un mismo sitio podría estar vinculado a distintas categorías. Como señala Bar-Ilan (2004: 236), "human-powered directories can be very useful as starting points for broad topical queries. On the other hand, such directories cannot compete in size and coverage with the crawler-based search engines". Actualmente, con la eclosión de la Web 2.0, comunidades para compartir favoritos, como es el caso de Delicious (http://delicious.com/), ofrecen una alternativa a este tipo de herramientas. Su funcionamiento se basa en la agregación de los millones de etiquetas empleadas por sus usuarios para describir el contenido de páginas y sitios web que consideran relevantes. En general, podemos decir que muchas herramientas 2.0, centradas en compartir contenidos, constituyen de algún modo "motores de búsqueda" especializados, que se basan en el empleo de los metadatos, generados por los propios usuarios, vinculados a los contenidos de la plataforma en cuestión.

Los meta-buscadores son herramientas de búsqueda que operan sobre otros

motores de búsqueda. Parten de la idea de que no hay ningún motor de búsqueda que, por sí solo, cubra toda la Web, debido a limitaciones de distinto tipo, relativas a las políticas de funcionamiento, indexación, obtención y jerarquización de resultados, etc. Bajo este razonamiento, la combinación de los resultados de todos ellos permitiría ofrecer una información más completa. Generalmente, los resultados se presentan en columnas distintas, siendo fruto de una búsqueda simultánea en distintos buscadores o bien se siguen políticas propias para jerarquizar el conjunto de resultados obtenidos, ofreciendo un único *ranking*. Entre los problemas con los que se enfrentan, están las limitaciones que los motores de búsqueda presentan para acceder a sus resultados. MetaCrawler o WebCrawler son dos ejemplos de meta-buscadores.

En ocasiones, los motores de búsqueda pueden combinar el empleo de arañas con los directorios, en cuyo caso se trata de motores de búsqueda híbridos (*hybrid search engines*).

Por último, cabe citar un caso especial, el de Internet Archive (www.archive.org), que consiste en un proyecto cuyo objetivo es archivar, con una perspectiva histórica, la mayor parte posible de la Web. En funcionamiento desde 1996, funciona como un motor de búsqueda que indexa la Web, almacenando copias en distintos momentos del tiempo de las páginas visitadas. Esta perspectiva histórica de la Web viene a mitigar la falta de "memoria" de los motores de búsqueda principales que, continuamente, actualizan sus bases de datos, no facilitando por lo general información histórica. Ello impide que se puedan realizar consultas referentes a momentos pasados. El proyecto Internet Archive cuenta con un buscador denominado Wayback Machine (http://www.archive.org/web/web.php) que permite consultar las copias de las páginas web guardadas. Esta función ha permitido realizar trabajos de investigación basados en, lo que podríamos denominar, indagaciones arqueológicas en la Web (Waite y Harrison, 2007). En cualquier caso, se trata de un servicio que se enfrenta a limitaciones similares a las del resto de motores de búsqueda, especialmente en lo relacionado con la

indexación mediante el empleo de arañas y sus problemas relativos al grado de cobertura de la Web en su conjunto y específicamente en determinados países (Thelwall y Vaughan, 2004). Un detalle de especial interés para la investigación webmétrica es que la información almacenada no es suficiente para obtener datos históricos sobre la estructura de enlaces. En distintos trabajos webmétricos (Vaughan y Thelwall, 2003; Vaughan, 2004a; Vaughan y Wu, 2004; Holmberg, 2009) se ha empleado el Internet Archive para fechar de forma aproximada el momento de creación de las páginas web. Vaughan y Thelwall (2003) y Vaughan (2004a) mostraron que el número de enlaces recibidos por un sitio web se correlaciona positivamente con el tiempo de existencia del sitio. Una razón para esto es que sitios con mayor tiempo de existencia tienden a ser más visibles y por tanto reciben más enlaces. Existe además un fenómeno de acumulación a la hora de recibir hipervínculos.

En adelante, al referirnos a motores de búsqueda o buscadores, haremos referencia únicamente a aquellos que indexan de manera automatizada la Web mediante el empleo de arañas, fundamentalmente a Google, Yahoo! y Bing.

3.3.3.4. El funcionamiento de los motores de búsqueda: el caso de Google

El objetivo de este apartado es ilustrar el funcionamiento de los motores de búsqueda basados en arañas, con el objeto de entender mejor cuál es el proceso de obtención de los datos que posteriormente emplearemos en los trabajos webmétricos. De modo sintético, los tres elementos clave de un buscador funcionan de la siguiente manera:

- La araña web (*crawler*) es la responsable de rastrear la red y recopilar los datos. Cada motor de búsqueda establece unas políticas distintas en cuanto a qué páginas visitar, con qué frecuencia y en qué orden. Estas políticas constituyen la principal diferencia entre buscadores en relación a su tamaño, cobertura y actualización de los resultados que ofrecen.

- La función del indizador o indexador (*indexer*) es la de extraer hipervínculos y palabras de las páginas web que descarga la araña. Junto a los enlaces y las palabras, se recopila información adicional relativa a la posición en la aparecen en la página y otras características (fuente, tipo de etiquetas HTML en la que aparece dicha palabra, etc.). La URL de la página indexada junto con las palabras que contiene dicha página se almacenan de forma inversa en pares en los que se vincula cada término con las URLs de todas las páginas en las que aparece. Se conoce como índice o archivo invertido (*inverted index or file*). Al margen de la anterior información, el buscador puede almacenar información adicional como metadatos, describiendo los contenidos o las URLs que enlazan a la página en cuestión.

- El motor de consultas (*query engine*) recibe las solicitudes de los usuarios y les da respuesta. Para ello formula una consulta al archivo invertido y ordena de acuerdo con determinados criterios los resultados, estableciendo un orden de relevancia a los miles o millones de URLs que pueden ser útiles para el usuario. Cada buscador dispone de sus propios algoritmos, los cuales son modificados constantemente con el fin de hacerlos más eficaces y evitar manipulaciones en los resultados.

Google, el buscador más popular, basa su diseño en la estructura anteriormente expuesta. Bar-Ilan (2004) apunta algunos de los factores que explican su éxito:

- La eficacia del sistema *Pagerank*, que establece la relevancia de los documentos en función del análisis de los enlaces entre las páginas web (Brin y Page, 1998);

- Su interfaz limpia y clara;

- El tamaño de su base de datos y su actualización relativamente frecuente; y,

- La posibilidad de acceder a las copias (*cached copies*) almacenadas de las páginas indexadas en su base de datos.

Googlebot es la araña que Google utiliza para indexar la Web. Actúa localizando y recuperando páginas que posteriormente serán incorporadas en la base de datos del motor de búsqueda. Funciona como un navegador normal, realizando consultas a los servidores web cuyas páginas descarga. El sistema está integrado por muchas computadoras que solicitan páginas web de una manera mucho más rápida a como puede hacerlo cualquier otro usuario con su propio navegador. Sin embargo, con el fin de evitar la saturación de los servidores o entorpecer al resto de usuarios, Googlebot realiza las peticiones de forma más lenta de lo que técnicamente podría hacerlo.

La localización de páginas web puede realizarse de dos formas: mediante la inclusión en el sistema de una URL (http://www.google.com/addurl.html) o de forma automática mediante los propios enlaces que Googlebot va encontrando en las distintas páginas. Cuando la araña encuentra una página, recopila todos los enlaces que aparecen y los añade a una lista para su posterior visita. En principio se podría afirmar que se localiza relativamente poco *spam* ya que la mayoría de los autores de las páginas web se encargan de realizar ese filtrado, es decir, los creadores de páginas web no suelen enlazar a páginas que contienen *spam*. De este modo, este sistema ayuda a que prevalezcan principalmente páginas web consideradas de "calidad". Con todo, ello no evita que el buscador deba emplear considerables recursos para combatir este problema así como otros intentos de manipular los resultados que ofrece. Con esta forma de proceder, Googlebot puede generar rápidamente un listado de enlaces que cubra amplias partes de la Web. Esta técnica, conocida como "*deep crawling*", permite al motor indagar con mayor minuciosidad dentro de los sitios web. Dado que el proceso se realiza a gran escala, las visitas "en profundidad" alcanzan a prácticamente todas las páginas potencialmente indexables. La lista de enlaces pendientes de visitar se debe examinar y comparar constantemente con aquellos enlaces (URLs) que ya se encuentran incluidos en el índice de Google. Los duplicados deben eliminarse con el fin de que Googlebot no revise las mismas páginas varias veces. La araña determina con qué frecuencia debe revisar una página web. Por una parte, es una

pérdida de recursos visitar páginas que no han cambiado; pero, por otra parte, Google debe actualizar aquellas páginas que son modificadas con el objeto de ofrecer resultados relevantes. Dada la inmensidad de la Web, el proceso es largo, por lo que algunas páginas son consultadas únicamente una vez al mes. En otros casos, para mantener el índice al día, debe revisar continuamente las páginas más populares, que son las que se modifican de forma constante. Este tipo visitas reciben el nombre de "*fresh crawls*". Por ejemplo, las páginas de los periódicos se visitan diariamente, mientras que las páginas con cotizaciones de bolsa se actualizan aún con mayor frecuencia. La combinación de *fresh crawls* y *deep crawls* permite a Google ser eficiente en el empleo de sus recursos, así como mantener el índice razonablemente actualizado.

El segundo elemento del buscador es el indexador o indizador (*Google's indexer*). Las páginas web visitadas y descargadas por Googlebot se almacenan en la base de datos. El índice se clasifica de manera alfabética por términos de consulta. A cada entrada del índice se asocia una lista de documentos en los que aparece el término, así como la localización del mismo dentro del texto. Esta estructura de los datos permite un acceso rápido a los documentos que contienen los términos de búsqueda requeridos por los usuarios. Para mejorar la búsqueda, se ignoran palabras comunes, denominadas *stop words*, tales como *the, is, on, or, of how*, etc. o sus equivalentes en otras lenguas, así como determinadas letras y números que aparecen de forma aislada. Este tipo de palabras o términos son tan comunes que no aportan información significativa a la hora de hacer una búsqueda precisa. También se ignoran signos de puntuación y espacios en el texto. Todas las letras se convierten en minúsculas para mejorar el desempeño del buscador.

El tercer elemento del motor de búsqueda es el motor de consultas, que está compuesto de varios elementos: la interfaz del usuario (la caja de búsqueda que vemos al utilizar el buscador), el "motor" que evalúa las consultas y las conecta con los resultados relevantes, y, por último, la función que permite la presentación de los resultados (*results formatter*). *PageRank* es el sistema que emplea Google para clasificar las páginas web de acuerdo a su relevancia. Fue desarrollado por los fundadores de Google, Larry Page y Sergey Brin (1998) en la Universidad de

Stanford y supuso una de las principales ventajas competitivas que permitió a Google convertirse en el motor de búsqueda de referencia en gran parte del mundo. El *PageRank* se basa en la información implícita contenida en la estructura de enlaces que da forma a la Web. El sistema interpreta que un enlace de una página A a una página B, es de alguna manera un "voto" favorable a la misma. Sin embargo, se tienen en cuenta otros factores, como, por ejemplo, la importancia de la página que emite el "voto". Al margen, los resultados de búsqueda están condicionados por las coincidencias con las palabras clave suministradas por los usuarios al realizar una consulta. Otras variables que se tienen en cuenta son: el número de veces que aparecen los términos en una página o el contenido de las páginas que la enlazan, entre otros. Una página con un *PageRank* más alto se considera más relevante y es probable que aparezca en un puesto superior a aquellas con un *PageRank* más bajo. Por tanto, al indexar el texto completo de una página web, Google va más allá de la simple coincidencia entre términos de consulta. Por ejemplo, da prioridad a las páginas que tienen palabras clave próximas entre sí y en el mismo orden que ha suministrado el usuario. Con el fin de mejorar la calidad de su servicio y evitar los intentos de alterar artificialmente los resultados, el buscador ajusta continuamente su algoritmo.

En este apartado, nos hemos centrado fundamentalmente en el funcionamiento de Google, que es el motor de búsqueda que, con más éxito, diseñó sus sistemas para explotar la información contenida en los hiperenlaces. Yahoo! y Bing en la actualidad también se basan en este sistema para desarrollar su servicio de búsquedas.

3.3.3.5. Limitaciones del empleo de motores de búsqueda en la obtención de datos para la investigación webmétrica

Los estudios webmétricos se enfrentan a numerosas dificultades derivadas de las características de la Web. La obtención de datos a través de los motores de búsqueda pone de relevancia muchas de estas limitaciones. Bar-Ilan (2004; 2008)

describe estas limitaciones, subrayando especialmente el hecho de que estas herramientas no están diseñadas para proporcionar información con fines de investigación, sino que persiguen propósitos comerciales. A pesar de ello, de entre la distintas opciones disponibles representan la mejor opción y, en ocasiones, la única disponible para efectuar determinados tipos de trabajos. Los fines comerciales de los buscadores pueden influir en las políticas de indexación de páginas web, así como en el posicionamiento de las mismas en los *rankings* de resultados (Thelwall, Vaughan y Björneborn, 2005).

En este apartado se abordan algunas de las principales limitaciones técnicas que presenta el empleo de motores de búsqueda para la investigación webmétrica. En primer lugar, podemos referirnos a la fiabilidad de los datos. Los buscadores no siempre facilitan o recuperan todos los elementos indexados en sus bases de datos (Mettrop y Nieuwenhuysen, 2001; Bar-Ilan, 2002). Además, el número total de páginas web que cumplen las condiciones de una determinada consulta no responde a un cálculo exhaustivo sino que está basado, por motivos de eficiencia, en estimaciones que permiten minimizar el tiempo de espera del usuario. En este sentido podemos decir que los motores de búsqueda, debido a sus propósitos comerciales, priman la eficiencia del servicio frente a su precisión y exhaustividad (Lipsman, 2007). Se ha asumido que el tiempo de respuesta a una consulta efectuada depende del volumen de tráfico de información que el motor de búsqueda esté procesando en un momento determinado. Por ello, en algunos trabajos se ha procurado realizar la obtención de datos en momentos en los que los buscadores soportan un menor nivel de uso (Thelwall, 2001d; Vaughan y Thelwall, 2003).

La falta de transparencia en el modo de funcionamiento de los motores de búsqueda es otro factor importante (Thelwall, Vaughan y Björneborn, 2005). Los algoritmos para elaborar los *rankings* de resultados no son públicos, como ya indicamos en el apartado anterior. Esto tiene repercusiones significativas, sobre todo si tenemos en cuenta que, debido a la enorme cantidad de información disponible en la Web, las páginas que alcanzan los primeros puestos en los resultados que ofrece un buscador son las que reciben la gran mayoría de visitas. Son escasos los usuarios que van más allá de los primeros diez o veinte resultados

(Silverstein et al., 1999; Spink et al., 2001; Wolfram et al., 2001).

Otra cuestión importante es el grado de cobertura que los motores de búsqueda hacen de la Web. Ya vimos en el Apartado 2.4.2 que el dinamismo de la Web, así como las características de determinadas páginas, hace imposible que los buscadores puedan indexar toda la información existente. Apuntamos también el fenómeno de la Web Invisible (Apartado 2.4.2.4). En este sentido, diversos trabajos han señalado que la cobertura que proporcionan los buscadores es poco uniforme, desigual y limitada (Lawrence y Giles, 1998; 1999b; Smith, 2003; Thelwall, 2000b; Vaughan y Thelwall, 2004; Jacsó, 2005; Spink et al., 2006), además de estar sujetos a fluctuaciones o a resultados poco fiables (Ingwersen, 1998; Snyder y Rosenbaum, 1999; Björneborn e Ingwersen, 2001; Mettrop y Nieuwenhuysen, 2001; Bar-Ilan, 2004). Se han detectado también sesgos en la indexación de determinadas áreas de la Web; por ejemplo, Vaughan y Thelwall (2004) indican que existe un sesgo a favor de Estados Unidos en la cobertura de la Web, frente a países como China, Taiwan y Singapur, debido a razones de tipo técnico e histórico.

En determinados casos, los motores de búsqueda han mostrado resultados poco consistentes (Risvik y Michelsen, 2002). Esto se ha comprobado realizando una misma consulta en momentos muy próximos, por ejemplo. Las causas pueden deberse, por ejemplo, al empleo de estimaciones, como apuntamos anteriormente. También cabe mencionar el hecho de que los buscadores tienden a eliminar automáticamente páginas que presentan contenidos duplicados. En otras ocasiones, las variaciones pueden deberse a actualizaciones en la base de datos , ya que continuamente se están indexando nuevas páginas. Otra razón que se ha esgrimido para explicar estas variaciones es que los motores de búsqueda no emplean una única base de datos sino múltiples con el objeto de enfrentarse a los grandes volúmenes de tráfico de consultas. Estas bases de datos pueden estar basadas en diversas arañas que indexan la Web de manera independiente. Google, por ejemplo, ha admitido la existencia de estas múltiples bases de datos (Google Librarian Central, 2007). En cualquier caso, lo que sí se debe dejar claro es que los resultados de una investigación que emplee este tipo de datos no se pueden reproducir ya que constantemente se detectan variaciones en el objeto de estudio y

en el instrumento empleado para medirlo (Rousseau, 1997).

La actualización de las bases de datos de los motores de búsqueda es un factor muy relevante, ya que parecen existir grandes diferencias en la frecuencia con la que las arañas de los distintos buscadores visitan las páginas web (Lewandowski, Wahlig y Meyer-Bautor, 2006). Como ya vimos en el caso de Google, es difícil encontrar un equilibrio adecuado en la frecuencia con la que se actualiza la información de la Web, ya que son diversos los factores que influyen de forma significativa: la periodicidad con la que determinadas páginas actualizan contenidos, las cuestiones éticas derivas de un incremento en el tráfico que soportan los servidores de las páginas web visitadas, los recursos limitados de los motores de búsqueda, etc.

Otro factor fundamental para la investigación webmétrica son los cambios en las funciones proporcionadas por los motores de búsqueda para realizar consultas. En ocasiones, incluso nos encontramos con motores de búsqueda que desaparecen. Si nos centramos en los tres grandes buscadores existentes en la actualidad (Google, Yahoo! y el buscador de Microsoft, Bing) podemos afirmar que a lo largo de los últimos cinco años se han producido modificaciones significativas que han obligado a los investigadores a adaptar sus prácticas para la obtención de datos. Por ejemplo, la API SOAP de Google dejó de dar claves de usuario a finales de 2006 (Google, 2006). Esta API proporcionaba unas buenas funciones para la obtención sistemática de datos de enlaces en la Web. Se puede encontrar más información sobre las APIs en el Apartado 3.3.3.6. Por otra parte, en 2007, Microsoft anunció la suspensión del funcionamiento de los operadores utilizados para obtener datos de enlaces (Live Search, 2007). Durante los primeros años de investigación webmétrica, Altavista fue el buscador de referencia por las funciones avanzadas que ofrecía y por su gran cobertura de la Web (Rodríguez Gairín, 1997; Ingwersen, 1998). Sin embargo, tras su adquisición por Yahoo!, éste pasó a convertirse en el buscador de referencia en este sentido. Google y el buscador de Microsoft se han empleado también en diferentes trabajos.

Otro problema son las limitaciones que los buscadores ponen a la realización de

búsquedas automatizadas, pudiendo derivar incluso en la cancelación o bloqueo del servicio. En este sentido basta con leer los *Terms of Service* de Google. Este tipo de búsquedas pueden ser interpretados como ataques a sus sistemas.

Todas estas limitaciones no son exclusivas de motores de búsqueda, sino que muchas responden a la naturaleza de la propia Web, lo cual nos permite afirmar que el empleo de métodos alternativos para la obtención de datos a gran escala no está exento de estos problemas. Para Li (2003), debido a lo anteriormente expuesto, el empleo de motores de búsqueda para la obtención de datos sólo puede proporcionar indicaciones generales, pero no conclusiones definitivas. En cualquier caso, dada la naturaleza de la Web, consideramos que el objetivo de la investigación difícilmente puede pretender generar un conocimiento que no esté sujeto a una continua revisión. En todo caso, como apunta Vaughan (2004a), en toda investigación se debe añadir la advertencia de que lo resultados son válidos y robustos desde el punto de vista del motor de búsqueda empleado y para el momento del tiempo en que se obtuvieron.

Concluimos este apartado subrayando que, al margen de las limitaciones, los motores de búsqueda proporcionan datos accesibles a todo el mundo de manera gratuita y constituyen la mejor aproximación al estudio de la Web en su conjunto, ya que son los únicos con capacidad para indexar amplias partes de la misma. Es importante recordar que los resultados ofrecidos por los buscadores no constituyen una medición exacta de la Web, sino una aproximación. A estas alturas, tras analizar las características de la Web y de las herramientas de que disponemos para acceder a ella, podemos afirmar que una medición exacta es inviable. Los motores de búsqueda se han desarrollado y mejorado considerablemente a lo largo de los últimos años; por ejemplo, en la capacidad para actualizar su base de datos, así como para indexar ficheros con distintos formatos. El ámbito en el que el desconocimiento es mayor es el de la formación de los *rankings* de resultados, los cuales constituyen el producto final de una consulta realizada por un usuario.

3.3.3.6. Empleo de las APIs para la obtención de datos

Las siglas API significan en inglés *Application Programming Interface*, expresión que puede traducirse al español como *interfaz de programación de aplicaciones*. Se trata de un conjunto de funciones y procedimientos que ciertos servicios de Internet ponen a disposición de los programadores y demás usuarios, con el fin de permitir un acceso automatizado a sus bases de datos. Su empleo permite realizar una recogida sistemática de datos, con un considerable ahorro de tiempo y recursos.

Las APIs permiten combinar distintas bases de datos para la creación de nuevos servicios, llamados *mashups*. Este es uno de los fenómenos más característicos de la Web 2.0, habiendo alcanzado una gran popularidad en los últimos años. El empleo de APIs y *mashups* ofrece un gran potencial para la investigación webmétrica ya que abre la posibilidad al acceso sistemático a grandes cantidades de datos en formato XML, lo cual permite trabajar con ellos directamente en programas estadísticos. La Tabla 3.1 incluye algunas de las APIs disponibles con mayor interés para la investigación webmétrica.

En relación con el empleo de las APIs, generalmente existen limitaciones al número de consultas que se pueden realizar por día, por lo general unas mil en el caso de los principales motores de búsqueda. Esta limitación, de algún modo, está también presente en las consultas a través de la interfaz web, ya que, en la práctica, solamente se puede acceder a alrededor de mil de los resultados estimados. Esta cuestión plantea problemas de cara al examen de los sitios que cumplen con los requisitos de una consulta, por ejemplo, de cara a tomar una muestra aleatoria de los mismos para llevar a cabo un análisis de contenidos. Por último, se han detectado diferencias entre los resultados obtenidos a través de la API y los que provienen directamente del interfaz web del motor de búsqueda (Thelwall, 2004; 2009). Mayr y Tosques (2005) investigaron este hecho para el caso de Google.

Existen diversos programas que emplean APIs para la obtención de los datos de investigación. Un ejemplo es el *LexiURL Searcher* (http://lexiurl.wlv.ac.uk/), desarrollado por Mike Thelwall en la University of Wolverhampton.

De cara a obtener datos, una alternativa que sugerimos es el empleo de fuentes RSS o Atom, las cuales permiten al investigador suscribirse a contenidos que recibe en un lector de manera automática conforme se van generando.

Tabla 3.1. Listado de algunas de las APIs más importantes facilitadas por los principales motores de búsqueda

API de servicios de búsqueda	Descripción	URL
Google SOAP Search	Servicios de búsqueda. A partir del 5 de diciembre de 2006 Google dejó de emitir claves de usuario para esta API.	http://code.google.com/apis/soapsearch/
Google AJAX Search	Servicios de búsqueda.	http://code.google.com/apis/ajaxsearch/web.html
Google Research Search	Servicios de búsqueda para investigadores.	http://research.google.com/university/search/
Yahoo! Search	Servicios de búsqueda.	http://developer.yahoo.com/search/web/

Yahoo! Boss	Servicios de búsqueda.	http://developer.yahoo.com/search/boss/
Yahoo! Site Explorer	Servicios de búsqueda de enlaces.	http://developer.yahoo.com/search/siteexplorer/
Bing	Servicios de búsqueda.	http://www.bing.com/developers/

Por último, acabaremos mencionando la existencia de otro tipo de programas, llamados *scrapers*, que permiten el procesamiento de los datos obtenidos a través de la interfaz web de los motores de búsqueda. En estos casos no se consulta directamente la base de datos del buscador, sino que lo que se hace es emplear un *software* que permite extraer la información relevante de las consultas efectuadas a través de la página web del motor de búsqueda. Algunos trabajos que han empleado este procedimiento son: Larson (1996), Heimeriks, Horlesberger y van den Besselaar, (2003), Heimeriks y van den Besselaar (2006). Sus posibilidades son, en todo caso, mucho más limitadas que las derivadas del empleo de APIs.

3.3.3.7. Términos de consulta

Cada uno de los tres grandes buscadores proporciona determinadas funciones para recabar información sobre enlaces en la Web. Este tipo de información, como hemos explicado en el apartado anterior, puede obtenerse bien a través del interfaz web del buscador o empleando una API. Cada uno de los buscadores presenta diferencias importantes en cuanto al tipo de operadores que se pueden emplear, bien a través de su página web o de su API.

Los principales operadores empleados para la obtención de información sobre enlaces son:

- *Link* – Es un operador que tanto en Yahoo! como en Google permite obtener las páginas que enlazan a una página específica, a una URL concreta.

- *Linkdomain* – Permite recuperar las páginas que enlazan a un dominio en particular, no únicamente a una página específica.

- *Linkfromdomain* – Este operador, que sólo se encuentra disponible en Bing, permite descubrir las páginas que están enlazadas desde un dominio en particular.

- *Site* – El operador *site* restringe la búsqueda a páginas que se encuentran dentro de un sitio web o dentro de extensiones determinadas. El empleo combinado de la función *-site* junto a los operadores *link* o *linkdomain* permite conocer el número de enlaces externos que apuntan a una página o sitio, esto es, excluyendo aquellos que provienen del mismo sitio web (auto-referencias).

Como ya se apuntó, las posibilidades de emplear ciertas funciones en los buscadores cambia continuamente; por ejemplo, Stuart (2008) señala que Live Search (Microsoft) canceló el servicio de los operadores *link* y *linkdomain* en 2007 (Live Search, 2007).

Se han detectado importantes diferencias entre los resultados que arrojan los diferentes buscadores, lo cual puede deberse a razones tales como la cobertura de la base de datos empleada (Vaughan y Zhang, 2007) o a variaciones en las estimaciones realizadas. Aunque la cobertura de los distintos buscadores difiere, en ocasiones, la elección del más apropiado para la investigación no depende de este criterio, sino, por ejemplo, del tipo de información que necesitemos utilizar. Si bien se podría considerar la combinación de los resultados de varios motores de búsqueda, en ocasiones es preferible emplear un único buscador que proporcione resultados homogéneos con el fin de poder efectuar comparaciones entre las distintas medidas empleadas, especialmente cuando los resultados difieren

significativamente entre ellos.

Resulta fundamental conocer las funciones avanzadas de los motores de búsqueda para diseñar consultas que sean lo más precisas posibles. Las consultas más usuales empleadas en la investigación webmétrica, junto con el tipo de enlaces que proporcionan, son:

- *linkdomain:abc.com -site:abc.com*: enlaces externos que apuntan a un dominio. Por enlaces externos se entiende que no proceden de páginas dentro del mismo dominio.

- *link:abc.com -site:abc.com*: enlaces externos que apuntan a una página web determinada.

- *linkdomain:abc.com site:xyz.com*: enlaces de páginas dentro del dominio *xyz.com* que apuntan a páginas dentro del dominio *abc.com*.

- *link:abc.com site:xyz.com*: enlaces de páginas dentro del dominio *xyz.com* que apuntan a la página *abc.com*.

- *linkfromdomain:abc.com -site:abc.com*: enlaces incluidos dentro de un dominio en particular y que no enlazan a páginas de ese mismo dominio, sino a páginas externas.

La Tabla 3.2 presenta las principales variables webmétricas empleadas en investigación, junto con los términos de consulta empleados en los motores de búsqueda: Google, Bing, Yahoo! y Site Explorer (Yahoo!). También se incluye una columna con los resultados de efectuar la consulta para el dominio indicado en la tabla: ugr.es (correspondiente a la Universidad de Granada) y us.es (perteneciente a la Universidad de Sevilla). Cuando algunas de las funciones no están operativas en algunos buscadores, se indica en la tabla. Site Explorer (https://siteexplorer.search.yahoo.com) constituye la herramienta a través de la cual Yahoo! proporciona información acerca de la presencia *online* de los sitios web.

Permite obtener datos sobre el número de páginas indexadas por Yahoo! en un sitio determinado, así como el número de páginas que enlazan una página o sitio web. Los datos obtenidos en la búsqueda se pueden exportar a un fichero con formato *tab-separated value* (TSV), hasta un límite de mil resultados por consulta. También es posible conseguir los datos a través de la API que ofrece el servicio. A pesar de ser una herramienta especializada en información sobre la estructura de la Web, las consultas complejas (por ejemplo, las referidas a co-enlaces) sólo pueden realizarse a través del motor de búsqueda general.

Tabla 3.2. Términos de búsqueda de información webmétrica aplicados en los principales motores de búsqueda (a 26 de enero de 2010)

Motor de búsqueda	Términos de consulta	Resultados
Número de páginas en un sitio web (empleando el dominio completo)		
Google	site:www.ugr.es	282.000
Bing (Microsoft)	site:www.ugr.es	139.000
Site Explorer (Yahoo)	http://www.ugr.es	132.930
No permitido: Yahoo!		
Número de páginas en un sitio web (empleando el dominio parcial)		
Google	site:ugr.es	1.480.000
Bing (Microsoft)	site:ugr.es	466.000
Site Explorer (Yahoo)	ugr.es	421.701
No permitido: Yahoo!		
Número de páginas que enlazan a un sitio web excepto aquellos enlaces que provienen del propio sitio web (sólo se tienen en cuenta los enlaces externos)		
Yahoo!	linkdomain:ugr.es -site:ugr.es	226.000

Site Explorer (Yahoo)	ugr.es (opciones: inlinks except from this domain to entire site)	231.003
No permitido: Google y Bing (Microsoft)		
Número de páginas que enlazan a un sitio web desde un sitio web en particular		
Yahoo!	linkdomain:us.es site:ugr.es	367
No permitido: Google, Bing (Microsoft) y Site Explorer (Yahoo)		
Número de páginas que enlazan a un sitio web (tanto enlaces internos como externos)		
Site Explorer (Yahoo)	ugr.es (opciones: inlinks from all pages to entire site)	427.755
No permitido: Google, Bing (Microsoft) y Yahoo!		
Número de páginas que enlazan a una página web		
Google Nota: sólo ofrece una muestra de todos los enlaces disponibles en su base de datos.	link:www.ugr.es	3.510
Site Explorer (Yahoo)	ugr.es (opciones: inlinks from all pages to only this URL)	223
No permitido: Bing (Microsoft) y Yahoo!		
Número de páginas que enlazan a una página web excepto aquellos enlaces que provienen del propio sitio web (*link:www.rbs.com -site:rbs.com)*		
Site Explorer (Yahoo)	ugr.es (opciones: inlinks except from this domain to only this URL)	151
No permitido: Google, Bing (Microsoft) y Yahoo!		
Número de páginas enlazadas desde un sitio web		
Bing (Microsoft)	linkfromdomain:ugr.es	1.020.000

Nota: los resultados parecen no ser consistentes.		
No permitido: Google, Yahoo!, Site Explorer (Yahoo)		
Número de páginas enlazadas desde un sitio web excepto lo que enlazan al propio sitio web		
Bing (Microsoft) Nota: los resultados parecen no ser consistentes.	linkfromdomain:ugr.es -site:ugr.es	516.000
No permitido: Google, Yahoo!, Site Explorer (Yahoo)		
Co-enlaces entrantes (de forma genérica, co-enlaces)		
Yahoo!	linkdomain:ugr.es linkdomain:us.es -site:ugr.es -site:us.es	21.900
Nota: a partir de mediados de 2009, se ha observado que tanto la versión internacional de Yahoo! como la API de Yahoo pueden ofrecer resultados erróneos. Las versiones nacionales parecen proporcionar por el momento resultados fiables (por ejemplo, Yahoo! UK, Yahoo! España o Yahoo! Canada).		
No permitido: Google, Bing (Microsoft), Site Explorer (Yahoo)		
Co-enlaces salientes *linkfromdomain:rbs.com linkfromdomain:barclays.com*		
No permitido: Google, Bing (Microsoft), Yahoo, Site Explorer (Yahoo)		
La siguiente tabla ha adaptado, actualizado y completado información incluida en: Search Engine Queries for Webometrics, Statistical Cybermetrics Research Group, University of Wolverhampton, la cual se encuentra disponible en: http://cybermetrics.wlv.ac.uk/QueriesForWebometrics.htm (consultado el 26 de enero de 2010). Las búsquedas incluidas en los ejemplos se realizaron el 26 de enero de 2010.		

Para los ejemplos se han tomado como referencia los sitios web de la Universidad de Granada (www.ugr.es) y de la Universidad de Sevilla (www.us.es). En el caso de emplear Site Explorer se detallan las opciones que hay que seleccionar para efectuar la consulta en cuestión.

Las APIs de los anteriores buscadores permiten acceder al mismo tipo de información, empleando también los citados operadores. Sin embargo, en determinados casos, se han observado variaciones significativas entre los datos obtenidos a través de este sistema y los provenientes de una consulta manual en la interfaz del motor de búsqueda (Mayr y Tosques, 2005).

Por lo general, al realizar las consultas, se recomienda eliminar la parte *www.* de las direcciones empleadas al objeto de incluir en la misma los subdominios, tales como *vpn.ugr.es*. Esto es especialmente importante en sitios web muy grandes. Las posibles búsquedas incluidas en la Tabla 3.2 no agotan todas las posibilidades de los motores de búsqueda, pero sí son las más habituales.

La consulta de información de enlaces basada en Google presenta limitaciones significativas (Barjak y Thelwall, 2008), pese a tratarse del motor de búsqueda con una mayor grado de cobertura de la Web. Por ejemplo, no permite combinar distintos operadores para refinar las consultas. Además, se ha observado que los resultados proporcionados incluyen únicamente una parte de la información contenida en su base de datos (Google, 2009). Vaughan (2004a) comparó los resultados obtenidos por diversos buscadores, señalando que éstos eran bastante consistentes entre sí, por lo que no mostraba ninguna preferencia clara hacia ninguno de ellos.

Cuando nos centramos en las consultas de otro tipo de información, no basada en enlaces sino en palabras clave, Google sí constituye una de las mejores opciones. Así, Thelwall (2008b) señala a Google como la mejor opción para el recuento del número de páginas referidas a palabras clave, mientras que Yahoo! representaría la opción preferida de cara a la obtención de otro tipo de información. Thelwall (2008a), a partir de un análisis cuantitativo de los resultados proporcionados por

Google, Live Search y Yahoo!, subraya que hay una gran consistencia entre los mismos, indicando la viabilidad de utilizar todos ellos para la investigación.

3.3.3.8. Preparación de los datos de enlaces para la investigación

La preparación de los datos obtenidos es fundamental para asegurar la validez de la investigación. En la mayoría de los casos, dadas las restricciones para acceder a todos los resultados existentes en las bases de datos de los motores de búsqueda, sólo se puede hacer un examen completo de aproximadamente mil páginas. En general se ha comprobado que es muy complicado filtrar las anomalías existentes en los datos y, en ocasiones, no compensa el tremendo esfuerzo que supone (Björneborn y Ingwersen, 2001). En estudios con un amplio alcance este filtrado puede ser claramente insuficiente ya que podemos encontrarnos con miles o cientos de miles de páginas que satisfacen los requisitos de la consulta. Por consiguiente, el mejor modo de enfrentarnos al problema es intentar precisar y delimitar la consulta realizada con el objeto de que los datos que obtengamos sean lo más precisos posibles.

Una de las alternativas que se han planteado para mejorar la calidad de los datos es el empleo de varios buscadores simultáneamente (Thelwall, Vaughan y Björneborn, 2005). Vaughan (2004a) aplicó esta alternativa con éxito, si bien no es siempre cierto que una combinación de los resultados obtenidos en distintos buscadores pueda mejorar los resultados. Actualmente, las divergencias existentes entre los buscadores complica mucho esta posibilidad. Otra de las alternativas más viables es la de emplear las funciones avanzadas de los buscadores al objeto de filtrar y delimitar al máximo los resultados obtenidos con el fin de obtener datos de mejor calidad y más precisos para la investigación. Esto se puede conseguir empleando operadores booleanos, siempre que sea posible, para, por ejemplo, eliminar páginas que contengan una determinada palabra clave o incluir páginas que sólo contengan dicha palabra. Vaughan y You (2008) llevaron a cabo, por ejemplo, un análisis de co-enlaces en el que restringían los resultados obtenidos

mediante el empleo de determinadas palabras clave para estudiar sectores concretos dentro de una industria más amplia. También se pueden restringir los resultados a dominios, lenguas, tipos de archivos, cadenas de palabras o, en ciertos casos y de manera limitada, a periodos de tiempo concretos.

Uno de los problemas más importantes que existen en la investigación basada en enlaces es la existencia de duplicados, esto es, páginas o sitios web que repiten un enlace o enlaces de manera automática, por ejemplo, al incluir una lista de "favoritos". Esta repetición no añade información significativa sobre la intensidad de la conexión entre sitios web y debería, por tanto, ser mitigada. En ocasiones son los propios buscadores los que eliminan estos duplicados, pero el tratamiento que realizan de ellos no es transparente. Si bien lo que se pretende con el filtrado de los duplicados es aumentar la calidad y validez de los datos (Thelwall, 2004), no siempre es posible realizarlo manualmente, especialmente cuando éstos proceden de motores de búsqueda, debido a las limitaciones para acceder a todos los resultados.

En muchos casos, el mejor modo de afrontar estos problemas es hacer una comprobación exploratoria de los problemas que pueden presentar los datos, partiendo de la información de que disponemos, por limitada que sea, y proceder de acuerdo con el juicio del investigador. En todo caso, se deben reconocer estas limitaciones en la investigación e interpretar los resultados de acuerdo con ellas.

3.4. El análisis de enlaces

Thelwall (2009) distingue en su libro *Introduction to Webometrics* dos tipos fundamentales de análisis basados en enlaces: el análisis de impacto de enlaces (*link impact assessments*) y la representación gráfica de relaciones a través de enlaces (*link relationship mapping*). Ambas perspectivas se han empleado en la investigación empírica de esta tesis, por lo que a lo largo de esta sección se pretende revisar los aspectos fundamentales de cada una de ellas.

Al margen de la investigación basada en hiperenlaces, existen otras perspectivas basadas, por ejemplo, en el análisis de impacto en la Web (*Web Impact Analysis*), que consiste en estudiar el impacto de ideas, conceptos, marcas, organizaciones, empresas, etc., a partir del recuento y análisis de las páginas web proporcionadas por los motores de búsqueda al formular una consulta. Un buen ejemplo es el reciente trabajo de Vaughan, Tang y Du (2010), en el que se emplea este tipo de análisis para elaborar distintos perfiles de empresas. Al margen, también está alcanzando cada vez una mayor importancia el estudio de distintos servicios y herramientas de la Web 2.0.

3.4.1. La investigación basada en enlaces: evidencia empírica y métodos

Nuestra investigación y la revisión de la literatura previa se centran en el estudio de la estructura de la Web, entendida esta como la red de enlaces que interconectan las distintas páginas o sitios web. La Web, como se ha expuesto anteriormente, constituye una red en la que se identifican una serie de nodos y conexiones entre ellos. Los nodos pueden estar representados por distintos tipos de elementos, en función del grado de agregación que consideremos, por ejemplo, partes de páginas web, páginas web, sitios web, etc. Las conexiones entre dichos nodos están representadas por los hiperenlaces, que igualmente pueden agregarse según convenga para el análisis. Anteriormente se ha expuesto la naturaleza y la importancia de los hiperenlaces, tanto desde un punto de vista histórico como más puramente conceptual y filosófico.

Como apunta Holmberg (2009), si se tiene en cuenta la naturaleza en apariencia caótica de la Web, es muy significativo el hecho de que, en determinados ámbitos, los estudios webmétricos hayan encontrado correlaciones significativas entre enlaces recibidos y otras variables ajenas a la Web y de reconocido valor; por ejemplo, la calidad de una publicación científica o determinadas variables financieras.

Mucha de la investigación realizada ha tenido un carácter exploratorio, buscando poner de manifiesto relaciones entre fenómenos observados en la Web y fenómenos propios del mundo físico. Por ejemplo, se han analizado las relaciones entre enlaces recibidos por sitios web de universidades con los niveles académicos alcanzados por las mismas (Thelwall, 2001a; Smith y Thelwall, 2002; Tang y Thelwall, 2003), enlaces a sitios web de facultades con su productividad investigadora (Chu, He y Thelwall, 2002), o enlaces a sitios web de revistas con la calidad de las mismas (Vaughan y Hysen, 2002; Vaughan y Thelwall, 2003). Otros trabajos han empleado los enlaces para medir la visibilidad y el impacto de un grupo de sitios web (Ingwersen, 1998; Vreeland, 2000; Chu, He y Thelwall, 2002), para analizar patrones de comunicación informal entre académicos (Li, 2003; Wilkinson et al., 2003; Tang y Thelwall, 2004), para analizar la relación con patrones geográficos o distancias entre organizaciones (Thelwall, 2002b), o para identificar áreas académicas con un mayor impacto en la Web (Thelwall et al, 2003). Otras investigaciones han estudiado países en concreto. Por ejemplo, Björneborn (2004) estudió los sitios web de universidades en el Reino Unido o Baeza-Yates, Castillo y López (2005) analizaron las características de la Web en España.

Por el momento no han sido muchos los investigadores que hayan prestado atención a la estructura de la Web en relación con el ámbito empresarial. Los trabajos se han centrado en el empleo de los enlaces recibidos como indicadores del desempeño empresarial y en el análisis de la situación competitiva en determinados sectores. Algunos de los trabajos más destacados son: Thelwall, (2001e); Vaughan (2004a); Vaughan y Wu (2004); Vaughan y You (2005; 2006; 2008).

La investigación sobre sitios o páginas web de empresas busca identificar relaciones entre variables web y variables económicas, así como patrones en las relaciones existentes entre empresas, los cuales permanecen ocultos a simple vista. El objetivo es determinar hasta qué punto el análisis de los hiperenlaces puede representar una nueva fuente de información sobre empresas y sus relaciones entre sí; por ejemplo, relaciones de competencia, alianzas, vínculos no evidentes entre organizaciones, relaciones con diversos actores sociales como la

administración pública o la universidad, etc. Investigaciones en la misma línea se han realizado para el estudio de gobiernos regionales con presencia en la Web o para el análisis de cómo la Web refleja las transferencias de conocimiento entre la universidad, las empresas y el sector público en su conjunto. Toda esta serie de trabajos se desarrollará con mayor detalle en los Apartados 3.5.2 y 3.5.3, si bien no han sido tan numerosos como los llevados a cabo en el campo de la investigación sobre la ciencia (Petricek et al., 2006).

La falta de interés por el estudio de sitios web de empresas por parte de investigadores que aplican técnicas webmétricas se puede explicar por la existencia de un prejuicio que Berners-Lee (1999: 100-101) ya identifica en los orígenes de la Web: "Los académicos que habían usado Internet desde el principio tenían la sensación de que era un espacio abierto, libre y puro para su propio uso, y les preocupaba que el generoso espacio de información del que habían disfrutado para esos correctos fines se convirtiese ahora en algo inasequible, lleno de correo basura y publicidad. Ciertas personas pensaban que el material comercial podía contaminar el Web. Yo no estaba muy de acuerdo con este punto de vista. El Web estaba diseñado como un medio universal. Un vínculo de hipertexto tenía que poder apuntar a cualquier cosa. La información que se incluyera con fines comerciales no podía ser excluida."

En este sentido, la Webmetría tiende a convertirse en una disciplina que proporciona un conjunto de metodologías transversales que permiten analizar cualquier fenómeno que tenga su reflejo en la Web. Por tanto, su componente interdisciplinar es fundamental (Chua y Yang, 2008).

Thelwall (2009) apunta dos tipos fundamentales de investigación webmétrica:

- Análisis de impacto de enlaces, que se basa en el análisis de las URL que incluyen un enlace a una página o a un sitio web determinado.
- Representación gráfica de relaciones a través de enlaces, donde podemos incluir el análisis de co-enlaces, que se centra en el estudio de páginas web que enlazan al mismo tiempo a dos páginas o sitios que son objeto de la

investigación. Dicho análisis puede hacerse también basándose en la trama de enlaces directos que vinculan un conjunto de sitios web o basándose en espacios web relacionados, esto es, dos páginas web distintas que comparten una misma referencia, es decir, que enlazan simultáneamente a un mismo sitio o página web.

Al margen de estos dos tipos de análisis de carácter cuantitativo, es recomendable completar la investigación con un trabajo cualitativo que permita comprobar, por ejemplo, si las páginas web en estudio contienen información significativa vinculada al fenómeno que investigamos. En este sentido, de poco valdría identificar una correlación positiva entre el número de enlaces a sitios web de empresas y sus resultados financieros si se desconoce qué tipo de información empresarial es la que proporcionan las páginas web que enlazan. Algunos estudios que han abordado esta perspectiva son, por ejemplo: Kim (2000); Thelwall, Harries y Wilkinson (2003); Thelwall y Harries (2004); Bar-Ilan (2005); Vaughan, Gao y Kipp (2006); y, Vaughan, Kipp y Gao (2007a; 2007b).

3.4.2. La hipótesis de proporcionalidad

Algunas de las relaciones puestas de relieve en los trabajos anteriores suscitan la siguiente pregunta: ¿qué idea sirve de base para vincular los enlaces recibidos por una página o sitio web con un fenómeno determinado?

Thelwall y Wilkinson (2008: 94) intentan dar una respuesta general a esta pregunta, planteando la *hipótesis de proporcionalidad*, en la que, según ellos, se basan la mayoría de trabajos que vinculan fenómenos *offline* con su impacto en la Web medido a través del número de enlaces. Esta hipótesis establece que un mayor número de enlaces o co-enlaces constituyen un indicador de la envergadura del fenómeno subyacente, que se pretende inferir a través del recuento de dichos enlaces.

Esta hipótesis de proporcionalidad se ha aplicado a la actividad académica e investigadora, a la similitud entre documentos o empresas dentro de un sector determinado, al impacto de un concepto o término, etc. También, directamente vinculado al objetivo del presente trabajo, la hipótesis de proporcionalidad toma forma en la idea de que el número de enlaces que reciben los sitios web de empresas está relacionado con la dimensión de la empresa o con su rendimiento financiero.

En cualquier caso, Thelwall y Wilkinson (2008) plantean una cuestión de gran interés que permite reflexionar sobre la naturaleza de la investigación webmétrica y de cómo está condicionada por el modo de captación de los datos en los que se basa. La investigación está muy influida por el modo en que los motores de búsqueda operan, por lo que se deberían emplear las mismas hipótesis en las que estos se basan con el objeto de que los resultados sean más robustos (por ejemplo, en línea con el funcionamiento del *PageRank* de Google, deberíamos suponer que un mayor número de enlaces recibidos constituye un indicador de que las páginas enlazadas son más relevantes).

Resumiendo, la hipótesis de proporcionalidad establece que el recuento de enlaces o co-enlaces está vinculado a la magnitud o intensidad del fenómeno investigado. Por supuesto, esto no quiere decir que exista una relación perfecta, sino que en términos generales, está relación tiende a verificarse.

3.4.3. Medidas de similitud en la Web

Mientras que los co-enlaces constituyen una medida indirecta de la relación entre dos páginas o sitios web, los enlaces directos entre páginas constituyen un evidente indicador de similitud. Siendo así, cabría preguntarse cuál de ellos proporciona una medida más adecuada. Thelwall y Wilkinson (2004) tratan de responder a esta pregunta y concluyen que el empleo de una combinación de enlaces, co-enlaces y espacios web relacionados permite identificar sitios similares. Sin embargo, su

combinación no mejora sustancialmente los resultados obtenidos al utilizar únicamente enlaces directos entre páginas. Los co-enlaces y espacios web relacionados serían especialmente útiles para ampliar el alcance del campo de estudio, ya que los enlaces directos, a estos efectos, son más escasos. Desde el punto de vista de la informática (Dean y Henzinger, 1999; Wang y Kitsuregawa, 2001; Hou y Zhang, 2003), los co-enlaces han sido empleados en la recuperación de información confirmando su utilidad para descubrir páginas similares o relacionadas.

Para Vaughan (2006), el análisis de co-enlaces para el estudio de relaciones entre sitios web es más prometedor que el análisis de enlaces directos, debido a que son potencialmente más robustos al ser un tipo de información mucho más difícil de distorsionar o manipular. En todo caso, siempre es mejor considerar ambas medidas como complementarias siempre que sea posible.

La cuestión acerca de qué medida de similitud utilizar al abordar el mundo de la empresa es especialmente importante. Cuando se hace referencia a empresas similares, especialmente dentro de un mismo sector, por lo general se trata de entidades competidoras, que comparten una serie de rasgos en común, como tener una oferta de servicios y productos similares. También puede tratarse de empresas asociadas, aliadas o con algún otro tipo de vínculo, si el conjunto de empresas que tomamos como referencia es heterogéneo, es decir, incluye empresas de distintos sectores. En los casos en los que la similitud adquiere la dimensión de relación competitiva, el empleo de co-enlaces es especialmente pertinente ya que los enlaces directos entre páginas web de empresas rivales son extremadamente escasos (Shaw, 2001; Vaughan, Gao y Kipp, 2006).

3.4.4. El análisis de impacto de enlaces

El análisis de impacto de enlaces (*link impact assessment*) (Thelwall, 2004; 2009) consiste en analizar las URL que incluyen un enlace a una página o a un sitio web determinado. Este tipo de análisis se enfrenta a problemas como son: el empleo de

varios dominios por una misma organización, dominios que están redireccionados a otros, etc. Por el carácter extensivo de este tipo de investigación, dado que los enlaces a una página web determinada pueden provenir de cualquier parte de la Web, los motores de búsqueda son la única opción factible para la obtención de los datos. Ello hace que las limitaciones propias de los motores de búsqueda sean relevantes en este tipo de trabajos. La investigación basada en co-enlaces que se incluye en el apartado siguiente, se puede considerar como un tipo de análisis de impacto de enlaces, si bien con un grado mayor de complejidad.

A continuación incluimos un ejemplo sencillo de este tipo de análisis aplicado a las páginas web de los candidatos a las últimas elecciones presidenciales en los Estados Unidos. Las consultas que realizaríamos en Yahoo!, motor de búsqueda con la mayor gama de funciones avanzadas en estos momento, serían:

- *linkdomain:barackobama.com -site:barackobama.com*
- *linkdomain:johnmccain.com -site:johnmccain.com*

Con estas consultas obtendríamos el número de enlaces que apuntan a cualquiera de las páginas alojadas dentro del dominio señalado, exceptuando los enlaces procedentes de páginas del propio sitio web. Este procedimiento permite obtener únicamente enlaces externos que son los que mejor permiten evaluar la relevancia de un sitio, ya que los internos pueden ser fácilmente manipulados por la propia organización. Es habitual, como han hecho muchos de los trabajos en esta área, intentar poner en relación la presencia en la Web, medida a través de los enlaces, con variables tales como, en este hipotético caso, el resultado en las elecciones, o en otros, por ejemplo, variables sobre la calidad de trabajos científicos y de universidades o variables económicas.

3.4.5. Representación gráfica de relaciones a través de enlaces: el análisis de co-enlaces

La representación gráfica de relaciones a través de enlaces (*link relationship mapping*) es un modo de visualizar las redes que se generan a través de los enlaces que vinculan las distintas páginas y sitios web (Thelwall, 2009). Un modo de hacerlo es estudiando los enlaces directos que vinculan las páginas y sitios entre sí; otro sería mediante el análisis de co-enlaces. Los co-enlaces son páginas web que enlazan al mismo tiempo a dos páginas o sitios que son objeto de la investigación. Se trata de un procedimiento especialmente indicado para el estudio de páginas o sitios web que, por su tipología, no establecen enlaces entre sí de forma habitual. En concreto es el caso de las páginas web de empresas, que evitan enlazar a la competencia con el fin de no desviar tráfico de visitas hacia ella (Shaw, 2001; Vaughan, Gao y Kipp, 2006). Las vinculaciones entre empresas pueden investigarse, sin embargo, mediante el empleo de co-enlaces. Este procedimiento es especialmente útil cuando se trata de empresas de un mismo sector, en cuyo caso priman las relaciones de competencia. Así se han analizado, por ejemplo, la industria de las telecomunicaciones en el mercado global y en el mercado chino en particular (Vaughan y You, 2006) y el subsector de la industria de las telecomunicaciones dedicado a la tecnología WIMAX (Vaughan y You, 2008).

En los siguientes subapartados se abordan algunas cuestiones específicas del análisis de co-enlaces.

3.4.5.1. De las co-citaciones a los co-enlaces

La investigación en co-enlaces surge como una aplicación a la Web de las técnicas de investigación basadas en co-citaciones, que se vienen empleando normalmente para elaborar mapas de conocimiento de la producción científica. Su origen se sitúa en 1973 a partir de las aportaciones de Small (1973) en Estados Unidos y de Marshakova (1973) en la Unión Soviética, que sientan las bases del empleo de co-citaciones como instrumento bibliométrico de investigación. Small y Sweeney

(1985), en los años 80, continuaron desarrollando las anteriores aportaciones. Como explican Faba Pérez, Guerrero Bote y Moya Anegón (2004a: 69), "[l]a premisa fundamental del análisis de cocitas es que cuanto mayor sea la cantidad de veces que dos documentos son citados conjuntamente, mayor es la probabilidad de que su contenido esté relacionado (Moya-Anegón, Jiménez-Contreras y Moneda-Corrochano, 1998)". El objetivo de este tipo de análisis es descubrir la estructura intelectual de las distintas disciplinas académicas, analizando las vinculaciones, principalmente, a nivel de documentos o de autores (Milman, 1994). Las co-citaciones se han empleado también para estudiar las redes de colaboración de investigadores. En esta área, destacan los trabajos de White, Wellman y Nazer (2004) y de Zuccala (2006b). Kretschmer y Aguillo (2004) han aplicado técnicas de análisis de redes al estudio de los patrones de colaboración *offline* y *online* de un grupo de investigadores, encontrando que ambas estructuras son similares.

Es precisamente en la traslación del concepto de co-cita o co-citación a la Web donde surge el concepto de co-enlace. Un co-enlace se genera cuando un sitio web A enlaza dos sitios web, B y C. En este caso, se dice que B y C están co-enlazados por A. Los co-enlaces se producen generalmente cuando dos páginas o sitios web están relacionados entre sí, por lo que su análisis permite determinar las vinculaciones entre ellos. El análisis de co-enlaces descansa sobre la hipótesis de similitud, es decir, se constituye en una forma de medir la semejanza o proximidad entre una serie de elementos en función de las relaciones que establecen terceros. Los análisis de co-enlaces difieren de los estudios que emplean enlaces directos en que los primeros constituyen una medida indirecta de similitud. Esto permite estudiar relaciones entre grupos de elementos que, por determinadas razones, no establecen conexiones directas entre ellos, característica por la que es un análisis muy indicado para los sitios web de empresas.

Si bien nos vamos a centrar en el análisis de co-enlaces (entendido como co-enlaces entrantes o *co-inlinks*), es posible realizar también un análisis de co-enlaces salientes (*co-outlinks*) o espacios web relacionados. La figura bibliométrica análoga es la del análisis de documentos relacionados (en inglés, *Bibliographic coupling* (Faba Pérez, Guerrero Bote y Moya Anegón, 2004a) o *Bibliometric*

coupling (Thelwall y Wilkinson, 2004)), popularizado por Kessler (1963). Como apunta Faba Pérez, Guerrero Bote y Moya Anegón (2004a: 70), "Kessler introdujo el término para señalar el hecho de que cuando dos documentos tienen, al menos, una referencia común existe entre ellos un lazo o enlace bibliográfico". Compartir referencias comunes, esto es, cuando dos documentos distintos citan conjuntamente a un tercero, se considera un indicador de proximidad temática o semejanza entre ambos documentos.

Existen diferencias considerables en cuanto a las posibilidades de obtención de datos en función de si se trata de un estudio de co-enlaces entrantes (*co-inlinks*) o salientes (*co-outlinks*). Nos referimos a continuación a este segundo caso. En primer lugar, si se contempla la opción de obtener los datos de motores de búsqueda, las limitaciones de funciones avanzadas de consulta en este campo prácticamente impiden la realización del trabajo. Actualmente el único buscador que permite conocer los enlaces que salen de una página o sitio web es Bing (Microsoft). Sin embargo, los resultados que proporciona no son consistentes y además no permite el empleo de operadores booleanos para conformar una consulta acorde con los objetivos de la investigación, es decir, no permite la recuperación de co-enlaces salientes. En segundo lugar, cabe considerar la utilización de un programa de *software* que rastree todos los enlaces existentes en un conjunto cerrado de páginas o sitios web en estudio. Esta opción sí sería válida ya que dicho programa únicamente debería recuperar los enlaces contenidos en las páginas o sitios web estudiados, lo cual, aún en el caso de que fuera un número elevado, sería técnicamente factible con un esfuerzo razonable. Sin embargo, esta opción no es posible para el análisis de co-enlaces entrantes ya que en estos estudios se analizan todos los enlaces que apuntan a un conjunto cerrado de sitios web, lo cual implica examinar todos y cada uno de las páginas web existentes ya que cada una de ellas puede establecer un enlace en potencia. Este esfuerzo titánico probablemente no está siquiera al alcance de los motores de búsqueda más potentes, cuanto menos al de un programa informático empleado a título particular, como es el que utilizaría un investigador. Ello implica que para este tipo de investigación se deba acudir a los motores de búsqueda, los cuales, a pesar de sus limitaciones, son los únicos con capacidad para indexar amplias partes de la Web.

Actualmente, Yahoo! es el único buscador que permite realizar este tipo de consultas.

3.4.5.2. Diferencias entre co-citaciones y co-enlaces

La investigación bibliométrica basada en co-citaciones, por el ámbito relativamente cerrado y bien definido al que se aplica, permite el análisis de variables que están claramente identificadas (por ejemplo, se pueden tener en cuenta sólo a primeros autores, sólo co-citaciones, co-citaciones y co-autorías, etc.). Este hecho ya apunta una de las principales diferencias en relación al objeto de estudio: la información contenida en la Web no está estructurada a diferencia de lo que ocurre con la producción académica.

Sin embargo, es pertinente establecer las analogías y diferencias principales con el objeto de poder determinar con mayor precisión las oportunidades y limitaciones que presenta la investigación en la Web en el momento actual. Fundamentalmente, los co-enlaces difieren de las co-citaciones en dos grandes rasgos: 1) por su naturaleza, y 2) por las razones que motivan la creación de los enlaces. Este último factor se aborda en el Apartado 3.4.5.2, que trata las limitadas posibilidades de desarrollo de una teoría que explique de modo global el hecho de la creación de enlaces. En relación con las diferencias relativas a la naturaleza del fenómeno, Prime, Bassecoulard y Zitt (2002) valoran la idoneidad de la analogía entre co-citación y co-sitación (término equivalente al de co-enlace), señalando las siguientes cuestiones:

- Las duplicaciones en sus más diversas formas constituyen un importante problema. Por ejemplo, la duplicación de páginas y listas de enlaces introduce niveles anormales de redundancia a la hora de cuantificar enlaces y co-enlaces en la Web. La propia estructura interna de determinadas páginas web también contribuye a ello, generando enlaces con motivos de navegación, pero sin aportar ningún valor añadido a la hora de evaluar la relevancia de los contenidos que enlazan.

- La falta de consenso a la hora de determinar la unidad idónea de estudio en la Web contrasta con la definición bien clara de conceptos en el ámbito de las publicaciones científicas. Por ejemplo, si en Bibliometría es posible realizar un estudio de co-citas en base a los autores de los artículos, en la Web no sería posible por carecer, en gran número de casos, de la información sobre la autoría del documento.

- Las direcciones web (URL) y el propio contenido de las páginas web no responden a modelos homogéneos y relativamente compartidos, a diferencia de los artículos académicos, que comparten una misma estructura. Ello dificulta la detección de patrones de similitud entre distintas páginas web.

- La Web tiene un carácter acrónico. Todos los documentos existentes en ella se encuentran siempre en un momento presente. No es posible volver a la Web de hace un año o siquiera de hace un día. No hay información precisa y públicamente disponible acerca de la fecha de creación, modificación o clausura de páginas. Se pierde así la dimensión temporal que tan importante es para la realización de investigaciones.

- La existencia de enlaces recíprocos entre páginas así como la posibilidad de que un documento enlace a otro posterior en el tiempo es una característica propia de la Web que no es concebible en la producción científica tradicional impresa. En la misma línea, un texto académico que se publica queda fijado, por lo que la modificación de un artículo constituye realmente la producción de otro distinto. En la Web un único documento puede modificarse continuamente sin dejar constancia de la versión primigenia.

Al margen de lo apuntado, Zuccala (2006a) subraya los problemas técnicos que presentan los buscadores comerciales a la hora de obtener información de co-enlaces, lo cual complica aún más la investigación. Todo este conjunto de factores hace que la interpretación del significado de enlaces y co-enlaces sea especialmente compleja, hasta el punto de que Prime, Bassecoulard y Zitt (2002: 306) señalan críticamente que "the transposition of co-citation or related techniques

(coupling) to the mapping of web topics seems promising but only with "down-sized" ambitions".

3.4.5.3. Interpretación del análisis de co-enlaces

Una vez expuestas analogías y diferencias entre co-citaciones y co-enlaces, es preciso afrontar las posibles formas de interpretación de estos últimos. Tras su tratamiento estadístico se visualizan a través de algún tipo de gráfico o diagrama que representa las relaciones entre las páginas o sitios web. Ha sido habitual en este tipo de investigación el empleo del escalamiento multidimensional para generar un mapa que muestre las posiciones relativas de los elementos en el estudio. De cara a la interpretación de los resultados nos centramos en el análisis de co-citaciones como fuente para obtener algunas ideas aplicables al análisis de co-enlaces. Zuccala (2006a: 1487) enumera algunas claves, adaptadas de White (1990):

- Los mapas de co-citas revelan la estructura cognitiva o intelectual de un campo o área científica, mostrando el consenso de los que citan sobre la relevancia de otros autores y sus contribuciones.

- Los mapas permiten mostrar cuáles son los elementos centrales y cuáles los periféricos en un área de estudio determinada, así como las dimensiones generales a partir de las cuales se agrupan los autores o trabajos.

- En la interpretación de los mapas es especialmente importante el conocimiento que atesore la persona que los interpreta en relación con el área de estudio en cuestión. Con ello se pretende que sea capaz de revelar o descubrir dimensiones adicionales, partiendo de la posición que los elementos en estudio (artículos, autores, revistas, etc.) ocupan en el mapa.

Dentro todavía del ámbito puramente bibliométrico, Noyons (2001) reflexiona sobre

la utilidad de los mapas como herramienta para la gestión científica, señalando que estos instrumentos, para verificar su efectividad, deberían ser objeto de una triple validación: por parte del investigador, por los propios científicos evaluados y por los usuarios de la información. Procesos de validación análogos podrían ser aplicables al análisis de co-enlaces en el campo empresarial. Prime, Bassecoulard y Zitt (2002) alertan contra la aplicación directa del análisis de co-citaciones a la Web, indicando que se debe proceder previamente a un filtrado de los datos. Sin embargo, este filtrado es en la mayoría de los casos muy difícil si no imposible de llevar a cabo.

Estos patrones de interpretación se pueden aplicar al análisis de co-enlaces. Sin embargo, las diferencias anteriormente apuntadas entre ambos conceptos y especialmente la heterogeneidad de razones por las que los enlaces o co-enlaces son creados genera importantes divergencias y dificultades. Zuccala (2006a) señala que la interpretación de mapas de co-enlaces aún se encuentra en su fase inicial, en pleno desarrollo. A la hora de determinar los principales factores que condicionan los *clusters* formados a partir de co-enlaces parece que el criterio geográfico es de los más relevantes, así como también la historia y el grado de adopción de las tecnologías de Internet (Zuccala, 2006a) y los factores lingüísticos y culturales (Vaughan, 2006). Vaughan (2006) explora la posibilidad de emplear los co-enlaces para la visualización de diferencias lingüísticas y culturales en Canadá. En concreto, se pretende evaluar si la información que proporcionan podría emplearse para formar *clusters* de páginas web que reproduzcan los fenómenos que se dan en el mundo *offline*; así como, si diferentes visiones (en concreto, basadas en perspectivas culturales y lingüísticas) se ven reflejadas mediante este procedimiento.

Una diferencia fundamental entre co-citaciones y co-enlaces está en los tipos de razones que motiva la creación de cada uno de ellos (Zuccala, 2006a). Frente a una teoría o pluralidad de teorías sobre la creación de citaciones, no existe todavía una teoría que explique la creación de hiperenlaces. Frente a un conjunto de razones, relativamente acotado, que pueden motivar la citación de un autor o trabajo, las posibles razones para enlazar una página web parecen ser tan amplias que

resultan casi imposibles de sistematizar de forma satisfactoria.

3.4.5.4. Métodos estadísticos y técnicas de visualización: el escalamiento multidimensional.

En general, el volumen de información obtenida en la investigación webmétrica es muy elevado, por lo que en su análisis se suelen emplear técnicas de visualización de datos que permiten interpretar la información de una manera más sencilla. La visualización de la información se emplea frecuentemente para facilitar las funciones de gestión, análisis, distribución e interpretación de grandes cantidades de datos (Baeza-Yates y Ribeiro-Neto, 1999; Card, Mackinlay y Shneiderman, 1999). Son varios los procedimientos estadísticos y de visualización que se pueden emplear (Faba Pérez, Guerrero Bote y Moya Anegón, 2004a), por ejemplo, análisis factorial, *clusters* jerárquicos, redes Pathfinder, redes neuronales, entre otros.

Como apunta Holmberg (2009), la visualización de la información constituye un proceso donde los datos se representan de forma visual para permitir la generación de conocimiento, aprovechando las capacidades humanas para detectar e interpretar patrones, tendencias y otras características que, de otro modo, serían difícilmente identificables (Gershon, Eick y Card, 1998; Börner, Chen y Boyack, 2003). Estas técnicas se han aplicado con éxito, por ejemplo, en la generación de mapas sobre las vinculaciones entre disciplinas científicas (Small y Garfield, 1985; Leydesdorff, 1987; Small, 1999; Cahlik, 2000; Boyack, Klavans y Börner, 2005). En Webmetría, su empleo ha sido diverso, por ejemplo, para el análisis de las interconexiones entre distintas áreas de la Web (Thelwall, 2001d; Thelwall y Smith, 2002; Thelwall, Tang y Price, 2003). Otro caso significativo es el de Ortega y Aguillo (2009), que han empleado los datos del proyecto *Ranking of the World Universities* (www.webometrics.info) para generar un mapa que permite visualizar cómo los distintos países se interrelacionan en el campo científico a partir de los hiperenlaces establecidos entre sus universidades.

A continuación nos centramos con más detalle en el escalamiento multidimensional, que es el procedimiento de análisis más utilizado hasta el momento en los trabajos basados en co-enlaces (por ejemplo: Larson, 1996; Faba-Pérez, Guerrero-Bote y De Moya-Anegón, 2003; 2004b; Musgrove et al., 2003; Kim, Park y Thelwall, 2006), incluidos los centrados en empresas (Vaughan y You, 2006; 2008). Se trata también del método que se emplea en los trabajos empíricos de esta tesis en los Apartados 4.2 y 4.3.

El escalamiento multidimensional (Kruskal y Wish, 1978) es un procedimiento que se utiliza para identificar las dimensiones que mejor explican las similitudes entre un conjunto de elementos, generando para ello un mapa de objetos que establece las posiciones relativas entre los elementos en estudio en función de una variable, o varias, que se emplea para medir la distancia entre ellos. Se basa en la idea de que los datos contienen información acerca de las similitudes o diferencias entre los distintos nodos considerados. La distancia entre cada nodo en el mapa generado representa el nivel de similitud entre los mismos en relación con el conjunto de los demás nodos incluidos en el estudio (Hanneman y Riddle, 2005). Esto quiere decir que si los nodos cambian o si algunos de ellos se eliminan las posiciones relativas que ocupan los elementos en el mapa se verían afectadas. El escalamiento multidimensional es aplicable para conjuntos de elementos no muy grandes. Holmberg (2009) indica que funciona únicamente para unos cientos de elementos. Sin embargo, de acuerdo con los trabajos incluidos en el Capítulo 4, podemos señalar que a partir de 50 nodos se hace más complicada la identificación de patrones y tendencias. Existe otro tipo de procedimientos, por ejemplo los basados en el *software* Pajek (Batagelj y Mrvar, 1998) o en el algoritmo Kamada-Kawai (Klavans y Boyack, 2006).

El escalamiento multidimensional se ha empleado en los estudios de empresa, por ejemplo, en marketing para el estudio de las preferencias de los consumidores. En el campo de la contabilidad y las finanzas han sido diversas sus aplicaciones: auditoría (Libby, 1979; Bailey, Bylinsky y Shields, 1983), fracaso empresarial (Mar Molinero y Serrano Cinca, 2001; Neophytou y Mar Molinero, 2004; Mar Molinero, Bishop y Turner, 2005), situación financiera de la economía europea (Serrano

Cinca, Mar Molinero y Gallizo, 2002), diferencias lingüísticas en la comunicación contable (Belkaoui, 1980), etc.

Para llevar a cabo el escalamiento multidimensional, se parte de una matriz de proximidades y se calculan unas coordenadas de manera que los distintos objetos se sitúen en un espacio de dos o más dimensiones. Las distancias mostradas en el mapa deben representar de la manera más precisa posible las distancias contenidas en la matriz de datos. Para verificar la bondad del ajuste se utiliza una medida estadística denominada *stress*, que indica en qué medida las distancias en el mapa se corresponden con las medidas de similitud obtenidas, con los co-enlaces en este caso.

De cara a la interpretación de los datos, cuando se cuentan con diferentes variables es posible utilizar dos técnicas de análisis multivariante: el *property fitting* (Pro-Fit) y análisis *cluster* jerárquico. En todo caso, siempre el primer paso en la interpretación es de carácter visual, observando y explicando las posiciones de los distintos elementos en el mapa. En los trabajos basados en co-enlaces, dadas las características de los datos, este es el tipo de análisis que se lleva a cabo. Los elementos representados en el centro del mapa serán aquellos que contengan un mayor número de enlaces comunes con el resto de elementos. Los que se encuentran más alejados del centro, en la periferia del mapa, serán aquellos con menor número de conexiones compartidas. En todo caso, esto depende de si se ha efectuado algún tipo de transformación en los datos o no. A la hora de procesar la matriz de proximidades es conveniente, dependiendo de los objetivos de la investigación, normalizar los datos que contienen con el fin de mitigar valores extremos o de ajustar los datos a través de una variable adicional. Así, por ejemplo, resulta obvio que no es igual que dos sitios web de empresas compartan cien enlaces recibidos cuando el número de enlaces recibidos total es muy pequeño, o cuando éste, por el contrario, es muy alto. En el primer caso, la similitud entre los dos sitios web es relativamente más significativa que en el segundo. Gmür (2003) efectuó una revisión de los trabajos sobre co-citaciones y plantea la conveniencia de normalizar los datos de distintas formas. Entre los métodos más frecuentes para normalizar los datos destaca el empleo del índice de Jaccard o del coeficiente de

correlación de Pearson.

Para concluir resumimos las ventajas del empleo del escalamiento multidimensional, de acuerdo con Neophytou y Mar Molinero (2004):

- No se hace ningún tipo de suposición sobre la distribución de los datos.
- No requiere la eliminación de observaciones extremas.
- No exige un ejercicio preliminar de reducción de los datos (*data reduction*).
- Es una técnica con unas sólidas bases teóricas.
- Permite obtener representaciones gráficas que permiten observar de forma sencilla las principales características de los datos.

3.4.6. Motivaciones para enlazar y análisis de contenidos

La Webmetría es una disciplina de carácter fundamentalmente cuantitativo. Sin embargo, se ha reconocido de manera generalizada la conveniencia de complementarla con estudios cualitativos que permitan conocer la naturaleza de las páginas y sitios web que establecen enlaces así como las motivaciones que explican su creación. Esto es importante para determinar si contienen información relevante con el objetivo de explicar los fenómenos a los que se vinculan. Thewall (2004) indica que el hecho de encontrar una proporción determinada de enlaces que no aportan información significativa no constituye necesariamente un problema. En cada trabajo el investigador debe determinar el porcentaje que puede considerar como aceptable.

Thelwall, Vaughan y Björneborn (2005: 122) señalan que "the quantitative techniques require complementary qualitative methods (user studies, link creation motivation studies) to understand the results completely". En esta línea, consideran que en el futuro se desarrollará una tendencia de colaboración con otras disciplinas,

como estudios culturales, análisis de redes sociales o comunicación mediada por ordenador, entre otras. A éstas se podrían añadir las aportaciones derivadas de la antropología, la etnografía y la sociología en general, que pueden complementar los estudios con una visión cualitativa más enriquecedora. Bar-Ilan (2005) propone un marco de referencia para llevar a cabo estudios cualitativos de enlaces. Análisis de contenidos de este tipo se han llevado a cabo principalmente para el caso de sitios web académicos y universitarios (entre otros ejemplos: Thelwall, 2002c; 2003b; Wilkinson et al., 2003; Vaughan, Kipp y Gao, 2007b).

Uno de los retos principales en el estudio de la Web es el desarrollo de una teoría que explique las motivaciones para crear enlaces, de forma análoga a los intentos que, desde la Bibliometría, se han llevado a cabo para elaborar una teoría que explique las motivaciones para citar en el ámbito académico. Thelwall (2003b) indica que aparte de los enlaces en revistas electrónicas y las copias en la Web de artículos académicos, muy pocos enlaces entre sitios se crean de forma similar a las citas académicas. Kim (2000), por su parte, piensa que las razones que subyacen al fenómeno de enlazar son, con frecuencia, el resultado de una compleja combinación de distintos motivos. Existen diferencias significativas entre el hecho de enlazar y el de citar, entre otras: en la Web los autores de las páginas no siempre pueden identificarse, las motivaciones son mucho más variadas, y no se puede determinar con demasiada claridad si las referencias tienen un carácter positivo o negativo. En todo caso, las similitudes entre citaciones e hiperenlaces han constituido un punto de partida para este tipo de trabajo en el ámbito webmétrico (Kim, 2000; Thelwall, Harries y Wilkinson, 2003; Harries et al., 2004). En el campo de empresa también se han realizado algunos análisis de contenidos de las páginas web que enlazan a sitios comerciales (Thelwall, 2001e; Vaughan, Gao y Kipp, 2006)

Thelwall (2003b) estudia las motivaciones para enlazar a páginas académicas, identificando fundamentalmente cuatro tipos:

- enlaces de información general: incluyen páginas que enlazan a una amplia variedad de información de carácter general, sin un tema específico;

- enlaces de propiedad: reconocen la autoría, co-autoría, pertenencia o co-pertenencia de la página o de un proyecto asociado;
- enlaces sociales: incluyen los creados con el propósito principal de reforzar los lazos sociales; y,
- enlaces sin motivo aparente: sin ninguna motivación clara en su creación.

Zuccala (2006a) plantea que la investigación basada en co-enlaces tiene, dentro de este marco, una función de reconocimiento y fortalecimiento de los lazos sociales. En su trabajo sobre institutos dedicados a la investigación matemática a nivel internacional, apunta, al menos, seis motivaciones para co-enlazar: social, de navegación, personal, geográfica, histórica y por cuestiones de prestigio.

Davenport y Cronin (2000) consideran el fenómeno de las citas académicas desde el punto de vista de la "confianza". Representan una forma de mostrar confianza en una determinada fuente, en un entorno considerado, de alguna manera, como virtual. De este modo las redes de citas pueden verse como redes de confianza y reconocimiento. Los autores proponen la traslación de esta teoría a la Web, lo cual resulta de especial importancia para nuestra investigación. Desde su punto de vista, dadas las analogías existentes entre la cita y el enlace, en aquellos casos donde enlazar no es un hecho trivial sino significativo, ambos elementos se basan en una idea común, la "confianza". Para Park, Barnett y Nam (2002), la idea clave que explica el hecho de enlazar es la búsqueda de "credibilidad" estableciendo, de algún modo, un contacto o afiliación con páginas consideradas creíbles o con una buena reputación.

En relación con estas teorías, es fundamental profundizar en las motivaciones para co-enlazar, ya que van a permitir determinar si el establecimiento de co-enlaces es un acto deliberado de carácter significativo o no. Es decir, si dichos co-enlaces proporcionan realmente información sobre similitud entre elementos, como sugiere la teoría. Por ello, los estudios más puramente cuantitativos suelen completarse con

estudios posteriores, consistentes en un análisis de contenidos, que ponen de manifiesto si los co-enlaces constituyen un fenómeno completamente aleatorio o, por otra parte, si son manifestación de un mecanismo social del cual son capaces de proporcionar información (Vaughan, Kipp y Gao, 2007a; 2007b).

3.5. Principales áreas de investigación

Las investigaciones webmétricas han prestado especial atención a los sitios web académicos, correspondientes a universidades, facultades, departamentos, investigadores, revistas científicas, etc. En este sentido, la Webmetría podría entenderse como la continuación de la investigación bibliométrica por otros medios. Sin embargo, la variedad de fenómenos que se desarrollan en la Web es tan amplia que los trabajos se han extendido rápidamente a otras áreas, entre las que destacan las empresas, el gobierno electrónico o la política. Thelwall, Vaughan y Björneborn (2005) también plantean otras áreas de estudio, como por ejemplo, el análisis de redes sociales o el estudio de aspectos geográficos en la Web. Junto a estas cuestiones, se ha tratado también la influencia de factores lingüísticos.

La ampliación de los campos de estudio en Webmetría pone de relevancia su fuerte componente interdisciplinar. Nuestra investigación se centra en el estudio de los sitios web de empresas, empleando una perspectiva fundamentalmente cuantitativa. En las páginas siguientes se van a abordar algunos de los principales trabajos y líneas de investigación webmétricas para contextualizar los avances realizados. Finalmente se prestará una especial atención a los trabajos realizados en el campo de empresas.

3.5.1. La investigación webmétrica en el ámbito académico

Los orígenes de la Webmetría, en estrecha vinculación con la Bibliometría, explican que la mayoría de la investigación llevada a cabo se haya centrado en el ámbito

académico. Se ha comparado la presencia en la Web de países, universidades, departamentos o incluso de investigadores individuales con sus indicadores de productividad científica, citaciones o redes de colaboración. Entre las variables empleadas con mayor profusión destacan el número de enlaces recibidos y los co-enlaces.

Probablemente el primer trabajo que se pueda considerar como webmétrico es el de Larson (1996), el cual lleva a cabo un análisis de co-enlaces con el fin de estudiar la estructura intelectual del Ciberespacio en relación a sitios web correspondientes a diversas temáticas científicas. El trabajo emplea el escalamiento multidimensional como procedimiento para la visualización de los datos. Otro trabajo, en los inicios de la disciplina, que emplea co-enlaces para la investigación es Dean y Henzinger (1999).

Buena parte de la investigación webmétrica en esta área ha tenido por objetivo explorar las vinculaciones entre las actividades académicas y su presencia en la Web, medida por lo general a través del número de enlaces recibidos por las páginas de la unidad en estudio. La obtención de evidencia positiva en este sentido permite emplear las variables de la Web para el análisis de la productividad científica, la calidad o el impacto de los artículos. Por ejemplo, se han detectado correlaciones significativas entre variables de productividad científica y variables que reflejan la visibilidad en la Web de las universidades; por ejemplo, en Estados Unidos (Tang y Thelwall, 2003; 2004), Taiwan (Thelwall y Tang, 2003), China (Tang y Thelwall, 2002; Thelwall y Tang, 2003) y el Reino Unido (Stuart y Thelwall, 2005), entre otros. El hecho de que las universidades más productivas reciban un mayor número de enlaces se explica porque son instituciones que generan también un mayor contenido en la Web, el cual atrae también un mayor número de enlaces. Otros trabajos, por ejemplo, Li (2005), han estudiado las mismas relaciones a nivel de departamentos, observando que aquellos con un mejor desempeño producen más contenido y crean y reciben más enlaces. Ello ha hecho que en algunos casos se haya tenido en cuenta la variable tamaño para normalizar los sitios web, por ejemplo, dividiendo el número total de páginas que componen el sitio web de un departamento por el número de investigadores de dicho departamento. Ello

presupone que, en el ámbito universitario, cada investigador tiene la capacidad de generar contenido web, algo que no ocurre habitualmente entre los trabajadores en un contexto empresarial.

Salvando las diferencias entre enlaces y citaciones, Vaughan y Shaw (2003; 2005) comparan las citaciones bibliográficas tradicionales con las que se producen en la web a través de los hipervínculos. En ambos trabajos, el número de enlaces era mayor que el de citaciones tradicionales, observando la existencia de correlaciones significativas positivas entre ambas variables. Ello evidencia que las actividades académicas tienen reflejo en la Web y que pueden ser estudiadas a partir de estas variables. Kretschmer, Kretschmer y Kretschmer (2007) adoptan una perspectiva de análisis de redes para estudiar la colaboración entre instituciones académicas, identificando que aquellas que son altamente productivas ocupan las posiciones centrales en las redes. Bordons y Gómez (2000) y Stuart, Thelwall y Harries (2007) también han explorado redes de colaboración entre investigadores a través de los enlaces que los vinculan. Por otra parte, el análisis basado en co-enlaces se ha empleado desde los inicios de la investigación webmétrica con la idea de elaborar mapas de sitios web similares. Por ejemplo, Thelwall y Wilkinson (2004) emplean enlaces directos entre sitios y co-enlaces entrantes y salientes para identificar similitudes entre páginas web.

Como se apuntó en el Apartado 3.4.6, los análisis de contenidos son muy importantes para determinar la validez de los datos que empleamos. Dado que ya se han mencionado otros trabajos de este tipo, sólo traemos aquí a colación el de Vaughan, Kipp y Gao (2007b) que clasificaron co-enlaces a sitios web de universidades de Canadá de acuerdo con el contenido de la página y el contexto en el que los enlaces se habían creado. Más del 90% de las páginas analizadas contenían temas académicos, con lo que se validaba la relevancia de la variable "enlaces" como medida válida de la similitud entre sitios web académicos.

Al tiempo que se han investigado actividades académicas, han surgido cuestiones adicionales de carácter más transversal que pueden afectar también a otros tipos de páginas web, no únicamente de tema académico. Se trata, por un lado, de la

existencia o no de patrones geográficos en el modo en que se distribuyen los enlaces en la Web y, por otro, de la influencia del factor lingüístico. Un trabajo que aborda, entre otras, ambas variables es el llevado a cabo por Vaughan y Thelwall (2005), que estudia los enlaces existentes entre universidades canadienses, identificando la lengua como una variable significativa a la hora de explicar los enlaces recibidos por los sitios web. En relación con la variable geográfica, observan una clara división entre las universidades de la costa este y las de la costa oeste, lo cual viene a reflejar cierta división política, social y económica del país. En todo caso, el patrón geográfico únicamente se limita a esta gran división.

Atendiendo a la geografía, se ha argumentado con frecuencia que Internet y la Web facilitan la superación de las barreras físicas y de las fronteras políticas entre países y regiones. Dos páginas web cualesquiera se encuentran potencialmente a un único enlace de distancia. Si la geografía física en la Web se difumina deberíamos esperar que otras variables adquirieran mayor protagonismo a la hora de explicar las redes que se forman. Sin embargo, distintos trabajos realizados ponen de manifiesto que esto no es así. Zook (2005) afirma que la estructura de la Web se encuentra directamente influenciada por cuestiones geográficas. Halavais (2000) señala también que es más probable que los enlaces se creen entre sitios web de un mismo país que entre sitios de países distintos. De algún modo, esto indica que los enlaces en la red reflejan las relaciones existentes en el mundo físico, ámbito en el cual la proximidad sigue siendo una variable significativa. Consideramos que si la Web respondiera únicamente a dinámicas propias sería más complicado de identificar estos patrones. Al mismo tiempo, ello implicaría una desconexión entre el mundo físico y el virtual que nos impediría analizar la Web para obtener información sobre el mundo *offline*. En cualquier caso, parece lógico añadir que la variable geográfica no puede aislarse y considerarse por sí sola ajena a otras variables influyentes, como son los factores lingüísticos, políticos y económicos, que aparecen generalmente de manera superpuesta. Por lo general, para visualizar la existencia de patrones geográficos se tienen en cuenta los co-enlaces a los sitios en estudio o, principalmente, los enlaces directos existentes entre dichos sitios.

En el contexto europeo, se ha observado la existencia de un factor geográfico

significativo en los enlaces directos existentes entre universidades en el Reino Unido (Thelwall, 2002b) y también a mayor escala en la región Asia-Pacífico (Thelwall y Smith, 2002). En Estados Unidos (Tang y Thelwall, 2003) no se ha obtenido evidencia significativa en este sentido. Heimeriks y Van den Besselaar (2006) han mostrado que las universidades europeas tienden a establecer mayor número de enlaces con instituciones de su propio país y en segunda instancia con universidades de países próximos. El contexto europeo es especialmente complejo en este sentido debido a la existencia, por un lado, de determinadas políticas comunes a través de la Unión Europea, y, por otro, de estados nacionales con distintos marcos institucionales, distintas lenguas y distintas culturas académicas. Ortega et al. (2008) también han investigado el contexto europeo, mientras que Ortega y Aguillo (2009) han prestado atención a universidades de todo el mundo, encontrando también patrones geográficos en la redes de enlaces directos entre sitios web. Thelwall y Zuccala (2008) en el contexto europeo identifican *clusters* regionales entre instituciones de países que son nuevos miembros de la UE. Esto indica que aún no existe una integración plena de estos países, al menos en este ámbito, ya que son las universidades de países más ricos las que presentan posiciones de dominio.

En relación con la lengua, el inglés ha sido tradicionalmente el principal vehículo de comunicación en la Web (Lavoie y O'Neill, 2001), especialmente en un contexto internacional, por ejemplo, en ámbitos empresariales o académicos. En la actualidad sigue siendo la *lingua franca* para las relaciones internacionales en la Web. El poderío del inglés queda claramente de manifiesto en el caso de aquellos países con varias lenguas, siendo ésta una de ellas. Es el caso de Canadá, donde las universidades en idioma inglés reciben un mayor número de enlaces que aquellas que emplean el francés (Vaughan y Thelwall, 2005). El empleo del inglés ha sido también estudiado en el contexto académico en China y Taiwan (Thelwall y Tang, 2003). En el polo opuesto, nos encontramos con lenguas minoritarias cuyo empleo supone un aislamiento debido, por ejemplo, a la dificultad para ser enlazadas por usuarios de lenguas distintas. Países cuyas lenguas se encuentran

en ese caso son, por ejemplo, Irán (Aminpour et al., 2009) o Grecia (Thelwall, Tang y Price, 2003). En el contexto europeo, Thelwall y Zuccala (2008) y Thelwall, Tang y Price (2003) han estudiado el impacto de la lengua en la Web.

Por último, cabe señalar que la lengua tiene un impacto significativo en el comportamiento de los usuarios y en la forma en que usan la Web (Kralisch y Mandl, 2006). Kralisch y Köppen (2005) han apuntado también la influencia que ejerce en el comportamiento de búsqueda de información de los usuarios y en el grado de satisfacción con los resultados que obtienen.

3.5.2. La investigación webmétrica en el ámbito del sector público

La investigación en el ámbito del sector público se refiere por lo general al gobierno electrónico o e-gobierno (*e-Government*). El e-gobierno aborda el modo en el que los gobiernos, en sus distintos niveles, emplean las tecnologías de la información, principalmente Internet y la Web, para su gestión interna y para su interacción con los ciudadanos. No debería consistir únicamente en el traspaso de una serie de servicios y procedimientos a la red, sino que este proceso debería conllevar también la transformación de los mismos, de modo que se puedan aprovechar al máximo las potencialidades de este medio (Burn y Robins, 2003). Otros autores, como Roy (2003) enfatizan la dificultad para alcanzar una definición precisa y adecuada de lo que es el e-gobierno. Pérez, Bolívar y Hernández (2008) apuntan que el gobierno electrónico se refiere principalmente a cinco aspectos: la interacción entre entidades públicas, el desarrollo de servicios basados en la Web, el comercio electrónico, la democracia digital y las finanzas electrónicas (Moon, 2002; Wang y Rubin, 2004).

Buena parte de los trabajos llevados a cabo en relación con el e-gobierno han estado basados en la teoría de la agencia; por ejemplo, Thompson (1998), Fisher, Laswad y Oyelere (2005) y Alvarez y Hall (2006). Desde el punto de vista de las tecnologías de la información, el e-gobierno puede abordarse desde distintas

perspectivas. Por ejemplo, Gascó (2003) analiza las transformaciones que está experimentando la administración pública como resultado del desarrollo de la sociedad del conocimiento. Davison, Wagner y Ma (2005) estudian el modo en que se produce la transición hasta el e-gobierno y el modo de hacerlo de manera efectiva; mientras que Ebrahim e Irani (2005) se centran en las dificultades de este proceso. Shackleton, Fisher y Dawson (2006) se centran en el progreso del e-gobierno en los gobiernos locales de Australia. Otros trabajos analizan como los sitios web se adecuan a la necesidades de personas con discapacidades (Shi, 2006) o de personas mayores (Becker, 2004).

Para un buen desarrollo del e-gobierno, es fundamental que las instituciones públicas sean sometidas al escrutinio de los ciudadanos. En este sentido, diversos trabajos se han centrado en la difusión de información financiera a través de la Web (por ejemplo: Caba Pérez, López Hernández y Rodríguez Bolivar, 2005; Rodríguez Bolivar, Caba Pérez y López Hernández, 2006; Caba Pérez, Rodríguez Bolivar y López Hernández, 2008), así como de información general de ayuntamientos (Gandía y Archidona, 2008). Algunos trabajos se han centrado en la evaluación de distintas iniciativas públicas de gobierno electrónico en el contexto de la Unión Europea (Torres, Pina y Royo, 2005; Torres, Pina y Acerete, 2006).

En relación con la actividad de carácter más político, Chadwick y Howard (2009) destacan el potente desarrollo de la investigación en política e Internet, lo que se ha dado en llamar *Internet Politics*. Han surgido conceptos como e-democracia (*e-Democracy*) o e-campaña (*e-Campaign*) y se han desarrollado trabajos en distintos campos (por ejemplo: Gibson y Ward, 2000; La Porte, Demchak y de Jong, 2002; King, 2006; Strandberg, 2006). Un buen número de estudios sobre política en Internet han gravitado en torno a dos teorías enfrentadas (Anstead y Chadwick, 2009): por un lado los que consideran que Internet permite reducir las diferencias entre distintas fuerzas políticas, permitiendo una mayor igualdad de oportunidades; y, por otro, los que consideran que en Internet siguen existiendo barreras que provocan que, sólo los partidos políticos y las instituciones con acceso a recursos, puedan aprovechar su potencial de una manera más eficiente (Pickerill, 2001; Ward y Gibson, 2003). Para Burt y Taylor (2001), las condiciones sociales y los distintos

sistemas de valores condicionan el modo en que se emplean las tecnologías. March (2006) considera que aquellas democracias en las que los sistemas políticos y de comunicación aún no están muy asentados podrían estar más abiertas a un mayor desarrollo de las nuevas tecnologías.

En relación con el activismo en Internet, investigaciones en Estados Unidos (Gibson, Kleinberg y Raghavan, 2003; 2005) y en Europa (Norris, 2003) ponen de relieve que muchos de los visitantes a los sitios web de partidos políticos son ciudadanos que ya están involucrados activamente en asuntos políticos, lo cual cuestiona de algún modo la capacidad de las nuevas tecnologías para captar y motivar a nuevos ciudadanos. En este sentido, el papel de la Web 2.0 es muy relevante en la práctica y, de manera progresiva, recibe una mayor atención en la investigación académica (Chadwick y Howard, 2009).

Tomando una perspectiva más propiamente webmétrica, destaca la tesis doctoral de Holmberg (2009), que emplea diversos métodos para analizar los gobiernos locales en Finlandia. Por otra parte, Yan y Zhu (2008) aplican técnicas webmétricas de análisis de impacto de enlaces al estudio de las páginas web de los gobiernos regionales chinos. Para ello, analizan las relaciones entre el número de enlaces recibidos por los sitios web, por un lado, y la capacidad económica (producto interior bruto de la provincia en cuestión) y la población, por otro, observando correlaciones positivas significativas. Los autores argumentan que las provincias con mayor población o actividad económica dispondrán a su vez de mayor número de personas o de empresas con presencia en la Web y, por tanto, existirán mayores posibilidades de que las páginas web de los gobiernos regionales sean enlazadas. Según plantean, el número de enlaces recibidos puede ser considerado un indicador útil para evaluar la situación económica de las regiones.

3.5.3. La triple hélice

El concepto de *triple hélice* (en inglés, *triple helix*) (Etzkowitz y Leydesdorff, 1995;

1997; Leydesdorff & Curran, 2000; Leydesdorff & Meyer, 2007) hace referencia a la transferencia de conocimiento entre tres ámbitos institucionales fundamentales que conforman cada una de las hélices del modelo: sector académico (universidad), sector privado (empresa) y sector público. Dicha transferencia de conocimiento se ha convertido en la última década en una política estratégica dentro de ámbitos tan importantes como la Unión Europea. En los últimos años se han empleado técnicas webmétricas con el fin de poner de relieve estas relaciones a través de la información contenida en la Web. La pregunta básica que se han planteado estas investigaciones es: ¿puede la Web proporcionar información relevante sobre la colaboración entre los sectores de la triple hélice? Hasta ahora la mayor parte de la investigación en la Web se ha desarrollado centrada en un único sector, al menos en el campo empresarial, debido a la dificultad para analizar la Web en su conjunto y las dificultades para obtener o sacar conclusiones de grupos heterogéneos de organizaciones (Stuart y Thelwall, 2006).

Stuart y Thelwall (2006) investigan las relaciones de triple hélice en relación con el sector automovilístico en las *Midlands* (Reino Unido). Para ello, examinan si los enlaces se pueden emplear como indicadores para determinar diversos niveles de colaboración entre organizaciones dentro de los tres sectores. La investigación confirma la existencia de importantes diferencias en el comportamiento de los sitios web a la hora de enlazar a otros sitios pertenecientes a alguna de las tres partes de la triple hélice: universidad, empresas y administración pública.

La escasez de enlaces no es exclusiva de las páginas web comerciales (Shaw, 2001), en su intento por no desviar la atención del cliente, sino que también se da en las pertenecientes a la administración pública. Stuart y Thelwall (2006) consideran que esto se debe al tipo de información que incluyen en sus páginas web y a quiénes son los encargados de incluir dicha información. Definitivamente, parece que las políticas de comunicación de las organizaciones son esenciales para entender dicho comportamiento. Otras posibles explicaciones serían: bien, el que dicha información de terceros no sea considerada por la organización como relevante para los usuarios de la información, o bien, el que las referencias a terceros no se hagan en forma de enlace sino textualmente.

García-Santiago y de Moya-Anegón (2009) emplean un análisis de espacios web relacionados o co-enlaces salientes (*co-outlinks*) para realizar un estudio basado en el modelo sociológico de la triple hélice. A través del análisis de los enlaces simultáneos a dos sitios web establecidos por el conjunto de sitios en estudio se permite visualizar las posiciones relativas que ocupa cada sitio web. Las redes permiten estudiar la imagen que cada institución desea atesorar a partir de su posición en la red de relaciones en la que se inscribe. Quizá uno de los aspectos más interesantes de este trabajo, aparte del empleo de co-enlaces salientes, sea la aproximación al análisis de la Web desde el reconocimiento de su heterogeneidad. El objetivo es observar cómo partiendo de la heterogeneidad de los sitios web estudiados, la estructura de enlaces existente permite establecer distintos grupos y subgrupos de sitios y sus relaciones. En total se seleccionan diez grandes áreas o sectores, algunos de ellos vinculados directamente con la triple hélice y otros ajenos a ella o no vinculados de forma directa (como, por ejemplo, medios de comunicación o fundaciones). De este modo se toman más 400 sitios como punto de partida y se emplea un *software* especializado (*Linkbot*) para la obtención de los datos necesarios. Los distintos análisis llevados a cabo muestran un comportamiento particular de las páginas web comerciales (bancos, asociaciones empresariales, cámaras de comercio, etc.) que aparecen especialmente vinculadas entre sí, así como una relación estrecha con los sitios de medios de comunicación.

3.5.4. La investigación webmétrica en el ámbito empresarial

Las empresas han adquirido progresivamente una presencia destacada en Internet, sobre todo desde la creación de la Web. En 2010 vivimos una simbólica efeméride en este sentido: los 25 años del registro de primer dominio *.com* por parte de un fabricante informático hoy desaparecido, Symbolics (Delclós, 2010). Actualmente existen más de 80 millones de sitios activos registrados bajo este dominio. No todos son empresas, pero sí muchos de ellos.

Es complicado llevar a cabo una revisión exhaustiva de toda la investigación

realizada en el ámbito económico y empresarial e Internet. Por ejemplo, en los veinte años que han pasado desde la creación de la Web, el panorama ha cambiado tanto que muchos trabajos carecen de interés hoy en día, mientras que otros se centran en perspectivas ajenas a la nuestra. En esta parte nos vamos a centrar fundamentalmente en los trabajos que estudian las relaciones establecidas entre las empresas y los usuarios a través de Internet, con un especial énfasis en la divulgación de la información financiera y empresarial en general así como en la evaluación de la presencia de las empresas en este medio.

3.5.4.1. Una aproximación general a la investigación de empresas e Internet

La investigación de las relaciones entre Internet y la economía en general puede adoptar muchos puntos de vista: desde el modo en que Internet puede influir en determinadas teoría económicas sobre los mercados (Brynjolfsson, Jeffrey y Smith, 2003; 2006; Ellison y Ellison, 2005), hasta las relaciones entre el comercio electrónico y el desarrollo socio-económico (Boateng et al., 2008), pasando por la divulgación de información financiera y empresarial en su conjunto a través de Internet, área en la que nos centraremos más por haber recibido una especial atención desde el ámbito de la contabilidad y las finanzas.

Una parte importante de la investigación analiza cómo la información es transmitida a los inversores en el contexto de Internet y cómo los inversores actúan ante dicha información (Barber y Odean, 2001). En relación con la divulgación de información financiera, Internet tiene la capacidad de revolucionar la transmisión de información (Jones y Xiao, 2004), eliminando intermediarios y costes de emisión, permitiendo un flujo de datos en tiempo real, enriqueciendo los medios y los formatos en los que se emite, permitiendo una mayor interacción por parte de los grupos de interés, etc.

Actualmente, la presencia de las empresas en Internet es muy heterogénea. García-Borbolla, Larrán y López (2005) identifican tres maneras que las empresas

pueden adoptar para estar en la Web:

1. Presencia ornamental. Se trata de sitios web que básicamente proporcionan información general de la empresa, sin dirigirse a un público en concreto ni implementar una política de transparencia informativa definida.

2. Presencia informativa. Se trata de empresas que en su sitio web incluyen información corporativa pero también secciones específicas con información dirigida a grupos de interés concretos, por lo general clientes actuales o potenciales.

3. Presencia relacional. Incluye empresas que emplean su sitio web para llevar a cabo operaciones comerciales con diferentes grupos de interés. Podríamos decir que se trata de una perspectiva enfocada al comercio electrónico.

En sintonía con lo apuntado por Healy y Palepu (2001), el incremento en la cantidad y calidad de información financiera que se divulga en Internet ha ido en aumento. Diversos trabajos han estudiado las relaciones de las empresas con los inversores en Internet en diversos países: Estados Unidos (Deller, Stubenrath y Weber, 1999), Reino Unido (Craven y Marston, 1999; Deller, Stubenrath y Weber, 1999), Alemania (Deller, Stubenrath y Weber, 1999; Marston y Polei, 2004), España (Giner y Larrán, 2002), Irlanda (Brennan y Kelly, 2000), Suecia (Hedlin, 1999), Japón (Marston, 2003) y los países de la zona Euronext (Geerings, Bollen y Hassink, 2003). En el caso de las entidades dedicadas a las microfinanzas, enfocadas a clientes con limitaciones para acceder a fuentes de financiación, Gutiérrez-Nieto, Fuertes-Callén y Serrano-Cinca (2008) han estudiado los incentivos para difundir información financiera y social en Internet. Otros trabajos han ofrecido guías para la divulgación de información financiera en Internet; por ejemplo: Ashbaugh, Johnstone y Warfield (1999); Lymer (1999); Bonsón-Ponte, Escobar-Rodríguez y Flores-Muñoz (2006); entre otros.

Otra serie de trabajos han analizado diversos tipos de fenómenos de Internet

vinculados a variables económicas y financieras. Por ejemplo, Tumarkin y Whitelaw (2001) estudiaron las relaciones entre los niveles de actividad de mensajes sobre valores negociados en plataformas de Internet y los rendimientos y volúmenes de negociación anormales de determinadas acciones. Das y Sisk (2005) aplicaron el análisis de redes sociales con propósitos parecidos. Otro trabajo relacionado es el de Antweiler y Frank (2004), que analizaron los mensajes publicados en Yahoo! Finance y en Raging Bull sobre valores del Dow Jones Industrial, concluyendo que es una información que ayuda a predecir la volatilidad del mercado, si bien su efecto en términos económicos es pequeño. Jin, Matsuo e Ishizuka (2009) utilizan también análisis de redes sociales para hacer un *ranking* de empresas a partir de datos recogidos de la Web. Diversos trabajos han analizado la eficacia de los correos electrónicos no solicitados (*spam*) para influir en el precio de las acciones cotizadas en bolsa (D'Alessio, 2007; Hanke y Hauser, 2008).

Otra área de investigación vinculada a la divulgación de información financiera es la que se centra en el empleo del XBRL (eXtensible Business Reporting Language), que es un lenguaje basado en XML, específicamente diseñado para la divulgación de información financiera en Internet (por ejemplo, Bonsón, Cortijo y Escobar, 2009).

Con el objeto de evaluar la presencia de las empresas en la Web, Calero, Ruiz y Piattini (2005) proponen un modelo *Web Quality Model* (WQM) con el que medir distintas dimensiones relativas al contenido, la presentación y la navegación en el sitio web. Algunos estudios de sitios web comerciales están vinculados al desarrollo del comercio electrónico; por ejemplo, los relativos a la estructura y atractivo del sitio web (White y Manning, 1998; Nel et al., 1999).

Algunos autores han discutido a nivel teórico el valor de la Web como medio para que las empresas obtengan inteligencia (Graef, 1997; McGonagle y Vella, 1999), mientras que otros han propuesto algunas técnicas específicas (Zanasi, 1998; Burwell, 1999; Nordstrom y Pinkerton, 1999; Bauer y Scharl, 2000; Fleisher y Blenkhorn, 2001). Al margen, se han desarrollado programas informáticos que tienen por objeto la aplicación de técnicas de minería de datos (*data mining*)

(Madnick y Siegel, 2002). Shaw (2001) analiza el comportamiento de sitios con dominios *.com* a la hora de establecer enlaces frente a sitios alojados en otros tipos de dominios. Si bien el dominio *.com* surge en principio para dar cabida a páginas de empresa o con contenidos comerciales, su empleo dista mucho de mantener esta coherencia. A pesar de dicha limitación, sí es cierto que el uso del dominio *.com* con fines comerciales es más intenso que el de otros dominios, especialmente en el momento en que se lleva a cabo el estudio. El estudio concluye que las páginas *.com* incluyen mayor número de enlaces, aunque la mayoría son de carácter interno con el objeto de evitar desviar la atención del visitante hacia otras páginas, especialmente hacia los competidores. Ello se debe a que la función principal de estos sitios comerciales es la de autopromocionarse. Las funciones predominantes en los sitios con dominios distintos al .com depende del objeto del sitio, pero, en general, se pueden resumir en alcanzar una mayor comprensión o conocimiento sobre un tema así como en establecer enlaces a distintas fuentes de información consideradas relevantes. Ha y James (1998) coinciden al concluir que los sitios web comerciales no enlazan normalmente a páginas externas, si bien sí que contienen muchos enlaces internos (auto-enlaces). Los enlaces a terceros son muy infrecuentes ya que la empresa pretende controlar el mensaje que se transmite a sus visitantes y evitar que desvíe su atención hacia otras páginas (Bentley, 1997).

Otro tipo de estudio es el de Park, Barnett y Nam (2002) que aplica la teoría de grafos al análisis de redes. En concreto, analizan una selección de los sitios web más visitados en Corea del Sur. El trabajo considera que la razón fundamental para enlazar a otras páginas es la búsqueda de "credibilidad" mediante la asociación con sitios que se consideran con buena reputación o "creíbles". Aunque el conjunto de entidades en el estudio no es exclusivamente de carácter comercial, se descubre que son precisamente los sitios vinculados a actividades financieras (por ejemplo, relativos a tarjetas de crédito, bancos, seguros, etc.) los que ocupan una posición central en los distintos *clusters* identificados. El modo en que estos sitios confieren credibilidad a los sitios web que los enlazan es proporcionando seguridad acerca de las transacciones comerciales electrónicas que se ofertan.

Uno de los más recientes intentos de explotar la información disponible es el

llevado a cabo por Choi y Varian (2009), que proponen el empleo de los datos que ofrece el servicio Google Trends para anticipar el comportamiento de compra de los consumidores. Este servicio facilita información de la evolución en el uso de un término de búsqueda determinado a lo largo del tiempo. De acuerdo con la evidencia obtenida hasta el momento, dicho indicador no presenta un carácter predictivo, sino que en principio permite anticipar datos comerciales cuyo cálculo y publicación conlleva un retraso y, por tanto, una consiguiente pérdida de relevancia. En su caso, por ejemplo, ensayan un modelo estadístico para anticipar los datos sobre ventas de coches mensuales de una determinada marca, adelantándose en el tiempo a la publicación oficial de los mismos por parte de la firma. De forma similar, fuera del ámbito comercial, investigadores de Google (Ginsberg et al., 2009) han empleado los términos utilizados por los usuarios en este buscador para tratar de prever la evolución en la incidencia de la gripe, permitiendo anticipar medidas preventivas.

La visibilidad en la Web es un factor crucial para el éxito de las empresas, especialmente en determinados sectores más intensivos en información. El número de enlaces recibido es una buena medida de la visibilidad de las empresas en la red, ya que constituye, además, uno de los criterios fundamentales para la indexación y buen posicionamiento de los sitios web en los buscadores, los cuales son los principales facilitadores del acceso de los potenciales clientes a las empresas.

En el subapartado siguiente nos centramos en la investigación webmétrica en empresas basada en hiperenlaces, conectando directamente con la investigación llevada a cabo en esta tesis doctoral.

3.5.4.2. La investigación webmétrica de empresas

En este apartado vamos a agrupar la investigación en empresas centrada fundamentalmente en el análisis de hiperenlaces, prestando una especial atención a los trabajos de Vaughan y otros, en tanto que han constituido una referencia para la investigación empírica llevada a cabo en esta tesis doctoral.

Hasta ahora, el estudio webmétrico de sitios comerciales no ha alcanzado el desarrollo de la investigación en el ámbito académico, lo cual resulta paradójico si tenemos en cuenta el predominio de los sitios comerciales en la Web y el incremento continuo de la importancia del comercio electrónico (Thelwall, Vaughan y Björneborn, 2005) y de la presencia de las empresas en Internet. A pesar de ello, Vaughan y Wu (2004) señalan que se conoce mucho más acerca de las páginas y sitios web académicos que de los comerciales, a pesar de que ya en el año 1999, el 83% de los servidores web incluían contenidos comerciales (Lawrence y Giles, 1999). La presencia comercial en la Web se ha incrementado aún más desde entonces, así como el número de operaciones que es posible llevar a cabo.

A finales de los 90 el empleo comercial de la Web estaba aún en una fase experimental, como confirma Thelwall (2000a) que encontró evidencia de que las empresas no diseñaban sus sitios web de manera que fueran fáciles de indexar por los motores de búsqueda. Como resultado, el 23% de los sitios no estaban registrados en cinco de los principales buscadores consultados. Thelwall (2000c) realizó un muestreo aleatorio de nombre de dominio *co.uk* (el dominio con fines comerciales en el Reino Unido) con el fin de determinar el tipo de empresas que de forma más destacada empleaban el sitio web para usos comerciales. El estudio permitió observar que las empresas relacionadas con los medios de comunicación y las profesiones (como, por ejemplo, las firmas de abogados) presentaban una fuerte presencia en la Web, aunque la industria informática era la que contaba con una mayor visibilidad. Thelwall (2001e) estudió los hiperenlaces en páginas comerciales, encontrando que apenas el 66% incluían enlaces apuntando a otras páginas web relacionadas. La mayor parte de los enlaces se dirigían a sitios con los que existían relaciones comerciales.

A continuación nos centramos en una serie de trabajos basados en el análisis de impacto de enlaces de sitios web de empresas. Vaughan y Wu (2004) investigan si existe algún tipo de correlación entre los enlaces que reciben, los sitios web de empresas y determinadas variables financieras. Las empresas analizadas son chinas y pertenecen a dos grupos: el primero incluye las cien primeras empresas de telecomunicaciones y tecnologías de la información en general, y, el segundo, a las

cien primeras empresas de titularidad privada. La elección del sector de telecomunicaciones responde fundamentalmente a la propia naturaleza de la actividad que llevan a cabo las empresas, con un componente altamente tecnológico, lo cual asegura una presencia destacada en la Web. La elección de un segundo grupo de empresas pertenecientes a sectores de actividad diversos tiene por objeto comparar resultados para determinar si la homogeneidad en cuanto a la actividad de las empresas es un factor destacado a la hora de obtener correlaciones significativas.

Las variables financieras recabadas para las empresas tecnológicas son: ingresos brutos, resultado, ingresos por exportaciones y gastos de investigación y desarrollo; mientras que para las empresas de titularidad privada el único dato disponible es el correspondiente a los ingresos brutos. La obtención de los datos de enlaces recibidos por las páginas web de las empresas se realiza empleando diversos motores de búsqueda comerciales (Google, AltaVista, AllTheWeb y MSN Search) que admitían esta tipo de consultas en 2002, fecha de la obtención de los datos.

Dado que Vaughan y Thelwall (2003) muestran que el número de enlaces que recibe un sitio está relacionado con el tiempo que lleva publicado (edad), este trabajo emplea una variable consistente en el tiempo desde la creación del sitio web de la empresa, para lo cual se emplean datos del Internet Archive (www.archive.org). Ello se debe a que las páginas web más antiguas reciben mayor número de enlaces por un simple proceso de acumulación a lo largo del tiempo.

Los resultados al calcular el coeficiente de correlación de Spearman para las empresas del grupo 1 se muestran en la Tabla 3.3.

Tabla 3.3. Coeficiente de correlación de Spearman para el grupo de empresas de tecnologías de la información en China (Vaughan y Wu, 2004)

Coeficiente de correlación de Spearman	Ingresos brutos	Resultado	Gasto en I+D
Nº enlaces	0,51**	0,3**	0,64**
Nº enlaces / edad	0,5**	0,3**	0,63**
** Coeficientes de correlación significativos al nivel 0,01. Fuente: Vaughan y Wu (2004)			

Los datos muestran una correlación significativa entre los enlaces recibidos y los ingresos brutos de las empresas, siendo especialmente alta en el caso del gasto en investigación y desarrollo. Esto parece indicar que las empresas que invierten más en I+D gozan también de una mejor presencia en la Web y por tanto sus sitios son más visibles y más enlazados. Por el contrario, el grupo de las mayores empresas de titularidad privada no muestra correlación significativa entre los ingresos brutos y los enlaces recibidos, lo cual apunta a que las medidas de desempeño de las empresas de distintos sectores no son directamente comparables.

En sintonía con este trabajo, Vaughan (2004b) investiga si las relaciones examinadas se verifican en países distintos, para lo cual estudia las empresas del sector de tecnología de la información en China y en Estados Unidos. Los resultados se pueden observar en la Tabla 3.4.

Tabla 3.4. Coeficiente de correlación de Spearman para empresas de tecnologías de la información en China y Estados Unidos (Vaughan, 2004b)

Coeficiente de correlación de Spearman	Enlaces & Ingresos	Enlaces & Resultado	Enlaces/edad & Ingresos	Enlaces/edad & Resultado
China	0,51**	0,30**	0,50**	0,30**
Estados Unidos	0,51**	0,35**	0,58**	0,37**
** Coeficientes de correlación significativos al nivel 0,01. Fuente: Vaughan (2004b)				

Se refuerza pues la evidencia de la relación entre enlaces recibidos y las magnitudes de ingresos y resultado, especialmente la primera de ellas. Las conclusiones de este trabajo sugieren que, con independencia del grado de desarrollo tecnológico de un país o con independencia del momento en que éste se produjo, la relación es consistente.

Por último, un tercer trabajo (Vaughan, 2004a) compara las empresas de tecnologías de la información de Canadá con las de Estados Unidos. En él se analiza la población completa de empresas del sector, a diferencia de los anteriores trabajos que seleccionaban las empresas entre las mayores del sector. Se toma en consideración una nueva variable, el tamaño de las empresas, medido a partir del número de empleados, al considerar que una compañía mayor obtendrá mayores ingresos, permaneciendo el resto de variables iguales. Calculando de nuevo el coeficiente de correlación de Spearman se obtienen los resultados que se indican en la Tabla 3.5.

Tabla 3.5. Coeficiente de correlación de Spearman para empresas de tecnologías de la información en Canadá y Estados Unidos (Vaughan, 2004a)

Coeficiente de correlación de Spearman	Enlaces & Número de empleados	Enlaces & Ingresos	Enlaces & Ingresos por empleado	Enlaces/ edad & Ingresos	Enlaces/ edad & Ingresos por empleado
Canada	0,57**	0,55**	0,35**	0,55**	0,36**
Estados Unidos	0,68**	0,71**	0,53**	0,67**	0,51**
** Coeficientes de correlación significativos al nivel 0,01. Fuente: Vaughan (2004a)					

Los datos indican que, aún incluyendo la variable tamaño, sigue existiendo una correlación significativa entre ambas.

Las conclusiones obtenidas a partir de los anteriores trabajos indican que los enlaces recibidos por páginas web de empresas podrían emplearse como un indicador complementario del desempeño de las mismas. Dada la correlación positiva existente, se puede afirmar que empresas con un mejor desempeño financiero muestran también una mayor presencia en la Web, atrayendo la atención de más páginas. De acuerdo con lo observado, parece que únicamente sería comparable el desempeño a través de los *inlinks* si el grupo de empresas seleccionado es homogéneo, es decir, pertenecen a un mismo sector de actividad.

Otra de las grandes líneas de investigación es la basada en el análisis de co-enlaces para obtener representaciones gráficas de las relaciones entre distintas entidades. Se ha utilizado en la investigación de empresas para el análisis de las posiciones competitivas de las firmas pertenecientes a un sector de actividad en concreto. Estos trabajos abren la posibilidad al empleo de los datos cuantitativos sobre la estructura de enlaces de la Web, con el objeto de obtener información,

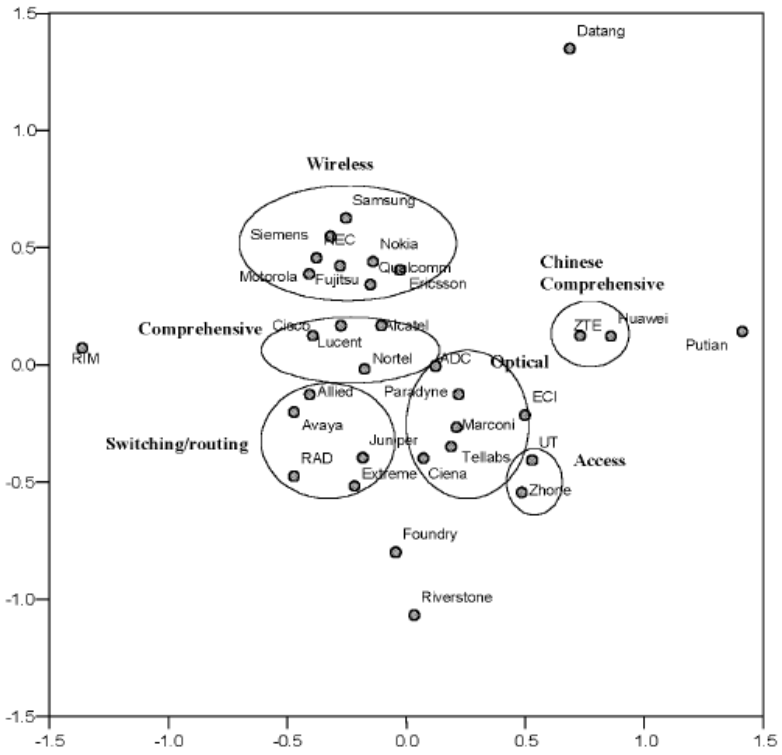
hasta el momento no explotada, sobre el entorno competitivo de la empresa con la vista puesta en la toma de decisiones estratégicas (Porter, 1980; 1985). La Inteligencia Competitiva (*Competitive Intelligence*) es un área de investigación que parte de la necesidad de conocer el entorno empresarial con el objetivo de explotar y mantener las ventajas competitivas de la organización (Heppes y du Toit, 2009). La investigación en Internet ha sido productiva en este sentido (Tan, Foo y Hui, 2002; Reid, 2003; Chau, et al., 2007). Kahaner (1996) considera que la Inteligencia Competitiva consiste en un plan sistemático para recabar y analizar información sobre los competidores y las tendencias generales de los sectores a los que pertenecen. La abundancia de información disponible en Internet y la Web abre nuevas posibilidades a las empresas y a otros agentes económicos para analizar datos que son accesibles de forma pública y generar conocimiento útil para, en última instancia, tomar decisiones. Los siguientes trabajos basados en co-enlaces pueden encuadrarse también en esta función de la empresa y de los analistas para controlar la estrategia competitiva en el sector.

En primer lugar, Vaughan y You (2006) examinan la viabilidad de emplear la información que proporcionan los co-enlaces con el fin de analizar las posiciones competitivas de un grupo de empresas. Una vez que Vaughan, Gao y Kipp. (2006) demuestran la inviabilidad de emplear a estos efectos enlaces directos entre empresas competidoras, es lógico suponer que sitios web independientes sí mostrarán mayor predisposición a enlazar conjuntamente sitios web de empresas competidoras. La investigación se centra en empresas del sector de las telecomunicaciones con presencia tanto en el mercado global como en el mercado chino. Debido a la inviabilidad de analizar un número muy elevado de empresas se opta por seleccionar un conjunto de 32 firmas, atendiendo a criterios como: nivel de ingresos, pertenencia a distintas zonas geográficas del mundo y especialización en distintas áreas dentro de las telecomunicaciones. Dado que el estudio pretende establecer las posiciones competitivas de las empresas en el mercado chino y en el global, para el primero se emplea Yahoo! China, filial de Yahoo! que opera con una base de datos propia especializada. Para comprobar las posiciones competitivas dentro del mercado chino únicamente, se limita la consulta a páginas en este idioma; esto implica que los datos pueden incluir enlaces que se originen fuera de

China con la condición de que sean en esta lengua.

Las matrices de co-enlaces obtenidas se normalizan aplicando el índice de Jaccard y se procesan mediante el escalamiento multidimensional para generar los mapas competitivos de la industria. Se prueba también a representar las matrices de co-enlaces sin normalizar, sin embargo los resultados no son tan adecuados para observar la realidad competitiva del sector. Los mapas obtenidos con este procedimiento, uno para el mercado global y otro para el chino, muestran que las empresas aparecen agrupadas entre sí formando *clusters* que reflejan la realidad competitiva del sector. La Figura 3.3 muestra el mapa correspondiente al mercado global.

Figura 3.3. Mapa correspondiente al mercado global del sector de las telecomunicaciones (Vaughan y You, 2006: 619)



Las principales diferencias entre los datos obtenidos para el mercado global y para el mercado chino es que los datos que se obtienen a través de Yahoo! China muestran un mayor número de enlaces, hasta diez veces superior, para las empresas chinas frente al resto. Ello pone de manifiesto la conveniencia de emplear esta fuente para dibujar el mapa competitivo del sector en China. Al tratarse de un mercado específico, se observa que dos empresas relevantes desde el punto de vista del mercado global no presentan co-enlaces con el resto, lo cual evidencia la escasa presencia de las mismas en el mercado chino. Se trata de información que proporciona una visión global del escenario competitivo en una industria, al tiempo que para los conocedores del sector proporciona una nueva perspectiva sobre el mismo.

En un segundo trabajo, Vaughan y You (2008) profundizan en el tipo de análisis realizado, enriqueciendo la consulta efectuada para la obtención de datos con un componente semántico que permite filtrar el contenido de las páginas que no incluyen el término clave indicado. Este método combina una doble perspectiva, cuantitativa y cualitativa, permitiendo precisar el ámbito de estudio. Se analiza un subsector de la industria de telecomunicaciones, el WiMAX (*Worldwide Interoperability for Microwave Access*). Las razones de la elección de este subsector son:

- Se conoce de manera global por su acrónimo WiMAX, lo cual permite realizar una labor de filtrado muy precisa. Al tratarse de un término que no posee otras referencias asociadas, se evita el ruido en los resultados obtenidos. Así, en principio, se excluirán de la investigación todos aquellos sitios web que co-enlacen dos empresas por razones ajenas al WiMAX. Este motivo evidencia las dificultades para emplear este método a un área determinada que no esté identificada claramente por un término clave que sea preciso y no polisémico.

- Hay una amplia variedad de empresas que desempeñan su actividad en esta área, de modo que se otorga al método empleado una buena oportunidad de mostrar su poder discriminante.

El estudio se centra en 39 empresas identificadas a partir de un informe sobre la situación en el subsector. En esta ocasión el motor de búsqueda seleccionado para la investigación es el de Microsoft, ya que tanto Google como Yahoo! presentan restricciones que impiden realizar busquedas de co-enlaces o combinarlas con un término clave adicional (WiMAX). Se han realizado dos conjuntos de búsquedas que han dado lugar a dos matrices de datos distintas: una búsqueda sin incluir la palabra clave y otra incluyéndola. Los términos de búsqueda empleados en la obtención de datos se pueden observar en la Tabla 3.6.

Tabla 3.6. Términos de consulta empleados por Vaughan y You (2008)

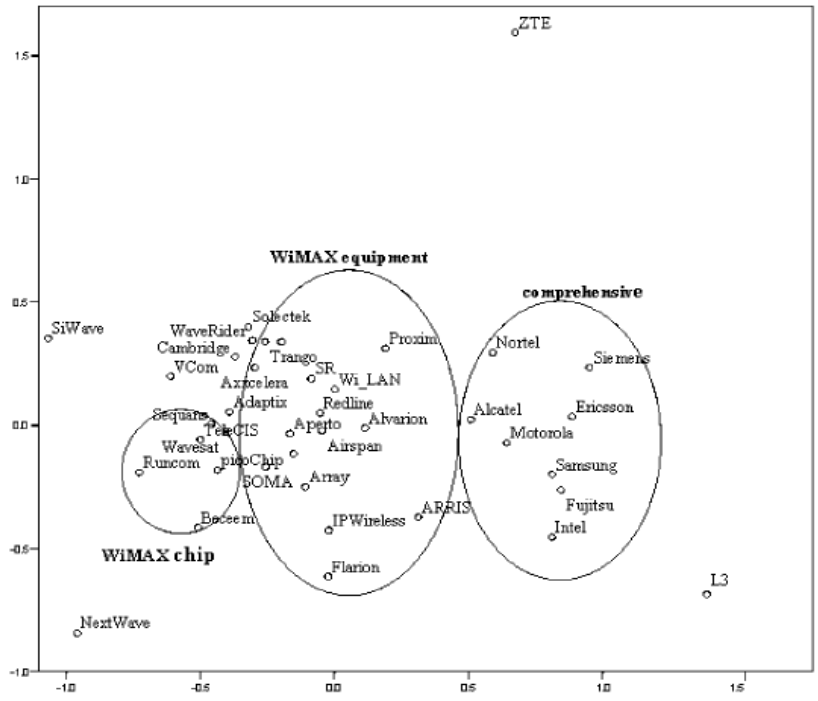
Tipos de datos recogidos	Consulta
Sin la palabra clave	(link:www.abc.com-site:abc.com) AND (link:www.xyz.com-site:xyz.com)
Con la palabra clave WiMAX	((link:www.abc.com-site:abc.com) AND (link:www.xyz.com-site:xyz.com)) WiMAX

Siguiendo los mismos procedimientos estadísticos utilizados en el artículo anterior, se obtienen dos mapas. Un primer mapa (Figura 3.4), en el que no se incluye la palabra clave WiMAX, permite establecer tres grupos de empresas que reflejan la fortaleza global de las mismas en el sector de las telecomunicaciones. Los tres grupos se identifican como:

- "comprehensive", que incluye las grandes firmas de telecomunicaciones (Ericsson, Alcatel, Nortel) con una amplia variedad de productos;
- "WiMAX chip", que comprende empresas dedicadas a dicha actividad (Beceem, picoChip, Runcom); y,

- "WiMAX equipment", con empresas especializadas en la producción de equipos (Alvarion, Airspan, Proxim). Los resultados coinciden con la clasificación llevada a cabo por el informe del sector tomado como referencia.

Figura 3.4. Mapa del sector de telecomunicaciones sin emplear el término "WiMAX" (Vaughan y You, 2008: 438)

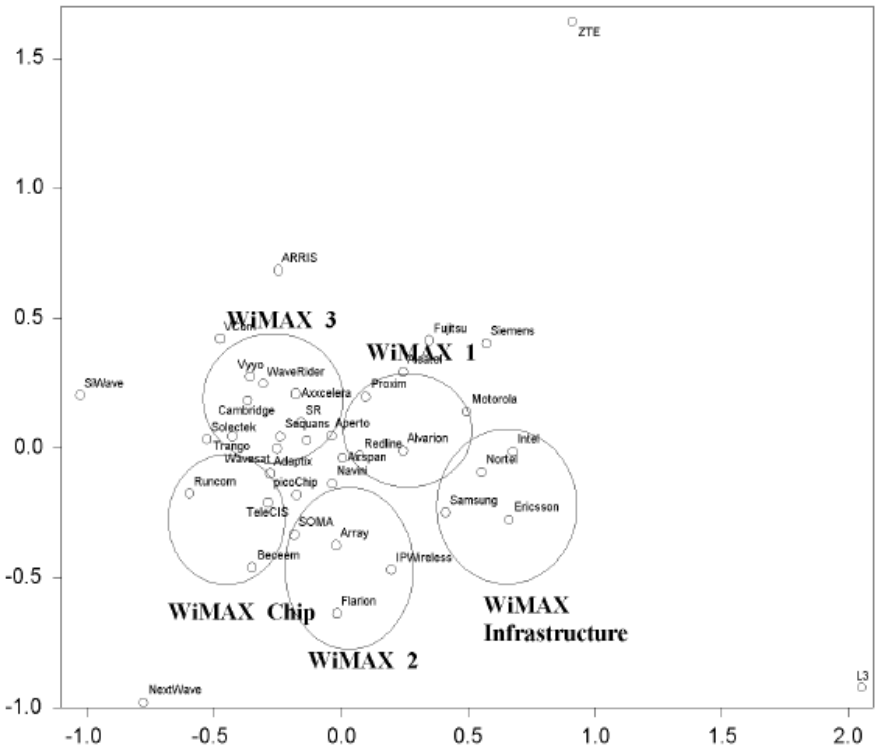


El segundo mapa (Figura 3.5) se obtiene a partir de las búsquedas realizadas, combinando los co-enlaces con el término "WiMAX". En este caso, las compañías se agrupan en cinco conjuntos que muestran niveles distintos de competencia en el sector. Los grupos identificados son los siguientes:

- "WiMAX 1": empresas de equipos líderes en el mercado WiMAX.

- "WiMAX 2": grupo de grandes firmas de equipos entre las cuales la competencia es más intensa que en las empresas en "WiMAX 1".
- "WiMAX 3": empresas de WiMAX más pequeñas que los dos anteriores grupos.
- "WiMAX chip": compañías centradas en el negocio de los chips para WiMAX.
- "WiMAX infraestructure": empresas que proporcionan infraestructuras para redes WiMAX, incluyendo grandes compañías como Ericsson, Nortel, Samsung e Intel.

Figura 3.5. Mapa del sector de telecomunicaciones empleando el término "WiMAX" (Vaughan y You, 2008: 439)



Estos resultados son consistentes con otros estudios realizados sobre el sector, por lo que los autores subrayan la grandes posibilidades del método empleado.

Vaughan y You (2009) aplican el mismo tipo de análisis a un producto concreto dentro del sector de las telecomunicaciones, en concreto el DSLAM, que se emplea en el acceso a Internet de banda ancha. Los resultados confirman la viabilidad del método empleado. Vaughan, Tang y Du (2009) investigan, en la misma línea, en el sector químico y el sector de electrónica en China.

Desde una perspectiva cualitativa, Vaughan, Gao y Kipp (2006) analizaron las motivaciones para enlazar páginas web comerciales. Los resultados concluyen que la mayoría de páginas web que establecen enlaces tienen carácter comercial (71,3%), seguidos de otros tipos como: organizacional (15,6%), educativo (5,2%), personal (6,1%) y gubernamental (1,9%). Esto confirma que los enlaces están creados por páginas que contienen información empresarial. En relación con las razones para enlazar destaca la presencia de directorios *online* (22,5%), listados de empresas (19,6%) y artículos en prensa y noticias (12,4%). Un último grupo considerado como "Otros", que incluye blogs, listados de favoritos, revisiones de productos, boletines, etc., representa el 20,3% del total, subrayando la creciente importancia de los contenidos creados por los propios usuarios o destinados de forma directa a ellos. Destaca de forma muy significativa la casi nula presencia de enlaces a empresas competidoras, lo cual confirma la evidencia mostrada por otros trabajos que señalan la oposición de las empresas a desviar la atención del visitante hacia los competidores (Bentley, 1997; Shaw, 2001).

Dado que tanto los directorios online como los listados de empresas son propicios para el establecimiento de co-enlaces, los autores llevan a cabo un trabajo adicional (Vaughan, Kipp y Gao, 2007a) en el que efectúan un análisis de contenidos de páginas web que establecen co-enlaces, en concreto de las mismas empresas estudiadas en Vaughan y You (2006). En este caso se distingue entre co-enlaces a las páginas de inicio (*homepages*) de las empresas y los coenlaces a otras páginas distintas dentro del sitio de la empresa.

Los resultados obtenidos según el tipo de sitio web que establece los co-enlaces muestran, de mayor a menor presencia, los siguientes resultados: comercial, 68,5%; organizacional, 12,5%; personal, 12,5%, educativo, 5,7%, gubernamental, 0,8%. Si atendemos a las razones para co-enlazar, los co-enlaces *estrechamente relacionados* con motivos empresariales (relativas a productos, empresas o servicios) representan el 61,4% del total frente al 23,8% de co-enlaces *marginalmente relacionados* y al 14,7% de co-enlaces *no relacionados*. En relación con el estudio descrito anteriormente (Vaughan, Gao y Kipp, 2006), que señalaba que el 20% de los *inlinks* no estaban vinculados a sitios comerciales, la cifra en este caso baja al 14,7%, lo cual indica que los co-enlaces están más relacionados con sitios comerciales que los enlaces directos, por lo que pueden ser empleados con mayor eficacia en determinados estudios webmétricos sobre empresas.

4. INVESTIGACIÓN EMPÍRICA

> *«¿Cómo no someterse a Tlön, la minuciosa y vasta evidencia de un planeta ordenado? Inútil responder que la realidad también está ordenada. Quizá lo esté, pero de acuerdo a las leyes divinas –traduzco: a leyes inhumanas– que no acabamos nunca de percibir. Tlön será un laberinto, pero es un laberinto urdido por hombres, un laberinto destinado a que lo descifren los hombres.»*
>
> *Jorge Luis Borges*
>
> Tlön, Uqbar, Orbis Tertius *(1944/2005b: 442)*

La investigación empírica se divide en tres secciones. La primera (Apartado 4.1) recoge trabajos basados en el análisis de impacto de enlaces; la segunda (Apartado 4.2) incluye estudios basados en el análisis de co-enlaces; por último, la tercera (Apartado 4.3) contiene una propuesta de investigación de un único sector empresarial, el de la banca internacional, a través de la combinación de los dos anteriores tipos de análisis. Con ello se pretende mostrar los avances hechos en estos campos, los más fértiles por el momento en la investigación webmétrica de empresas, así como abordar una propuesta para llevar a cabo una aplicación práctica de estas técnicas para el análisis de sectores de actividad específicos.

4.1. Análisis de impacto de enlaces

Como se ha desarrollado en el Apartado 3.4, los estudios centrados en el análisis de impacto de enlaces han explorado las relaciones entre el número de enlaces que apuntan a un determinado sitio o página web y otras variables vinculadas a la actividad de la entidad a la que hace referencia. En la revisión de la literatura, hemos podido comprobar que este tipo de investigación se ha aplicado con éxito a ámbitos como la producción científica o la empresa. En este apartado se incluyen distintas investigaciones que amplían la evidencia encontrada hasta el momento en la búsqueda de relaciones entre variables empresariales y variables web, en concreto, el número de enlaces recibidos.

Los siguientes trabajos pretenden ser los primeros que, desde el ámbito de los estudios de empresa, aborden las relaciones entre la variable "número de enlaces recibidos por el sitio web de una empresa" y una serie de variables financieras. Su oportunidad es doble. Por un lado, pone de manifiesto relaciones entre variables no exploradas anteriormente en contabilidad y finanzas, ampliando, desde una perspectiva interdisciplinar, una línea de investigación que, aplicada a otras áreas de conocimiento, se ha desarrollado con éxito durante los últimos años en el campo de Ciencias de la Información. Por otro, presenta nuevas fuentes de información que permiten la apertura de perspectivas adicionales para la investigación, especialmente en el campo de las nuevas tecnologías.

Los dos primeros trabajos incluidos en este apartado intentan dar respuesta a la primera pregunta de investigación señalada en el Apartado 1.2.1: *¿En qué medida se relacionan las variables financieras clave de una empresa con su presencia en la web, medida a través del número de enlaces que reciben sus sitios web?*

Los trabajos previos (Vaughan 2004a; 2004b; Vaughan y Wu, 2004) se han concentrado geográficamente en empresas de Canadá, China y Estados Unidos. El ámbito europeo aún no ha sido abordado. Adicionalmente, la regulación mercantil en la Unión Europea nos permite disponer de datos públicos de empresas no cotizadas, con lo cual es posible comprobar qué ocurre con las relaciones entre

variables cuando estudiamos empresas de tamaño más reducido. El Apartado 4.1.2 incluye un trabajo en el que se analizan empresas españolas y británicas pertenecientes a distintos sectores de actividad.

Por otra parte, las investigaciones previas se han limitado fundamentalmente al sector de tecnologías de la información, el cual por su propia naturaleza tiene una presencia destacada en la Web. De momento se desconoce si las correlaciones observadas se producen también en otros sectores de actividad. En este sentido, el Apartado 4.1.1 estudia diversos sectores empresariales dentro de los Estados Unidos, país de referencia en este campo al contar con una amplia penetración de Internet y ser también líder en la industria de los motores de búsqueda. El trabajo sobre empresas españolas y británicas también es relevante en este sentido.

Un tercer trabajo incluido en el Apartado 4.1.3 pretende contestar la segunda pregunta de investigación: *¿Cuáles son las variables financieras que explican el número de enlaces que recibe el sitio web de una empresa?*

Se han realizado algunos análisis de regresión para intentar mostrar cuáles son las variables que explican el número de enlaces recibidos por los sitios web académicos. Por ejemplo, Vaughan y Thelwall (2005) investigan las universidades en Canadá. Sin embargo, en el ámbito de los sitios web de empresas esta perspectiva aún no ha sido muy explorada. Con el objeto de llenar este vacío se ha llevado a cabo un análisis de regresión múltiple sobre empresas de España y Reino Unido, las mismas utilizadas en el Apartado 4.1.2.

4.1.1. Análisis de impacto de enlaces en diversos sectores empresariales en los Estados Unidos

En esta sección se analizan cinco conjuntos de empresas que pertenecen a distintos sectores de actividad en los Estados Unidos. También se tiene en consideración al conjunto de empresas que componen el índice *Dow Jones Industrial*. Mientras que en los cinco grupos las empresas son homogéneas en

cuanto a la actividad que desempeñan, no ocurre lo mismo con las empresas del índice *Dow Jones*. Este grupo servirá de referente para comprobar si las correlaciones significativas encontradas sólo se producen entre empresas con un mismo perfil de actividad o, por el contrario, si este factor no es relevante. La evidencia mostrada por el anterior trabajo de Vaughan y Wu (2004) apunta a que únicamente en conjuntos de empresas que realizan la misma actividad y, por tanto, están sujetos a modelos de negocio similares y a una exposición a Internet parecida, se detectan correlaciones significativas entre número de enlaces y variables económico-financieras.

Al margen, el presente trabajo pretende comprobar hasta qué punto las relaciones sugeridas entre las variables se verifican en sectores con una presencia menos intensa en Internet.

4.1.1.1. Método

En primer lugar se aborda la descripción de las empresas seleccionadas para la investigación. La elección de Estados Unidos como país de referencia para realizar la primera investigación que tiene en cuenta múltiples sectores empresariales se debe a las siguientes razones:

- Es un país donde la presencia comercial de las empresas en la Web es muy alta. Además, Estados Unidos es un país pionero en el desarrollo de nuevas tendencias en el campo tecnológico y especialmente de Internet.

- Se trata de la principal economía mundial, con lo que es más probable que las empresas cuenten con la dimensión y envergadura suficientes para tener en la Web un fiel reflejo de sus actividades, su desempeño y su situación financiera.

- Se trata de uno de los países cuyas empresas de telecomunicaciones han sido incluidas en trabajos previos (Vaughan, 2004a; 2004b), lo cual permite

establecer comparaciones con los resultados obtenidos anteriormente.

Las empresas seleccionadas cotizan en bolsa, ya que estas entidades están obligadas en los Estados Unidos a publicar su información financiera. Los sectores considerados en el trabajo son los siguientes: Bancos (*Commercial banks*), Construcción (*Construction of Buildings*), Comercio general (*General Merchandise Stores*), Minería (*Mining*) y Servicios (*Utilities*). Al margen, como se ha indicado, se incluye el conjunto de empresas que conforma el índice *Dow Jones Industrial*.

La Tabla 4.1 describe la muestra de empresas en el estudio, obtenida a partir de la base de datos Mergent (www.mergent.com). La segunda columna incluye el total de empresas recogidas por la base de datos para el código NAICS (sistema de clasificación de actividades comerciales en Estados Unidos) indicado, mientras que la tercera columna recoge el número de esas empresas que se encuentran en activo. La muestra analizada está formada por cien empresas de cada sector, siempre que el número de empresas activas fuera superior. En caso contrario, se ha tomado toda la población. Posteriormente se ha procedido a excluir algunas empresas que no cumplen con las condiciones exigidas, que son:

- Las empresas cuentan con página web.
- Disponer de datos financieros para el ejercicio 2007, que es el último de los años que está disponible en la base de datos.
- Las empresas no comparten una misma dirección web, generalmente por tratarse de empresas pertenecientes a un mismo grupo.

Por último, una vez efectuadas las correspondientes exclusiones, se indica el número final de empresas válidas de la muestra para cada sector y el porcentaje que supone sobre el total de empresas activas en el mismo.

Tabla 4.1. Muestra descriptiva de las empresas de Estados Unidos incluidas en el estudio

Sectores comerciales (código NAICS)	Total de empresas (activas)	Tamaño de la muestra	Exclusión (falta de página web)	Exclusión (falta de información)	Empresas válidas	% del total empresas "activas"
Bancos (52211)	1.099 (581)	100	6	-	94	16 %
Construcción (236)	88 (57)	57	16	17	24	42 %
Comercio general (452)	71 (35)	35	6	5	24	69 %
Servicios (221)	426 (275)	100	15 + 3 (por compartir URL)	7	75	27 %
Minería (21)	897 (671)	100	21	26	53	8 %
Dow Jones Industrial	30	30	-	-	30	100 %

La información en la tabla se ha obtenido de la base de datos Mergent entre el 30 de enero y el 5 de febrero de 2009, periodo en el que se efectúan las consultas. La composición del índice *Dow Jones Industrial* corresponde al 4 de diciembre de 2008 (consultado en http://www.djindexes.com/mdsidx/?event=showAverages). La información financiera empleada en el estudio se ha tomado de la citada base de datos, incluyendo, entre otras, las siguientes variables: número de empleados, resultado, ingresos, volumen de activos y pasivos, ROA (*Return on assets*) y ROE

(*Return on equity*). Los datos se han obtenido para los tres últimos años disponibles 2005, 2006 y 2007.

Por su parte, en relación con la cuantificación del número de enlaces recibidos, se emplea Google para determinar las páginas web de las empresas seleccionadas en la muestra. En aquellos casos en los que una empresa cuenta con varios dominios se selecciona aquel que mayor número de enlaces recibe. La Tabla 4.2 indica las consultas empleadas para la recuperación de los datos de enlaces, indicando las fechas y los métodos empleados.

Tabla 4.2. Momento y modo de obtención del número de enlaces recibidos

Sintaxis (Yahoo!)	linkdomain:abc.com -site:abc.com	linkdomain:abc.com -site:abc.com	link:http://www.abc.com -site:abc.com
Método de recogida de datos	Interfaz web del buscador	API de Yahoo!	API de Yahoo!
Sector de actividad (código NAICS)	Fecha de recogida de datos (en los tres casos fueron obtenidos en Canadá)		
Bancos (52211)	30-01-2009	06-05-2009	06-05-2009
Construcción (236)	02-02-2009	06-05-2009	06-05-2009
Comercio general (452)	02-02-2009	06-05-2009	06-05-2009
Servicios (221)	01-02-2009	06-05-2009	06-05-2009
Minería (21)	05-02-2009	06-05-2009	06-05-2009
Dow Jones Industrial	08-02-2009	No disponible	No disponible

El empleo de Yahoo! como fuente de obtención de datos se debe a que es el motor de búsqueda que proporciona las mejores opciones para el diseño de consultas complejas en el momento de la recogida de la información, así como a que ofrece resultados más completos que Google y MSN, los otros dos grandes competidores.

Se han empleado dos consultas para realizar el trabajo:

- *linkdomain:abc.com –site:abc.com*
- *link:http://www.abc.com –site:abc.com*

La primera está basada en el operador *linkdomain*, que proporciona todos aquellos enlaces (URLs) que apuntan a cualquier página perteneciente al dominio especificado. La segunda, por su parte, emplea el operador *link*, que proporciona las páginas que enlazan únicamente a la dirección URL especificada. Los resultados ofrecidos por el primero son más numerosos ya que incluyen los enlaces que apuntan a la URL principal del dominio de la empresa más todos aquellos que apuntan a URLs dentro de directorios del dominio o de subdominios. Si bien los resultados que producen ambos operadores son similares, en el trabajo se utiliza como referencia el operador *linkdomain*, ya que conceptualmente recoge las referencias hechas por otras páginas al sitio web de la empresa de forma más completa. Al margen, cabe indicar que el componente de la consulta *-site:abc.com* tiene por objeto eliminar de los resultados obtenidos aquellos que proceden del mismo dominio, es decir, auto-referencias, las cuales, al ser creadas por la propia empresa, conllevan un sesgo en los resultados.

Para obtener los datos se han empleado dos sistemas. Por un lado, se ha empleado la interfaz web del motor de búsqueda, esto es, se han introducido en Yahoo! las consultas correspondientes y se han anotado los resultados ofrecidos. Por otro, se ha empleado la API de Yahoo!. Si bien los resultados parecen diferir en función del empleo de la interfaz web o de la API, las diferencias no son significativas en cuanto a las implicaciones que pudieran tener en el análisis.

En trabajos anteriores (Vaughan 2004a; 2004b), se tuvieron en cuenta variables adicionales, como por ejemplo, la edad de la página web. En esta investigación, sin embargo, no se ha considerado relevante ya que la evidencia muestra que los resultados no difieren en función de si se emplea o no esta variable.

4.1.1.2. Resultados

Dado que la mayoría de las variables en el estudio, especialmente las correspondientes a enlaces recibidos, presentan una distribución muy asimétrica, no satisfaciendo el supuesto de normalidad, se opta por utilizar el coeficiente de correlación de Spearman con el fin de determinar si existen relaciones significativas entre la variable de enlaces recibidos y las variables económico-financieras.

De las tres mediciones de enlaces que se han realizado para el estudio, los resultados obtenidos mediante el operador *linkdomain* proporcionan una correlación significativa y más alta que los derivados del empleo de *link*. Por otra parte, salvo excepciones, en la mayoría de los casos las correlaciones con los datos de enlaces utilizando el operador *linkdomain*, tomados entre enero y febrero, son más altas que con los datos de mayo. En cualquier caso, puede observarse en la Tabla 4.3 que las correlaciones entre los distintos conjuntos de datos son muy elevadas, lo cual indica que los resultados que proporcionan los buscadores son consistentes y relativamente estables en el tiempo.

Tabla 4.3. Correlaciones entre el número de enlaces a principios de año (enero y febrero 2009) y en el mes de mayo de 2009

	Mayo 2009 "linkdomain"	Mayo 2009 "link"
Bancos	,964**	,903**
Construcción	,923**	,890**
Minería	,940**	,949**
Comercio general	,988**	,977**
Servicios	,982**	,973**
** Coeficientes de correlación significativos al nivel 0,01. Se toma como referencia el número de enlaces obtenido entre los meses de enero y febrero de 2009 empleando el operador *linkdomain*.		

La Tabla 4.4 muestra los coeficientes de correlaciones entre los enlaces recibidos y las variables económico-financieras consideradas en el trabajo. Para cada sector se tiene en cuenta, en columnas, la primera medición de enlaces realizada entre enero y febrero y la segunda correspondiente al mes de mayo. En ambos casos se empleó el operador *linkdomain*. En filas, para cada variable se incluyen los datos correspondientes a los últimos tres ejercicios para los que se dispone de información. Por lo general, los coeficientes de correlación para los distintos años no varían demasiado debido a que las variables contables de un año a otro también están muy correlacionadas.

En función de lo apuntado anteriormente, vamos a tomar, como referencia para el análisis, los coeficientes de correlación entre número de enlaces correspondientes a la medición de principios del año 2009 y las variables contables de 2007, último ejercicio para el que hay datos.

Tabla 4.4. Coeficientes de correlación de Spearman para los distintos grupos de empresas

		Bancos		Construc-ción		Minería		Comercio general		Servicios		Dow Jones
		1ª	2ª	1ª	2ª	1ª	2ª	1ª	2ª	1ª	2ª	1ª
Emplea-dos	2007	,741**	,699**	,292	,335	,522**	,559**	,845**	,868**	,855**	,847**	,239
Activos	2007	,736**	,696**	,129	,136	,503**	,534**	,868**	,884**	,837**	,849**	,206
	2006	,719**	,678**	,165	,114	,495**	,498**	,861**	,878**	,794**	,807**	,201
	2005	,744**	,708**	,255	,189	,457**	,532**	,857**	,873**	,759**	,763**	,191
Resulta-do Neto	2007	,627**	,623**	,099	,134	,106	,175	,941**	,929**	,794**	,804**	,319
	2006	,682**	,663**	,218	,140	,116	,196	,897**	,889**	,711**	,709**	,133
	2005	,719**	,684**	,102	,069	,114	,188	,921**	,918**	,576**	,553**	,102
Ingresos	2007	,745**	,707**	,240	,242	,525**	,530**	,886**	,908**	,801**	,809**	,114
	2006	,730**	,684**	,181	,161	,466**	,489**	,863**	,885**	,786**	,790**	,093
	2005	,750**	,711**	,224	,141	,471**	,514**	,839**	,845**	,767**	,771**	,100
EBITDA	2007	-	-	,075	,075	-	-	,944**	,935**	,820**	,826**	,385*
	2006	-	-	,246	,207	-	-	,943**	,939**	,737**	,731**	,379
	2005	-	-	,271	,194	-	-	,901**	,894**	,679**	,662**	,255
ROA	2007	,134	,183	,133	,181	,237	,274	,560**	,476*	,389**	,404**	,123
	2006	,278*	,304**	,074	,005	,101	,197	,570**	,524*	,344**	,316**	-,015
	2005	,242*	,246*	,197	,165	-,114	-,027	,616**	,590**	,194	,166	,031
ROE	2007	,154	,174	,344	,375	,227	,163	,649**	,573**	,050	,032	,023
	2006	,298**	,290**	,174	,081	,236	,206	,671**	,622**	,179	,137	-,119
	2005	,221*	,207	,197	,132	-,002	,015	,626**	,567**	,049	,011	-,048

* Coeficientes de correlación significativos al nivel 0,05.

** Coeficientes de correlación significativos al nivel 0,01.

En columnas, 1ª y 2ª hace referencia a la primera medición del número de enlaces realizada entre enero y febrero y a la segunda medición correspondiente al mes de mayo, respectivamente. Los guiones indican que no existen datos disponibles.

La Tabla 4.5 indica la posición que ocupa cada sector en cuanto al coeficiente de correlación entre número de enlaces y cada una de las variables contables, siendo 1 el sector con el coeficiente de correlación más elevado.

Tabla 4.5. Posición que ocupa cada sector en función del coeficiente de correlación entre variables

	Emplea-dos	Activos	Resultado Neto	Ingresos	EBITDA	ROA	ROE
Bancos	3	3	3	3	N/A	-	-
Construcción	-	-	-	-	-	-	-
Minería	4	4	-	4	N/A	-	-
Comercio general	2	1	1	1	1	1	1
Servicios	1	2	2	2	2	2	-
El sector con el número 1 es aquel que presenta el coeficiente de correlación más alto entre el número de enlaces y la variable indicada. En todos estos casos, el coeficiente de correlación es signicativo al nivel 0,01. El guión indica que la relación encontrada no es signicativa y N/A significa que no se dispone de datos.							

A continuación se examinan los resultados para cada uno de los grupos de empresas considerados.

- **Bancos**

En el sector bancario las correlaciones son significativas para las principales variables, tanto de posición financiera (activos, 0,736) como de desempeño empresarial (resultado neto, 0,627; ingresos, 0,745). Sin embargo, para el año

2007 no se verifica la existencia de una relación significativa entre los enlaces recibidos y la variables de desempeño financiero en términos relativos, como son los ratios ROA y ROE. En relación con los datos de 2005 y 2006, el coeficiente de correlación es muy bajo, pero sí es significativo. Ello parece indicar que, en este sector, los enlaces constituyen un claro indicador de la dimensión de la entidad, si bien no se puede decir lo mismo en cuanto a su rentabilidad.

- **Construcción**

En el sector de la construcción no se han identificado relaciones significativas entre las variables en ninguno de los casos.

- **Minería**

En el sector de la minería, existen relaciones significativas con determinadas variables en términos absolutos: empleados (0,522), activos (0,503) e ingresos (0,525). En el caso de los ratios de rentabilidad no existe relación significativa.

- **Comercio general**

En este sector, los coeficientes de correlación entre las variables estudiadas son significativos y altos en todos los casos. Cabe destacar que si bien para variables en términos absolutos los coeficientes rondan entre el 0,85 y el 0,95, en el caso de los ratios de rentabilidad disminuyen hasta situarse para el ROA en 0,56 y para el ROE en 0,649. Estos resultados coinciden con el hecho de que las empresas en la muestra son las que mayor número de enlaces reciben por término medio, lo cual puede explicarse por el hecho de que se trata de un sector muy orientado hacia el usuario y, por tanto, también objeto de mayor atención en Internet.

- **Servicios**

En este caso, las variables empleadas -activos, resultado neto, ingresos y EBITDA- presentan coeficientes de correlación significativos y altos. En el caso del ROA el coeficiente es significativo (0,389), sin embargo el ROE no presenta ningún tipo de relación que sea estadísticamente significativa.

- ***Dow Jones Industrial***

Por último, de manera congruente con lo mostrado por Vaughan y Wu (2004), no encontramos ningún tipo de relación significativa al tratarse de conjuntos de empresas no homogéneas en términos de actividad.

4.1.1.3. Conclusiones

Los resultados analizados grupo a grupo indican que, en la mayoría de los sectores, por muy diversos que éstos sean, existe una vinculación entre el número de enlaces que reciben las páginas web de las empresas y su dimensión económica, medida a través de sus variables de posición y desempeño financieros. Únicamente el sector de la construcción no presenta ningún tipo de relación. Ello puede deberse a que dicho sector aún no hace un uso generalizado de Internet. Al margen debe de considerarse como factor importante la naturaleza de la actividad y el modo en que ello condiciona la presencia de las empresas en la Web. Las empresas de comercio general están muy enfocadas al usuario final y, por tanto, es lógico pensar que la atención que reciben en la Web permite reflejar de una manera más adecuada la situación real de las mismas, aunque también, debido al mayor volumen de información, es posible que exista un mayor "ruido" a la hora de determinar qué información es relevante de cara a los propósitos de la investigación. Esta orientación al usuario final se traduce en que este sector es el que mayor número de enlaces recibe entre todos los sectores. Cabe también subrayar que la

existencia de correlaciones entre variables únicamente parece verificarse entre grupos homogéneos de empresas, como corroboran los datos correspondientes al índice Dow Jones y las investigaciones anteriores.

La Tabla 4.5 permite observar los sectores para los cuáles el número de enlaces parece constituir un indicador más efectivo. El sector Comercio general es el que presenta correlaciones más elevadas para todas las variables contables, seguida de Servicios. Sin embargo, el primero es el único que muestra unos coeficientes de correlación significativos para los ratios ROA y ROE. El sector bancario se sitúa en tercera posición, mientras que el de la minería en cuarta. Cuantos mayores son los coeficientes de correlación también es mayor el número de variables con el que se detectan relaciones significativas. Así, mientras que las empresas de comercio general tienen correlaciones significativas con las siete variables consideradas, las empresas de minería sólo con tres. Ello parece indicar que sería posible identificar una serie de características propias de cada sector que permitan establecer distintos grados de relevancia a la hora de emplear los enlaces como un indicador complementario sobre la dimensión o el desempeño de la empresa.

El número de empleados presenta un comportamiento y unos coeficientes de correlación similares a los del volumen de activos en todos los sectores, demostrando ser una variable que pone de manifiesto de forma adecuada la dimensión de la empresa con independencia del sector en que nos encontremos.

Al margen, hay que apuntar que las variables financieras están altamente correlacionadas entre un ejercicio y otro. Igual ocurre con las mediciones de enlaces realizadas con cuatro meses de diferencia (Tabla 4.3). Se debe señalar que, por lo general, el número de enlaces suele ser estable, no cambiando de manera considerable en periodos cortos de tiempo.

La comparación con los resultados de los trabajos anteriormente realizados en el sector de tecnologías de la información (Vaughan, 2004a; 2004b; Vaughan y Wu, 2004) permite destacar que los coeficientes de correlación, cuando son significativos, se han incrementado. Una de las posibles razones sería el gran desarrollo de Internet en los últimos cinco años, lo cual ha motivado que la actividad

en la Web y fuera de ella se haya aproximado más en determinados contextos. El surgimiento y desarrollo de la Web 2.0 (O'Reilly, 2005) constituye un importante factor en este contexto.

Entre las limitaciones del trabajo podemos identificar algunas provenientes de la elección del país, ya que nos ha obligado a limitar el estudio a entidades cotizadas, que son aquellas para las que existe información financiera disponible. Esto ha hecho que el perfil de las empresas sea de dimensión media o grande y que en determinados sectores el número de empresas considerado en la muestra sea pequeño. En cualquier caso, como ya señalamos, esta elección se ha fundamentado en que, tomando como referencia los estudios previos (Vaughan, 2004a; 2004b), Estados Unidos es el país de referencia sobre el que establecer comparaciones con los resultados obtenidos en ellos.

Por otro lado, debe señalarse que la comparación entre enlaces recogidos a principios de 2009 y datos financieros correspondientes al ejercicio 2007, los cuáles estaban disponibles a mediados de 2008, se debe a la falta de datos empresariales más actuales, en el momento de la realización del trabajo. Sin embargo dado el alcance del estudio y considerando el elevado grado de correlación entre las variables contables de un grupo de empresas de un mismo sector de un año a otro, se pueden alcanzar los objetivos planteados, es decir, establecer evidencia exploratoria de la correlación entre dos variables de cara a facilitar su empleo en trabajos de investigación posteriores en esta línea. Idéntica situación se ha producido en la investigación desarrollada previamente. Cabe destacar que en la presente tesis se incluye un trabajo, en el Apartado 4.3.1, donde se abordan los principales bancos a nivel mundial en el que se emplean datos de enlaces correspondientes a diciembre de 2008 y datos financieros correspondientes al mismo ejercicio.

4.1.2. Análisis de impacto de enlaces en diversos sectores empresariales en España y Reino Unido

El objetivo de este trabajo es obtener evidencia de las relaciones existentes entre el número de enlaces recibidos por los sitios web de empresas y sus variables económico-financieras en el contexto europeo. Hasta ahora todas las investigaciones realizadas se han centrado únicamente en Estados Unidos, Canadá y China. El ámbito europeo constituye un reto especial dado que confluyen un gran número de países con diferentes economías, lenguas y culturas que comparten un mercado único con una normativa relativamente armonizada o, en algunos aspectos, común. Además, esta investigación tiene especial importancia porque, a diferencia de los Estados Unidos, en Europa la información financiera de empresas no cotizadas es pública.

4.1.2.1. Método

Las empresas seleccionadas pertenecen a cinco sectores que desarrollan actividades económicas muy diferentes, y que tienen un grado de presencia en la Web también heterogéneo. Únicamente se han tenido en cuenta dos países europeos, España y Reino Unido. Ambos países constituyen dos de las mayores economías de la Unión Europea y además emplean dos de las lenguas más importantes a nivel mundial, el español y el inglés. La información relativa a las empresas en el estudio se obtuvo en octubre de 2009 de la base de datos Amadeus (www.bvdep.com/en/amadeus.html), que contiene información financiera de empresas europeas, tanto cotizadas como no cotizadas. Los casos disponibles se han filtrado con el objeto de que los mismos cumplan con los siguientes requisitos:

- El último año de información financiera disponible es 2007.
- Todas las empresas tienen, al menos, información correspondiente a las variables: activos totales y resultado después de impuestos.

- Las empresas tienen página web, de acuerdo con la información contenida en la base de datos.

La Tabla 4.6 muestra los sectores incluidos en el trabajo, su código NAICS correspondiente a la clasificación sectorial y el número de empresas para cada país.

Tabla 4.6. Distribución de las empresas por país y sector

Sectores (Código NAICS 2007)	España	Reino Unido	Total
Construcción (*Construction*) (23)	93	103	196
Hostelería y restauración (*Accommodation and Food Services*) (72)	70	180	250
Servicios (*Utilities*) (221)	56	60	116
Industria editorial (*Publishing Industries, except Internet*) (511)	72	201	273
Telecomunicaciones (*Telecommunications*) (517)	26	71	97
Total	317	615	932

La página web de cada empresa ha sido obtenida a partir de la base de datos Amadeus y posteriormente verificada al objeto de asegurar su corrección. Las empresas que compartían una página web con otras entidades han sido omitidas, a menos que una de ellas pudiera ser identificada de forma clara como la principal

prestadora de los servicios o la que tiene un tamaño significativamente mayor, medido a través de sus activos totales. En el caso de aquellas empresas que disponen de más de una URL, se han examinado todas ellas, seleccionando para la investigación aquella que presenta el mayor número de enlaces recibidos.

El conjunto de variables financieras tomadas de la base de datos Amadeus incluye las siguientes: activos totales, resultado después de impuestos, cifra de negocios y activos fijos intangibles. El número de casos disponibles para cada variable puede comprobarse en las Tablas 4.7, 4.9 y 4.10.

La cuantificación del número de enlaces recibidos por las páginas web se ha llevado a cabo utilizando Yahoo!, ya que ni Google ni MSN Bing proporcionan la información necesaria para llevar a cabo el trabajo. Entre las limitaciones que presentan estos grandes buscadores cabe señalar que la búsqueda de información sobre enlaces recibidos en Google únicamente proporciona una muestra de todos los enlaces indexados en la base de datos del buscador (Google, 2006). Además, los operadores para diseñar consultas que Google ofrece están más limitados, ya que no permiten filtrar los enlaces internos al no permitir combinar el operador *link* con otros operadores. Siendo esto así, no sería posible eliminar los enlaces internos o auto-referencias (por ejemplo, los enlaces creados para facilitar la navegación entre unas páginas y otras). Nuestro estudio requiere tener en consideración únicamente los enlaces externos creados fuera del propio sitio web, con el fin de evitar un sesgo en los datos. En el caso del buscador de Microsoft (ahora, Bing), anteriormente podía realizarse este tipo de búsquedas. Sin embargo las funciones para recuperar información sobre enlaces fueron limitadas en marzo de 2007 (Live Search, 2007). En noviembre de 2009, momento en que se obtienen los datos, Yahoo! representa la única opción viable para llevar a cabo este tipo de consultas requeridas para la investigación.

Dado que las versiones de los motores de búsqueda en cada país pueden disponer de bases de datos que primen a los sitios web de su ámbito geográfico al prestar una mayor atención a las páginas de su mercado local (Vaughan y Thelwall, 2004), se ha considerado el empleo de Yahoo! España y Yahoo! UK de forma respectiva

para las empresas de cada país. Sin embargo, las pruebas realizadas en ambas versiones de Yahoo! arrojan idénticos resultados, lo cual indica que la base de datos empleada es la misma, si bien las interfaces empleadas están adaptadas al mercado local. En tales circunstancias se optó por emplear Yahoo! UK para la cuantificación de los enlaces recibidos.

Como en la investigación incluida en el apartado 4.1.1, empleamos el operador *linkdomain* ya que todos los enlaces que apuntan a un dominio determinado son de relevancia para cuantificar la atención que este recibe. La sintaxis de la consulta efectuada es: *linkdomain:abc.com –site:abc.com*. Al objeto de recabar también los enlaces dirigidos a subdominios (por ejemplo, *mail.abc.com*), se omite el componente *www* de la URL. La parte *-site:abc.com* del término de búsqueda permite filtrar los enlaces internos creados en alguna de las páginas del dominio *abc.com*.

4.1.2.2. Resultados

Correlaciones por sector y por país de manera independiente

La funciones de distribución de las variables "enlaces recibidos" (Tabla 4.14 y Figura 4.1), así como de las variables financieras, son muy asimétricas, por lo que se opta por llevar a cabo el test de correlación de Spearman en lugar del de Pearson con el objetivo de explorar las relaciones entre ambos grupos de variables.

Los coeficientes de correlación entre el número de enlaces y las variables financieras de las empresas por sector y por país se incluyen en la Tabla 4.7. Las columnas bajo la etiqueta "sectores" muestran las correlaciones para empresas del mismo sector, sin distinguir el país; mientras que las columnas bajo la etiqueta "país" muestran las correlaciones para empresas pertenecientes a un mismo país, con independencia del sector de actividad al que pertenezcan. Como indican trabajos previos (Vaughan y Wu, 2004) así como los resultados del Apartado 4.1.1, las correlaciones son significativas cuando se tienen en consideración empresas de

un único sector. Por tanto, se espera que las correlaciones sean más significativas y elevadas cuando hablamos en términos de empresas homogéneas. La Tabla 4.7 muestra que aunque las correlaciones son significativas en ambos grupos de columnas, es decir, tanto cuando se agrupan por actividad como por país, los coeficientes son más elevados, para la mayor parte de variables, cuando atendemos al sector como criterio de homogeneidad.

Las correlaciones con la variable "activos fijos intangibles" no son significativas (caso de los sectores construcción y servicios) o son muy bajas. Sin embargo, en el caso de las industrias de publicación y de telecomunicaciones, que están muy centradas en información y tecnología, los coeficientes de correlación son los más elevados, 0,178 ($p<0,01$) y 0,250 ($p<0.05$), respectivamente. Esto podría apuntar hacia las debilidades en el reconocimiento y valoración de los activos intangibles, una de las cuestiones más relevantes en la regulación contable. En todo caso, aunque la presencia en la Web, medida en este caso a través del número de enlaces recibidos, es un importante recurso intangible para la empresa y podría ser también un indicador de la existencia de otros activos intangibles, no todos los activos fijos intangibles a los que se refiere la variable tienen porqué estar relacionados de algún modo con Internet.

Los coeficientes de correlación son altamente significativos cuando tomamos el sector como factor homogeneizador, con independencia de que empresas españolas y británicas se consideren de forma conjunta. Este mayor grado de correlación parece indicar que existe un cierto nivel de homogeneización entre empresas de un mismo sector, con independencia de su localización geográfica. Así, podemos indicar que el factor sector de actividad es un elemento homogeneizador más relevante que el factor país. Esta mayor homogeneidad entre empresas de un mismo sector puede ser consecuencia del proceso de globalización económica que, especialmente en ámbitos como la Unión Europea, hace que los mercados se encuentren cada vez más integrados. Este proceso es especialmente acusado en empresas de tamaño grande o muy grande, como son las que componen la muestra, si bien, la mayoría de ellas, no son empresas cotizadas. La pertenencia a la Unión Europea constituye un elemento importante a

tener en cuenta, dado que las empresas compiten en un único mercado europeo. En cualquier caso, vale la pena destacar el hecho de que mientras España es un país de la Eurozona, no ocurre lo mismo en el caso del Reino Unido.

La variable "activos totales" parece ser el mejor indicador del número de enlaces en los cinco sectores incluidos en el estudio, siendo el coeficiente de correlación más bajo el correspondiente al sector de la hostelería y restauración, 0,369 (p=0,001). La siguiente variable relevante es la "cifra de negocios". La correlación con "resultado después de impuestos" es significativa al nivel 0,1% para los sectores de construcción y publicación, al nivel 5% para hostelería y servicios, y no significativa para telecomunicaciones. El menor nivel de correlación para la variable "resultado después de impuestos", en comparación con otras variables, se podría explicar por la naturaleza de esta variable que puede ser tanto positiva como negativa, mientras que las otras tienen un límite mínimo de cero. El "resultado" es una de las mejores medidas del desempeño empresarial, dependiendo más de la evolución financiera de un año determinado. Aunque éste es también el caso de la cifra de negocios, esta magnitud permanece más estable, pudiendo incluso incrementarse, periodo tras periodo, aunque el resultado de la compañía descendiera o fuera negativo debido a unos elevados gastos.

Tabla 4.7. Correlación entre número de enlaces recibidos y variables financieras agrupadas de forma independiente por sectores y por país

Correlación rho de Spearman	Sectores					País	
	Construcción	Hostelería	Servicios	Industria editorial	Telecomunicaciones	España	Reino Unido
Cifra de negocios	,513***	,260***	,362***	,414***	,367***	,338***	,262***
N	185	235	112	265	95	314	578
Resultado después de impuestos	,305***	,143*	,219*	,267***	,090	,129*	,198***
N	196	250	116	273	97	317	615
Activos totales	,470***	,399***	,488***	,369***	,407***	,224***	,351***
N	196	250	116	273	97	317	615
Activos fijos intangibles	,140	,153*	,122	,178**	,250*	,317***	,221***
N	181	245	114	265	91	313	583

* Coeficientes de correlación significativos al nivel 0,05.

** Coeficientes de correlación significativos al nivel 0,01.

*** Coeficientes de correlación significativos al nivel 0,001.

Al objeto de determinar si existen diferencias significativas entre las variables de cada país se ha empleado el estadístico U de Mann-Whitney para las empresas incluidas en cada sector. La Tabla 4.8 indica el grado de significación del estadístico (con dos colas).

Tabla 4.8. Significación de los tests de Mann-Whitney para las diferencias de cada variable en España y Reino Unido

	Enlaces	Cifra de negocios	Resultado después de impuestos	Activos totales	Activos fijos intangibles
Construcción	,001	,001	,818	,100	,000
Hostelería y restauración	,727	,015	,046	,453	,000
Servicios	,012	,000	,008	,001	,005
Industria editorial	,004	,825	,986	,197	,070
Telecomunica-ciones	,208	,081	,195	,005	,081

La tabla muestra que existen diferencias significativas entre variables en función del país. Así, por ejemplo, se puede afirmar que existe una diferencia significativa entre el número de enlaces recibidos por las empresas en España y en Reino Unido para los sectores Construcción, Servicios y Editorial, mientras que no existe dicha diferencia en el caso de los sectores de Hostelería y Telecomunicaciones. A continuación se examinan las correlaciones en los distintos sectores, tomando España y Reino Unido de forma separada.

Correlaciones por sectores en España y Reino Unido

Como se señaló anteriormente, se espera que el nivel de correlación sea mayor para cada industria, considerada de manera independiente en cada país, que cuando todas las industrias en un país se consideran de manera agregada, mezclando empresas con distintos tipos de actividad (Tabla 4.7).

Los resultados en la Tabla 4.9 muestran que ésto es únicamente cierto en el sector

de la construcción y, únicamente para determinadas variables, en los otros sectores. La variable "activos totales" es la que está correlacionada de forma más significativa con el número de enlaces recibidos, mejorando en todos los casos los coeficientes de correlación incluidos para España en la Tabla 4.7. Cuando comparamos los resultados de la Tabla 4.9 con las correlaciones por sectores (es decir, mezclando empresas de ambos países) en la Tabla 4.7, detectamos que, en general, sólo los sectores de la construcción y las telecomunicaciones tienen niveles de correlación más altos cuando se analizan a nivel de España únicamente.

Tabla 4.9. Correlaciones entre el número de enlaces recibidos y los datos financieros por sector en España

Correlación rho de Spearman	España				
	Construcción	Hostelería	Servicios	Industria editorial	Telecomuni-caciones
Cifra de negocios	,554***	,255*	,207	,374**	,481*
N	91	69	56	72	26
Resultado después de impuestos	,343**	,028	,130	,341**	-,022
N	93	70	56	72	26
Activos totales	,518***	,317**	,439**	,275*	,612**
N	93	70	56	72	26
Activos fijos intangibles	,345**	,324**	,311*	,052	,486*
N	90	69	56	72	26

* Coeficientes de correlación significativos al nivel 0,05.

** Coeficientes de correlación significativos al nivel 0,01.

*** Coeficientes de correlación significativos al nivel 0,001.

En la Tabla 4.10 tenemos los resultados de los distintos sectores en el Reino Unido. Los coeficientes de correlación a nivel de país son más elevados para todas las variables, excepto para los activos intangibles, que cuando comparamos con los resultados para el Reino Unido en la Tabla 4.7, esto es, considerando todas las empresas del Reino Unido de forma conjunta con independencia de su actividad. Ello indica que grupos de empresas homogéneas son más propensas a obtener coeficientes de correlación significativos elevados.

Cuando comparamos con los sectores en la Tabla 4.7, es decir, cuando consideramos conjuntamente a las empresas españolas y británicas que pertenecen a un mismo sector, observamos que el análisis a nivel del Reino Unido es más significativo, en términos generales, para los sectores de la construcción y la hostelería, muy similar para el editorial y ligeramente inferior para el caso del sector servicios y telecomunicaciones. En todo caso, los resultados dependen de las variables específicas que tomemos en consideración, no existiendo ningún patrón bien definido. Como en el caso español, "activos totales" es la variable que está correlacionada de manera más significativa con el número de enlaces recibidos.

Tabla 4.10. Correlaciones entre el número de enlaces recibidos y los datos financieros por sector en Reino Unido

Correlación rho de Spearman	Reino Unido				
	Construcción	Hostelería	Servicios	Industria editorial	Telecomuni-caciones
Cifra de negocios	,396***	,277***	,318*	,413***	,341**
N	94	166	56	193	69
Resultado después de impuestos	,301**	,180*	,158	,240**	,190
N	103	180	60	201	71
Activos totales	,524***	,417***	,467***	,418***	,398**
N	103	180	60	201	71
Activos fijos intangibles	,280**	,113	,121	,184*	,179
N	91	176	58	193	65

* Coeficientes de correlación significativos al nivel 0,05.

** Coeficientes de correlación significativos al nivel 0,01.

*** Coeficientes de correlación significativos al nivel 0,001.

La Tabla 4.11 indica qué país presenta el coeficiente de correlación significativo más alto para cada industria y variables, tal y como se muestran en las Tablas 4.9 y 4.10. Es importante notar como, en general, los sectores de Hostelería, Servicios y Editorial presentan correlaciones más altas en el Reino Unido, mientras que los sectores de Construcción y Telecomunicaciones lo hacen en España. También, la correlación entre el número de enlaces recibidos y la variable activos totales es más elevada en general para el caso del Reino Unido.

Tabla 4.11. País con el mayor coeficiente de correlación

Comparación entre España (ES) y Reino Unido (UK)	Construcción	Hostelería	Servicios	Industria editorial	Telecomuni-caciones
Cifra de negocios	ES	UK	UK	UK	ES
Resultado después de impuestos	ES	UK	-	ES	-
Activos totales	UK	UK	UK	UK	ES
Activos fijos intangibles	ES	ES	ES	UK	ES

4.1.2.3. Conclusiones

Esta es la primera investigación webmétrica que se realiza para un conjunto de diversos sectores, en dos países europeos e incluyendo empresas no cotizadas. Estas características parecen complicar mucho el análisis de los datos, dado que no se pueden observar patrones claros en los resultados.

En primer lugar, es importante subrayar que cuando distintos sectores en un mismo país se toman de manera agregada, mezclando empresas heterogéneas en cuanto a actividad, las correlaciones siguen siendo significativas, aunque los coeficientes son inferiores a cuando se consideran por separado. Esto puede deberse, entre otras cuestiones, al elevado número de casos incluidos en cada grupo, que llega hasta los 317 para España y a 615 para Reino Unido. Ello hace que el punto crítico

a partir del cual se considera un coeficiente de correlación significativo se reduzca considerablemente. Así, si bien la correlación puede ser estadísticamente significativa, puede resultar tan débil que no sea de gran utilidad. En todo caso, los datos indican que es conveniente realizar estudios más amplios en economías concretas con el fin de explorar el alcance de las correlaciones cuando se tienen en consideración empresas de distintos sectores de actividad.

En segundo lugar, es relevante hacer notar que, cuando el análisis por sectores realizado en la Tabla 4.7 se desagrega por países, los resultados no mejoran de manera significativa. En algunos sectores y para algunas variables los coeficientes de correlación son incluso menores o menos significativos. Se ha encontrado evidencia de que los niveles de correlación son similares en países distintos, por ejemplo, es el caso de Estados Unidos y China (Vaughan, 2004b) y de Estados Unidos y Canadá (2004a). Esto está aún más justificado en el caso presente en el que se trata de dos países que comparten mercado único como miembros de la Unión Europea. En base a ello, es posible que para determinados análisis fuera más útil tener en cuenta regiones que países como tales.

Si comparamos los resultados de este estudio con los resultados del estudio en la Sección 4.1.1, relativo a diversos sectores en los Estados Unidos, encontramos que el comportamiento de los sectores en ambos casos es muy poco similar. El sector de la construcción en la economía americana no presenta correlaciones para ninguna de las variables consideradas, mientras que en este estudio es uno de los sectores con mayores coeficientes de correlación, especialmente en el caso de las empresas españolas. De hecho, esto viene a reflejar la importancia que este sector tiene en España, siendo uno de los principales motores de la economía durante muchos años y causa importante de la grave crisis económica y de empleo que atraviesa el país. También el sector servicios tiene un nivel de correlación mucho más bajo en España y Reino Unido que en los Estados Unidos. Por ejemplo, los coeficientes de correlación entre enlaces recibidos y activos totales son 0,439 en España, 0,467 en Reino Unido y 0,794 en los Estados Unidos. Es conveniente recordar, en cualquier caso, que las empresas estadounidenses en el estudio son mayores que las europeas.

Este trabajo presenta algunas limitaciones específicas que se derivan de la naturaleza de la información disponible en la Web. No es posible obtener información de enlaces correspondiente a un tiempo pasado, por lo que el trabajo emplea datos de 2009 mientras que los datos financieros más actualizados en ese momento corresponden al ejercicio 2007. Estudios de tipo longitudinal podrían resolver este problema en los próximos años. Es importante señalar también que los sectores industriales seleccionados podrían segmentarse en grupos más homogéneos y específicos. Por ejemplo, el código NAICS 511, *Publishing Industries*, incluye editoriales de libros, prensa y fabricantes de *software*. Sin embargo, haber realizado ésto con la información disponible hubiera obligado a reducir mucho el tamaño de la muestra con la que trabajamos.

El presente trabajo exploratorio constituye el primer trabajo webmétrico sobre empresas realizado en el ámbito europeo. Si bien se ha encontrado evidencia que apoya la existencia de relaciones entre las variables de enlaces recibidos y las económico-financieras, algunas de las conclusiones contradicen otros resultados puestos de relieve anteriormente, abriendo nuevas oportunidades de investigación con el fin de clarificar estas cuestiones.

4.1.3. Análisis de regresión multivariante del número de enlaces recibidos por los sitios web de empresas en España y Reino Unido

El trabajo incluido en este apartado va más allá de los análisis de correlación simple que, con fines exploratorios, se han llevado a cabo anteriormente. Una vez identificadas correlaciones positivas significativas entre las variables, pretendemos elaborar un modelo explicativo de la variable enlaces recibidos mediante un análisis de regresión multivariante. No hemos encontrado hasta el momento ningún trabajo publicado en este sentido. Vaughan y Thelwall (2005) llevaron a cabo un análisis similar para estudiar los patrones de enlaces que reciben los sitios web de universidades canadienses. En dicho trabajo emplean como variables independientes la calidad de los estudiantes y de los centros en estudio. También estudian el factor lingüístico representado por las universidades que enseñan en inglés y las que lo hacen en francés.

4.1.3.1. Método

Partiendo de los datos del estudio anterior, correspondientes a empresas en diversos sectores de actividad en España y Reino Unido, en esta sección vamos a proceder a realizar un análisis econométrico de los mismos que permita establecer algunas de las variables financieras que explican el número de enlaces que reciben las página web de las empresas.

Para ello se va a emplear como técnica estadística el análisis de regresión multivariante, que nos permitirá estudiar si existen o no relaciones lineales entre las variables que integran el modelo teórico propuesto. El análisis de regresión multivariante es una técnica estadística comúnmente utilizada en los trabajos de investigación empresarial, puesto que permite verificar un conjunto de relaciones entre variables de modo que, cumpliendo determinados supuestos, las conclusiones puedan ser generalizables. Permite analizar la relación entre una única variable dependiente (variable explicada) y un conjunto de variables independientes (variables explicativas). Los coeficientes de regresión del modelo

se corresponden con las ponderaciones de cada variable explicativa, que indican la contribución relativa a la predicción global de la variable explicada.

El modelo planteado cuenta con las siguientes variables explicativas de carácter financiero: activos totales, activos fijos intangibles, cifra de negocios y resultado después de impuestos. Las dos primeras son variables de posición financiera mientras que las dos últimas de desempeño o rendimiento financiero. Han sido elegidas por considerarse entre los indicadores más representativos de la envergadura y rendimiento de una empresa. La variable de activos intangibles se ha seleccionado por su posible vinculación con la presencia de la empresa en la Web, lo cual constituye un intangible, en ocasiones muy valioso, para la empresa. Al margen se ha incluido una variable ficticia dicotómica que permite distinguir entre Reino Unido (la variable toma valor 0) y España (toma valor 1). Esta variable puede responder tanto a criterios geográficos, como a criterios lingüísticos, ya que todas las empresas del Reino Unido tienen su página en inglés y todas las empresas españolas en español, salvo alguna excepción en la que presentan una versión inglesa alternativa. Para la descripción de los datos empleados en este apartado remitimos al trabajo anterior (Apartado 4.1.2.1).

Con el objeto de valorar el grado en qué son generalizables las conclusiones obtenidas, la Tabla 4.12 contiene los supuestos en los que se basa el análisis de regresión multivariante, los cuales examinaremos una vez elaborado el modelo.

Tabla 4.12. Supuestos para el análisis de regresión multivariante

Supuesto	Significado	Pruebas de diagnóstico	Rasgos
Linealidad	Relación lineal entre cada variable independiente y la variable dependiente.	Gráficos de regresión parcial.	No deben existir pautas curvilíneas en los gráficos.

Homocedas-ticidad	Varianza constante del término de error.	Diagrama de residuos estandarizados observados y esperados.	No deben existir ningún tipo de patrón de comportamiento en los gráficos.
Normalidad	Las distribución de los residuos del modelo deben seguir una distribución normal.	Histograma de residuos. Gráfico p-p o gráfico de normalidad de los residuos estandarizados de la regresión.	La distribución de los residuos del modelo debe aproximarse a una distribución normal. La distribución de los puntos debe de coincidir con la diagonal trazada en el diagrama.
Multicolinea-lidad	Grado en que cada variable independiente se explica por otras variables independientes.	Valor de Tolerancia.	Cercano a 1. Umbral fijado en 0,1.
		Factor de inflación de la varianza (FIV). [Inverso del valor de tolerancia.]	Cercano a 1. Umbral fijado en 10.

Para llevar a cabo el análisis econométrico, empleamos el paquete estadístico SPSS v.17. Dado que existen diversos procedimientos para llevar a cabo la regresión, se ha empleado el método *Stepwise*, que selecciona de manera automática qué variables independientes son significativas a la hora de explicar la variable dependiente. El método *Stepwise* emplea un procedimiento matemático que tiene en cuenta las correlaciones parciales entre variables al objeto de incluir aquellas con mayor poder explicativo en primer lugar. Se trata de un procedimiento especialmente indicado en aquellos casos en los que no existe literatura previa o un

modelo bien definido, por lo que es preciso seleccionar entre un conjunto de variables independientes que teóricamente podrían explicar la variable dependiente. Cada una de las variables se incluye progresivamente de modo que se puede observar como el poder explicativo del modelo cambia en cada paso.

Con el fin de confirmar los resultados obtenidos, se ha llevado también a cabo la regresión múltiple mediante el procedimiento *Enter*, que introduce las cinco variables seleccionadas simultáneamente en la ecuación del modelo.

El modelo teórico que se va a comprobar se incluye en la Tabla 4.13.

Tabla 4.13. Modelo teórico explicativo del número de enlaces recibido

Ecuación genérica	$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$
Variable explicada	Y_1: Enlaces
Variables explicativas	X_1: Activos totales X_2: Activos fijos intangibles X_3: Cifra de negocios X_4: Resultado después de impuestos X_5: País Reino Unido (0), España (1)

La hipótesis que subyace al modelo es que el número de enlaces que recibe el sitio web de una empresa puede explicarse a partir de variables de posición financiera (activos totales y activos fijos intangibles), variables de desempeño financiero (cifra de negocios y resultado neto) y el país al que pertenece la empresa. De manera más específica, las hipótesis del comportamiento de cada una de las variables independientes son:

- Hipótesis 1: Los activos totales de una empresa se relacionan de manera

positiva y significativa con el número de enlaces que recibe el sitio web de la empresa.

- Hipótesis 2: Los activos fijos intangibles de una empresa se relacionan de manera positiva y significativa con el número de enlaces que recibe el sitio web de la empresa.

- Hipótesis 3: La cifra de negocios de una empresa se relaciona de manera positiva y significativa con el número de enlaces que recibe el sitio web de la empresa.

- Hipótesis 4: El resultado neto de una empresa se relaciona de manera positiva y significativa con el número de enlaces que recibe el sitio web de la empresa.

- Hipótesis 5: El país al que pertenece la empresa se relaciona de manera significativa con el número de enlaces que recibe el sitio web de la empresa.

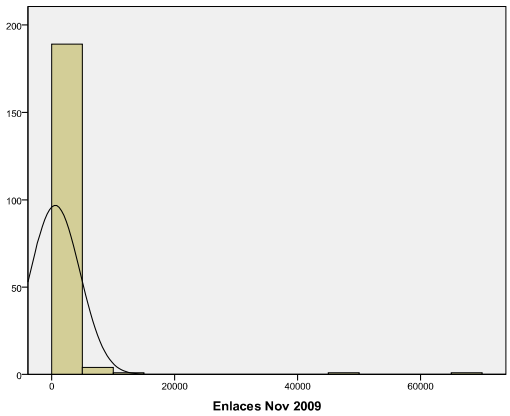
Dado que los datos de empresas corresponden a cinco sectores empresariales distintos, se va a examinar el modelo anterior en cada uno de ellos. Un primer cálculo del modelo nos permite observar que la distribución de los residuos de los modelos se aleja considerablemente del supuesto de normalidad. Lo mismo ocurre también con las variables en su conjunto y especialmente la variable explicada. El número de enlaces recibido presenta una distribución muy asimétrica, como puede observarse en la Tabla 4.14 y la Figura 4.1. Es por ello que decidimos aplicar una transformación logarítmica (logaritmo natural) a todas las variables, excepto a la variable dicotómica. En determinados casos, se ha añadido una constante a la variable original con el fin de evitar valores negativos o iguales a cero.

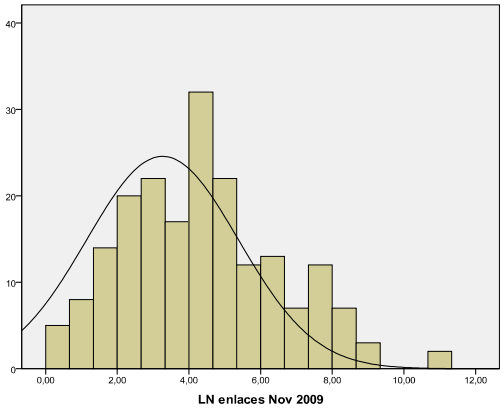
Tabla 4.14. Descriptivos de la variable dependiente "Enlaces" sin transformar y tras aplicar la transformación logarítmica

	Construcción		Hostelería y restauración		Servicios		Industria editorial		Telecomunica-ciones	
	No tr.	Tr.	No tr.	Tr.	No tr.	Tr.	No tr.	Tr.	No tr.	Tr.
N: Validos (Perdidos)	196 [ES =93; UK =103]		249 [ES =70; UK =179]		113 [ES =56; UK =57]		273 [ES =72; UK =201]		97 [ES =26; UK =71]	
Media	1.104,75	4,33	7.074,55	6,04	7.141,65	5,93	275.256,92	8,54	55.487,46	6,25
Mediana	62,00	4,14	414,00	6,03	336,00	5,82	3.990,00	8,29	235,00	5,46
Desviación estándar	5.779,34	2,21	58.125,68	2,19	29.543,36	2,32	918.548,21	3,29	338.445,60	2,81
Asimetría	9,52	,38	14,44	,053	6,56	,55	5,49	,14	8,74	,86
Error estándar de asimetría	,174	,174	,154	,154	,23	,23	,147	,147	,245	,245
Curtosis	96,79	-,075	218,35	,36	48,23	-,065	40,03	-,54	80,88	,50
Error estándar de curtosis	,346	,346	,307	,307	,451	,451	,294	,294	,485	,485

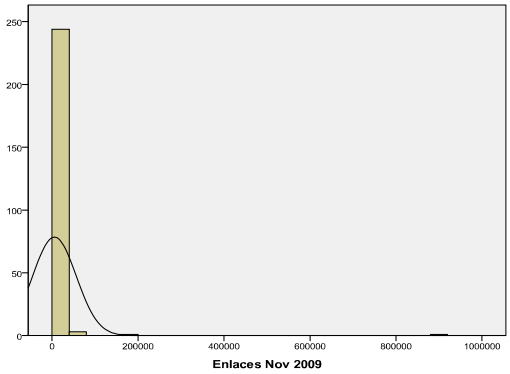
Figura 4.1. Distribución de frecuencias de la variable dependiente "Enlaces" sin transformar y tras aplicar la transformación logarítmica

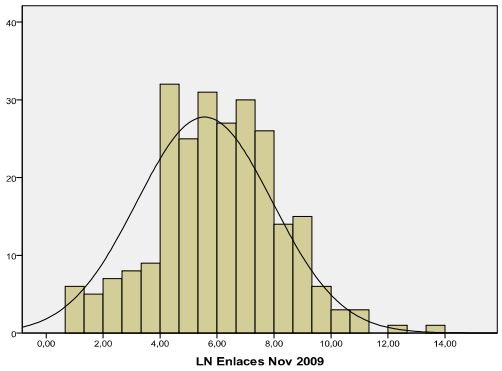
Sector: *Construcción*



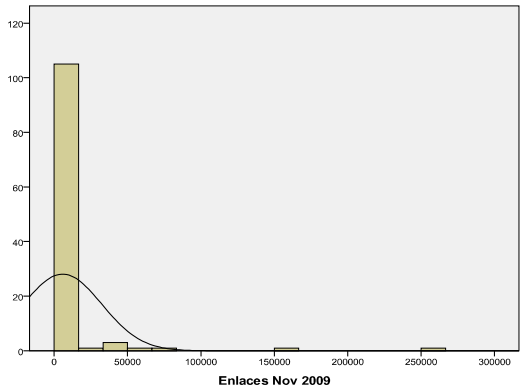


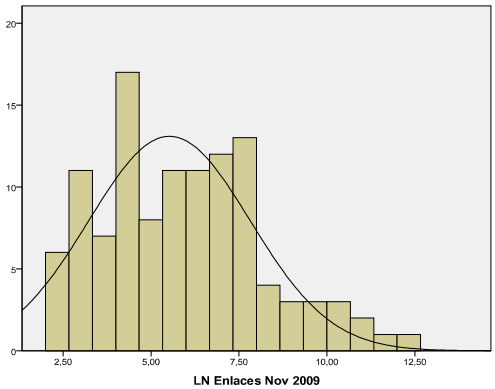
Sector: *Hostelería y restauración*



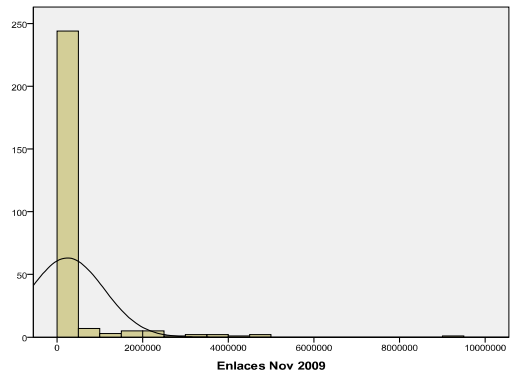


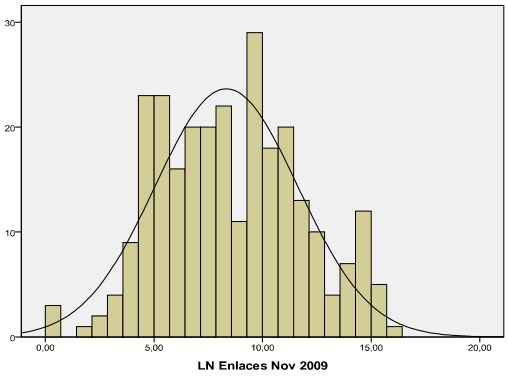
Sector: *Servicios*



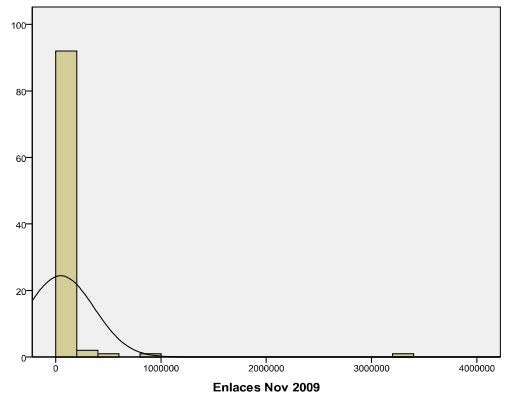


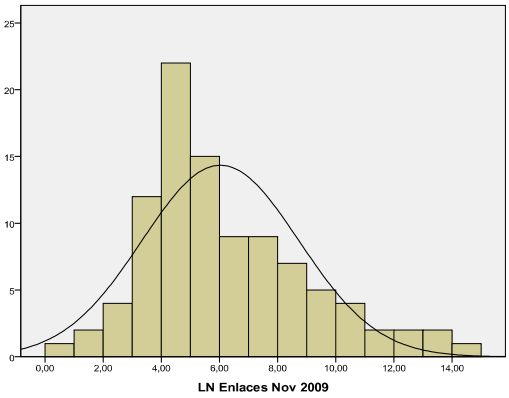
Sector: *Industria editorial (excepto Internet)*





Sector: *Telecomunicaciones*





Tras llevar a cabo la transformación de las variables, se vuelve a calcular el modelo de regresión, obteniendo una distribución de los residuos que se aproxima satisfactoriamente a la normalidad. Antes de proceder al cálculo del modelo definitivo, se verifican los casos cuyo valor observado difiere más de dos desviaciones típicas del valor esperado según el modelo. Tras un análisis detallado de los casos, se eliminan algunos valores que presentan algunos problemas (debidos a la imposibilidad de acceder a la página web, a la redirección del dominio, a la existencia de dominios web alternativos, etc.) que no fueron detectados en el momento inicial de la recogida de datos o que se han generado con posterioridad y que podrían distorsionar los resultados. En concreto, los valores descartados

corresponden a: un caso al sector de Telecomunicaciones, tres a la Construcción, cuatro a Hostelería y restauración, cuatro a Servicios y ninguno a industria Editorial.

4.1.3.2. Resultados

La Tabla 4.15 muestra los resultados obtenidos aplicando el método *Stepwise*. Se procedió también al cálculo del modelo mediante el método *Enter*, confirmando los resultados obtenidos. Para cada uno de esos modelos correspondientes a los distintos sectores se incluye la F y el R^2. Por su parte, para cada una de las variables se facilitan cuatro valores; por orden de primera a última corresponden a: coeficiente de regresión, error típicos (entre paréntesis), coeficiente de regresión estandarizado y valor de la t.

Tabla 4.15. Modelo de regresión multivariante para los cinco sectores empresariales

	Constante	LN Activos	LN Intan.	LN Rdo. después impuestos	LN Cifra negocios	País	F	R^2
Construcción	-4,494***	,582***			,274**	-1,117***	36,636***	,398
	(,974)	(,101)			(,098)	(,280)		
		,446***			,215**	-,256***		
	-4,613	5,790			2,790	-3,988		
Hostelería y restauración	-1,353	,696***					73,870***	,248
	(,884)	(,081)						
		,498***						
	-1,531	8,595						

Servicios	-,874	,554***				44,311***	,305
	(,998)	(,083)					
		,552***					
	-,875	6,657					
Industria editorial	3,763*			1,132***	1,344**	33,561***	,209
	(1,622)			(,151)	(,402)		
				,417***	,186**		
	-2,320			7,478	3,342		
Telecomu-nicaciones	17,126	,794***	-1,336*		-1,831***	19,187***	,401
	(8,771)	(,110)	(,612)		(,502)		
		,632***	-,184*		-,320***		
	1,953	7,231	-2,184		-3,650		

Variable dependiente: LN enlaces

Las casillas muestran: coeficiente de regresión, error típico (entre paréntesis), coeficiente de regresión estandarizado y t.

* Variable significativa al 5%.

** Variable significativa al 1%.

*** Variable significativa al 0,1%.

Los modelos resultantes pueden verse en la Tabla 4.16. La información en ella contenida proviene de la Tabla 4.15, donde puede ampliarse.

Tabla 4.16. Modelos explicativos del número de enlaces recibidos por los sitios web de las empresas para los cinco sectores empresariales

Sector	Modelo	R^2
Construcción	-4,494 + 0,582 LN Activos + 0,274 LN Cifra de negocios -1,117 País	,398
Hostelería y restauración	-1,353 + 0,696 LN Activos	,248
Servicios	-,874 + 0,554 LN Activos	,305
Industria editorial	3,763 + 1,132 LN Cifra negocios + 1,344 País	,209
Telecomuni-caciones	17,126 + 0,794 LN Activos -1,336 LN Resultado después impuestos -1,831 País	,401

A la luz de los modelos anteriores podemos observar varias cuestiones. En primer lugar, la variable "LN Activos" es la que mejor explica de forma global el número de enlaces que reciben los sitios web de las empresas. El único sector en el que no aparece esta variable en el modelo es el Editorial, ya que no es significativa al 5%, aunque sí lo es al 10,4%. Si bien el nivel de significación es superior a la barrera del 5%, se encuentra relativamente próximo por lo que creemos que se trata de una variable que, de modo consistente, debería aparecer en modelos explicativos de enlaces. Si atendemos a los coeficientes de regresión estandarizados (Tabla 4.15) podemos observar como "LN Activos" es la variable de mayor importancia en todos los casos.

La variable "LN Intangibles" no es significativa en ningún caso. Existen varias razones que pueden explicar este extremo. La primera es de carácter fundamentalmente práctico. Muchas empresas indican que sus activos intangibles son cero, habiendo otras muchas que no incluyen información de este tipo, por lo que la casilla aparece como un dato perdido. No estamos seguros si a la hora de divulgar la información financiera en ambos países el que no se incluya el dato de

intangibles significa que no se dispone de este dato o de que por contra se trata de cero. Al margen, la contabilización de los activos intangibles es un área que presenta muchas dificultades debido a los problemas de identificación, reconocimiento y valoración de los mismos. Su naturaleza inmaterial complica, por ejemplo, la satisfacción del criterio de control o su valoración fiable. Así, una empresa para la cual su presencia y visibilidad en la Web constituye un recurso muy valioso, no tendrá una partida en su balance que refleje dicho activo.

En el sector de la Construcción y en el Editorial, la variable "LN Cifra de negocios" aparece en el modelo; mientras que la variable "LN Resultado después de impuestos" sólo aparece en el sector de Telecomunicaciones con un coeficiente estandarizado relativamente bajo, 0,184, y un nivel de significación también reducido.

Así en términos generales, podemos afirmar que las variables financieras que mejor explican el número de enlaces recibidos son "LN Activos" (variable de posición financiera) y "LN Cifra de negocios" (variable de desempeño financiero). Ello nos permitiría aceptar las hipótesis 1 y 3 planteadas anteriormente, si bien habría que atender específicamente a cada uno de los sectores de actividad. Las hipótesis 2 y 4, relativas a "LN Intangibles" y a "LN Resultado después de impuestos" se rechazan. En relación con esta última variable, la relación existente, de acuerdo con el modelo incluido para las Telecomunicaciones, es negativa en vez de positiva como se planteaba. Esto indicaría que un mayor resultado implica un menor número de enlaces, lo cual en principio no tendría sentido. Debemos señalar en este caso que dado que el número de enlaces tiende a ser mayor con el tiempo, por un proceso de acumulación, sería más lógico pensar que variables como los activos totales o la cifra de negocios permiten explicar mejor esta magnitud. El resultado de un ejercicio podría ser negativo un ejercicio a pesar de que la cifra de negocios se hubiera seguido incrementando, por lo que se trata de una magnitud más sujeta a variaciones, pudiendo ser positiva o negativa en años alternos dentro de toda lógica.

Por último, la hipótesis 5 hacía referencia a que la variable país fuera significativa,

algo que ocurre en tres de los cinco sectores. En el caso del sector de la Construcción y de Telecomunicaciones, su valor es negativo, indicando que el hecho de que la empresa sea española supone una disminución en el término constante frente al hecho de que sea británica, mientras que en el sector Editorial ocurre lo contrario.

De cara a la interpretación de los datos, al haber realizado una transformación logarítmica tanto de la variable dependiente como de las independientes, exceptuando la variable ficticia, los coeficientes han de interpretarse como elasticidades. En el contexto de la regresión, interpretaremos la elasticidad como el porcentaje de cambio en la variable dependiente cuando la variable independiente se incremente en un 1%. Así, podemos decir, en relación con el modelo del sector de la Construcción que un incremento del 1% de los activos provoca un incremento del 0,582% en el número de enlaces recibidos.

Finalmente se ha comprobado el cumplimiento de los supuestos indicados en la Tabla 4.12 anterior. Para ello hemos empleado una serie de gráficos y diagramas que pueden observarse en el Anexo 2, organizados en función del sector empresarial al que se refiere el modelo. Por un lado, el examen del histograma de residuos estandarizados y del P-P Plot de los residuos estandarizados de la regresión, nos muestra en los cinco casos que el supuesto de normalidad se cumple de manera aceptable. El examen de los diagramas de residuos estandarizados observados y esperados nos permite afirmar que no se aprecia claramente ningún problema de heterocedasticidad. Por su parte, a través de los diagramas de regresión parcial de las variables del modelo podemos indicar que, en términos generales, la variable dependiente y las independientes mantienen una relación de linealidad. La multicolinealidad, como puede observarse en la Tabla 4.17 no representa un problema para ninguno de los modelos en los que se incluyen más de una variable.

Tabla 4.17. Indicadores de multicolinealidad para los cinco sectores empresariales.

Sector	Variables	Tolerancia	VIF
Construcción	LN Activos	,612	1,635
	País	,878	1,139
	LN Cifra de negocios	,613	1,632
Hostelería y restauración	LN Activos	-	-
Servicios	LN Activos	-	-
Industria editorial (excepto Internet)	LN Cifra de negocios	1,000	1,000
	País	1,000	1,000
Telecomunicaciones	LN Activos	,913	1,095
	País	,904	1,107
	LN Resultado después de impuestos	,987	1,013

4.1.3.3. Conclusiones

En el apartado anterior hemos examinado con más detalle los resultados de aplicar el modelo de regresión propuesto a cada uno de los cinco sectores estudiados. Existen determinadas variables, como el total de activos y la cifra de negocios, que en la mayoría de casos permiten explicar en buena medida el número de enlaces recibidos por los sitios web de las empresas. Futuros trabajos deben explorar cómo se pueden utilizar modelos de este tipo para detectar, por ejemplo, empresas cuya presencia en la Web es inferior a la esperada o empresas cuyo impacto en la red es superior al que correspondería por su dimensión económica. Esto puede constituir una herramienta para las empresas de cara a gestionar sus activos intangibles vinculados a Internet.

Sería conveniente explorar otras variables, distintas a las financieras, que permitieran incrementar el R^2, enriqueciendo de ese modo los modelos. En este

trabajo se ha utilizado una variable *dummy* para analizar el impacto de que las empresas fueran españolas o británicas, encontrando que, en tres de los cinco sectores, la procedencia constituye un factor significativo.

Es preciso también recoger datos de enlaces de forma periódica, de modo que se pueda seguir una evolución en el comportamiento de los modelos cuando la medición de ambos tipos de variables se aproxima en el tiempo. Debido a la imposibilidad de obtener datos de enlaces de momentos pasados, en el trabajo se combinan datos correspondientes a 2009 con datos financieros del ejercicio 2007, último año para el que la mayoría de empresas contaba con información en la base de datos empleada.

4.2. Análisis de co-enlaces

La investigación centrada en el análisis de co-enlaces (Apartado 3.4.5) pone de relieve las relaciones entre elementos, utilizando el punto de vista de aquellas páginas web que enlazan simultáneamente a los sitios web de las organizaciones en estudio. Hasta ahora los trabajos realizados en el ámbito de empresas se han circunscrito a la situación competitiva en sectores individuales (Vaughan y You, 2006) y en subsectores mediante el empleo de términos clave para filtrar la búsqueda (Vaughan y You, 2008; 2009). Los trabajos incluidos en este apartado se centran en el estudio de conjuntos de empresas heterogéneos en cuanto a su sector de actividad (Apartado 4.2.1) y en el empleo de términos clave para el estudio de situaciones económicas específicas (Apartado 4.2.2). En este sentido ambos estudios exploran nuevas aplicaciones del análisis de co-enlaces.

El primer trabajo pretende dar respuesta a la tercera pregunta de investigación de de la tesis (Apartado 1.2.3): *¿Cómo se interrelacionan empresas pertenecientes a sectores de actividad distintos analizadas a través de su presencia en la Web?* Los co-enlaces proporcionan información sobre la similitud o grado de relación entre dos elementos. En el contexto empresarial, esta similitud, considerada al nivel de un mismo sector, se interpreta como una medida del grado de competencia entre dos empresas. Empresas similares en un mismo sector son aquellas que compiten de algún modo por un mismo mercado. El trabajo en el Apartado 4.2.1 analiza las relaciones entre las empresas de algunos de los principales índices bursátiles del mundo en los que coexisten actividades de muy distinta índole. Dado que los co-enlaces proporcionan información sobre el grado de similitud entre empresas, nuestra expectativa es ver cómo se forman *clusters* basados en sectores de actividad y observar cómo los sectores interaccionan entre sí.

El segundo trabajo aborda la cuarta pregunta de investigación de la tesis (Apartado 1.2.4): *¿Podemos emplear la información sobre enlaces para estudiar eventos económicos determinados?* Para contestar a esta pregunta, hemos llevado a cabo un análisis de co-enlaces, filtrado mediante el empleo de palabras clave, para

analizar la evolución de la crisis financiera de los bancos en los Estados Unidos.

4.2.1. Análisis de co-enlaces de empresas heterogéneas pertenecientes a algunos de los principales índices bursátiles del mundo

Este trabajo amplía la aplicación del análisis de co-enlaces a sitios web de empresas que pertenecen a algunos de los principales índices bursátiles del mundo. Son empresas que están entre las mayores en sus respectivas economías y que pertenecen a una amplia variedad de sectores de actividad. Dada su relevancia económica, cuentan en general con una presencia destacada en la Web y están expuestas a la atención de los usuarios, otras empresas y otros agentes económicos. Hasta ahora, se ha aplicado el análisis de co-enlaces para investigar las relaciones competitivas entre empresas de un mismo sector; sin embargo, consideramos que al combinar empresas de distintos sectores es posible detectar otros tipos de relaciones, tales como alianzas u otros acuerdos de cooperación o el grado de vinculación entre sectores y su naturaleza. El alcance de este estudio nos permite analizar el comportamiento de las empresas cuando se consideran junto a otras entidades de naturaleza diversa. El objetivo es fundamentalmente explorar estas relaciones y abrir nuevas posibilidades de investigación en este sentido.

El número de co-enlaces constituye una medida de la similitud que se ha empleado anteriormente para evaluar las posiciones competitivas de empresas dentro de un sector específico (Vaughan y You, 2006; 2008; 2009). El trabajo actual pretende ampliar los anteriores para avanzar en el análisis de empresas que son heterogéneas en cuanto a la actividad a la que se dedican. Partiendo de los supuestos anteriores, nuestra expectativa inicial es la de identificar *clusters* de empresas pertenecientes a un mismo sector de actividad.

4.2.1.1. Método

Para el presente trabajo se han tenido en cuenta las empresas que componen los siguientes índices bursátiles: Dow Jones Industrial (Estados Unidos), Dow Jones Euro Stoxx 50 (empresas de la Eurozona), CAC 40 (París), FTSE 100 (Londres) e Ibex 35 (Madrid). Una selección de las empresas más señaladas de cara al análisis realizado en el artículo (Figuras 4.2 a 4.8) se incluye en el Anexo 3. La Tabla 4.18 incluye información descriptiva sobre los diferentes índices incluidos en el estudio: país, número de empresas, fecha de la composición del índice qué aparece en el trabajo y la URL de la empresa. Las páginas web incluidas en la tabla permiten obtener información sobre la composición del índice, aunque es posible que ésta no coincida con la composición tenida en cuenta en el estudio debido a que se modifica a lo largo del tiempo. Al objeto de facilitar la identificación de las empresas que están omitidas en los anexos, las etiquetas aplicadas a las empresas en los mapas son las mismas etiquetas empleadas en sus cotizaciones bursátiles.

Tabla 4.18. Información sobre los índices bursátiles en el estudio

Índice	País / Región	Número de empresas	Fecha de la composición	URL
Dow Jones Industrial	Estados Unidos	30	17/02/2009	http://www.djaverages.com/
FTSE 100	Reino Unido	100	24/02/2009	http://uk.finance.yahoo.com/q/cp?s=^FTSE
Euro Stoxx 50	Países de Eurozona	50	17/02/2009	http://www.stoxx.com/indices/components.html?symbol=SX5E
CAC 40	Francia	40	19/02/2009	http://www.euronext.com/trader/indicescomposition/composition-4411-EN-FR0003500008.html?selectedMep=1
IBEX 35	España	35	19/02/2009	http://www.bolsamadrid.es/ing/mercados/acciones/accind1_1.htm

El sitio web de cada una de estas empresas fue obtenido de la página web del índice bursátil y posteriormente verificado para comprobar su corrección. La gran mayoría de empresas dispone únicamente de un único sitio web. En aquellos casos en que existen varios dominios vinculados a una misma empresa intentamos incluir las distintas URLs en el estudio. Sin embargo, Yahoo!, el buscador empleado en la investigación, no permite combinar operadores para realizar la consulta requerida, por lo que finalmente se optó por seleccionar la URL con mayor número de enlaces recibidos.

Es necesario señalar que la clasificación sectorial en cada bolsa de valores es distinta. Así, por ejemplo, el *Industry Classification Benchmark* (ICB) está muy extendido (por ejemplo, lo utilizan Dow Jones, FTSE y otros mercados en distintas partes del mundo), aunque no se emplea de manera universal. Por ejemplo, el mercado español, representado a través del IBEX 35, emplea su propia clasificación sectorial. Esta heterogeneidad, junto con el escaso número de empresas en determinados sectores dentro de algunos de los índices considerados, obliga a agregar empresas dentro de grupos más generales tales como tecnologías de la información (*IT*), medios de comunicación (*Media*) o Finanzas (*Financial*). Se trata de una agregación de carácter práctico para hacer viable el análisis, que, en todo caso, se encuentra abierta a discusión y revisión. Algunas notas sobre la descripción de la clasificación sectorial original de los distintos índices se incluyen en el Anexo 3.

Para la obtención de información de co-enlaces se utiliza Yahoo!, que es el único gran motor de búsqueda que permite combinar operadores para formular una consulta apropiada para nuestro trabajo. Tanto Google (Google, 2006; 2009) como MSN Live Search (Live Search, 2007) presentan limitaciones en el momento de la recogida de los datos, a principios de 2009. Estas limitaciones han sido abordadas previamente en la parte metodológica de las anteriores investigaciones expuestas.

Dado que los buscadores comerciales cuentan con versiones en distintos países es posible que las páginas de un mismo país sean tratadas de forma más extensa y

pormenorizada que las de otros, proporcionando resultados sesgados (Vaughan y Thelwall, 2004). Esto ocurre por ejemplo con China, país para el que Yahoo! opera con una base de datos propia. En nuestro caso, se hicieron las oportunas comprobaciones de resultados empleando Yahoo! España y Yahoo! France; sin embargo, los resultados obtenidos fueron los mismos, por lo que concluimos que pese a emplear distintas interfaces para cada país, la base de datos de búsquedas es común junto con la de Yahoo! Global, por lo que se emplea esta versión (www.yahoo.com) para efectuar las consultas.

De los dos operadores con los que cuenta Yahoo!, *linkdomain* and *link*, explicados en diversas secciones anteriores, empleamos *linkdomain* ya que teóricamente nos interesa conocer todos los enlaces que apuntan al dominio de la empresa y no únicamente a su página de inicio. La sintaxis de las consultas relizadas aparece en la Tabla 4.19, usando como referencia las siguientes URLs: www.abc.com y www.xyz.com. La parte www se elimina al objeto de incluir los subdominios en la búsqueda. Por otra parte, el componente *-site:abc.com* permite filtrar los enlaces internos que provienen del propio dominio en estudio.

Tabla 4.19. Términos de búsqueda empleados en Yahoo!

Tipos de enlaces	Términos de búsqueda
Enlaces (inlinks) que apuntan a www.abc.com	linkdomain:abc.com –site:abc.com
Co-enlaces entre www.abc.com y www.xyz.com	(linkdomain:abc.com –site:abc.com) (linkdomain:xyz.com –site:xyz.com)

En tanto que los co-enlaces constituyen páginas que enlazan a dos sitios web incluidos en el estudio, los datos se recogen en una matriz simétrica, en la que el dato de cada cruce representa el número de co-enlaces entre el sitio web *x* y el sitio web *y*. Los datos de co-enlaces fueron obtenidos para cada índice bursátil; esto es, las empresas en un mismo índice están en la misma matriz de co-enlaces. Por lo

tanto, hay cinco matrices de co-enlaces para cada una de las cinco bolsas de valores. Cada una de las ellas ha sido analizada mediante escalamiento multidimensional (*multidimensional scaling,* MDS en adelante) con el fin de generar unos mapas que visualicen las relaciones entre las distintas empresas medidas a través del número de co-enlaces. El escalamiento multidimensional se basa en un método heurístico, situando las empresas con un mayor número de co-enlaces más próximas en el mapa resultante. En tanto que las empresas similares o vinculadas entre sí son más proclives a recibir co-enlaces (por ejemplo, dos empresas con poca relación, por ejemplo, dedicadas al sector de la alimentación y a las telecomunicaciones tienen una menor probabilidad de ser enlazadas simultáneamente o co-enlazadas), el número de co-enlaces entre dos empresas constituye una medida de su grado de relación o proximidad. De este modo, podemos concluir que las empresas que estén vinculadas entre sí estarán situadas más cerca la una de la otra que aquellas que no. La situación de las empresas en el mapa solamente muestran sus posiciones relativas entre sí. Esto quiere decir que si modificamos la lista de empresas en el análisis, las posiciones de las empresas también se varían modificadas. Por lo tanto, es importante recalcar que no existe un significado absoluto en la posición de las empresas ni en las coordenadas de las empresas en el mapa, sino que depende de la posición de las demás empresas en el mapa. Se trata pues de un posicionamiento relativo.

El objetivo que se pretende en el trabajo es emplear los mapas MDS para generar grupos de empresas y poner de relieve sus posiciones de mercado. También esperamos evaluar la efectividad de nuestros métodos comparando los resultados de los distintos mapas entre sí, al objeto de determinar si existen patrones comunes entre los distintos países o regiones contempladas en los índices.

Al objeto de procesar la matriz de co-enlaces obtenida, ésta se ha normalizado para obtener una medida relativa del grado de relación existente entre las distintas empresas. Esta normalización se considera necesaria en este caso ya que un número de co-enlaces de 10 entre dos empresas se podría considerar muy grande si el número de enlaces que recibe el sitio de cada empresa considerado individualmente es pequeño, por ejemplo, 15. Sin embargo, si el sitio de cada

empresa recibiera miles de enlaces, entonces 10 co-enlaces representarían únicamente una pequeña proporción. La normalización se lleva a cabo aplicando el índice Jaccard que se calcula de la siguiente manera:

Número de co-enlaces normalizados $=n(A\cap B)/n(A\cup B)$, donde:

- A es el conjunto de páginas web que enlazan al sitio web X,
- B es el conjunto de páginas web que enlazan al sitio web Y,
- $n(A\cap B)$ es el número de páginas que enlazan simultáneamente a los sitios web X e Y, y
- $n(A\cup B)$ es el número de páginas que enlazan al sitio web X o al sitio web Y.

Las matrices de co-enlaces normalizadas fueron procesadas en SPSS para llevar acabo el análisis de escalamiento multidimensional (MDS). El mapa MDS para el índice FTSE 100 contenía un número tan elevado de empresas que hacía inviable la observación de grupos bien diferenciados de cara a la interpretación de los resultados. Por lo tanto, en este caso, se decidió reducir el número de empresas de 100 a 80 y posteriormente a 35. Para ello se ordenaron las empresas por el número de enlaces recibidos y se seleccionaron las empresas con mayor número de enlaces. Ambos mapas, basados en las 80 y 35 empresas con mayor número de enlaces, se incluyen en el trabajo para su análisis (Figuras 4.4 y 4.5). Una de las empresas incluida en el Euro Stoxx 50, Generali, no presenta co-enlaces con la mayoría del resto de empresas en el índice por lo que se omite del análisis.

Los *stress values* de los mapas MDS generados (valores que indican la bondad del ajuste de los datos al representarlos en el mapa) son: 0,03 para Dow Jones; 0,07 para Euro Stoxx 50 (49 empresas); 0,06 para CAC 40; 0,05 para Ibex 35; 0,05 para FTSE con 80 empresas; y, 0,06 para FTSE con 35 empresas. Son unos valores suficientemente bajos, lo cual indica que existe un buen ajuste entre los datos y las

posiciones determinadas para las empresas en los mapas.

4.2.1.2. Resultados

Los cinco índices bursátiles son representativos de diferentes economías y difieren en su composición sectorial y en los marcos sociales, culturales y políticos en los que se desarrollan. En este apartado presentamos una descripción de cada índice y posteriormente realizamos una comparación entre ellos con el fin de obtener unas conclusiones globales.

Dow Jones Industrial

El índice Dow Jones Industrial está formado por las 30 principales empresas de los Estados Unidos. La composición del índice en el momento en que se lleva a cabo el estudio y la clasificación sectorial de las empresas que lo componen se pueden consultar en el Anexo 3. La Figura 4.2 representa el mapa correspondiente a este índice. Únicamente los *clusters* de empresas considerados significativos aparecen señalados con el fin de facilitar la interpretación.

Se pueden distinguir tres grupos principales de empresas:

- El grupo de tecnologías de las información (*IT*, en el mapa) incluye las siguientes empresas: telecomunicaciones basadas en líneas fijas (Verizon, AT&T), servicios de informática y *hardware* (IBM, HP), semiconductores (Intel) y *software* (Microsoft). Las posiciones de estas empresas dentro del *cluster* son significativas. Las tres empresas de *hardware* se encuentran próximas entre sí y lo mismo ocurre con las dos empresas de telecomunicaciones. Microsoft es la única empresa de *software* incluida en el Dow Jones, ocupando una posición dominante y claramente diferenciada en el mercado, por lo que se mantiene distante del resto de empresas.

- El grupo financiero (*Financial*, en el mapa) está compuesto por bancos y otras empresas de servicios financieros. Hay tres bancos en el índice: Bank of America (BAC), Citigroup (C), JP Morgan (JMP), que se encuentran localizados dentro de un triángulo en la parte superior del mapa. El *cluster* que conforman no se puede considerar perfecto ya que hay otras dos empresas que aparecen incluidas en el mismo espacio. Una cuarta empresa que hay que tener en consideración es American Express, que se encuentra situada en el mapa cercana al grupo de tecnologías de la información. Una característica común a las empresas financieras en este trabajo es que se localizan en una posición similar, como puede observarse también en las Figuras 4.6 y 4.8.

- Se pueden identificar claramente otros tres *clusters*: farmacéutico, petróleo y gas y transporte. Distribuidas en torno a estos grupos se sitúan otras empresas que pertenecen a diversos sectores. Una característica común a todas ellas es que pertenecen a industrias más tradicionales, refiriéndonos con ello a que su actividad no se encuentra tan centrada en la información.

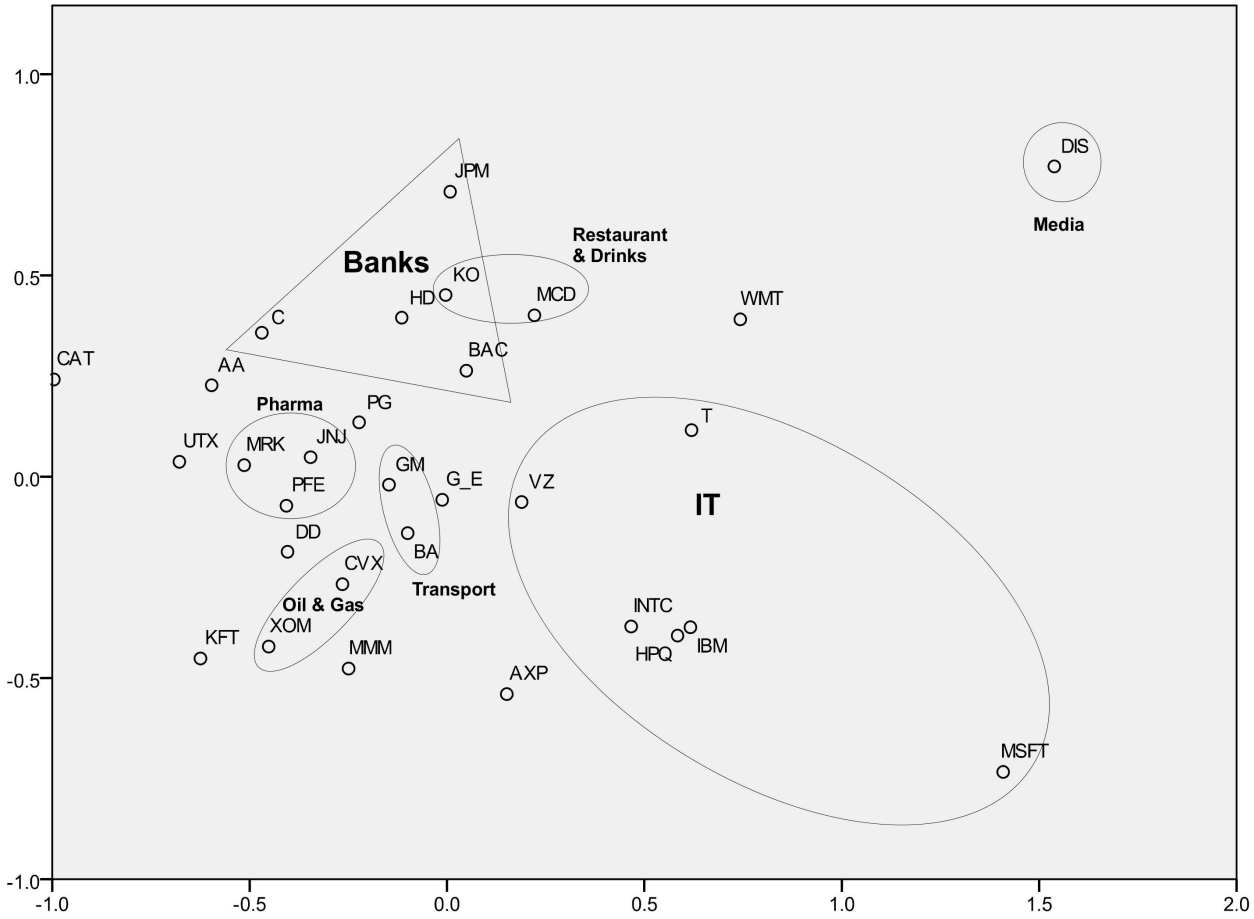
Se han de destacar también las posiciones de otras dos empresas. Disney es el único conglomerado de medios incluido en el índice, por lo que no compite directamente con ninguna otra empresa en el mapa. La naturaleza de su actividad, con la mayoría de sus productos y servicios pudiendo ser digitalizados, implica una fuerte presencia en la Web, que es también utilizada como uno de sus principales canales de distribución. La empresa recibe 6.070.000 enlaces externos, situándose en segundo lugar únicamente por detrás de Microsoft, con 46.600.000. El hecho de que la empresa reciba millones de enlaces, pero que únicamente un pequeño porcentaje de ellos sean co-enlaces con otras empresas, junto con la falta de competencia con otras compañías incluidas en el índice, explica el porqué Disney se sitúa en unos de los límites del mapa. Aunque se trata de una única empresa, Disney aparece también señalada dentro de un círculo en la Figura 4.2 con el objeto de comparar mejor su posición con otras empresas de medios presentes en

los demás índices.

Al igual que Disney, Wal-Mart se sitúa en una posición muy externa en el mapa, con un total de 1.210.000 enlaces externos recibidos. El hecho de que una empresa de comercio generalista con centros comerciales por todo Estados Unidos reciba tal cantidad de enlaces se puede explicar por la evolución de su modelo de negocio que ha derivado hacia la venta y distribución de productos por Internet. Se trata de la única empresa, aparte de las de tecnologías de la información y medios, que recibe más de un millón de enlaces.

En relación con el número de enlaces recibidos, es importante señalar que las empresas en la parte derecha del mapa son las que reciben un mayor número de enlaces. Todas las empresas (excepto Verizon) reciben más de un millón de enlaces. Las empresas con un menor número de enlaces (menos de 150.000 enlaces) se sitúan en la parte izquierda del mapa. Las empresas localizadas en el centro del mapa oscilan en su mayoría entre 200.000 y 300.000 enlaces. No es sorprendente que las empresas de tecnologías de la información y las orientadas al consumidor final (por ejemplo, Disney y Wal-Mart) sean más visibles en la Web y por tanto atraigan un mayor número de enlaces. Un estudio realizado en el mercado chino también ha encontrado evidencia de que las empresas con una mayor orientación al consumidor final reciben más hiperenlaces (Vaughan, Tang y Du, 2009).

Figura 4.2. Mapa MDS del Dow Jones Industrial

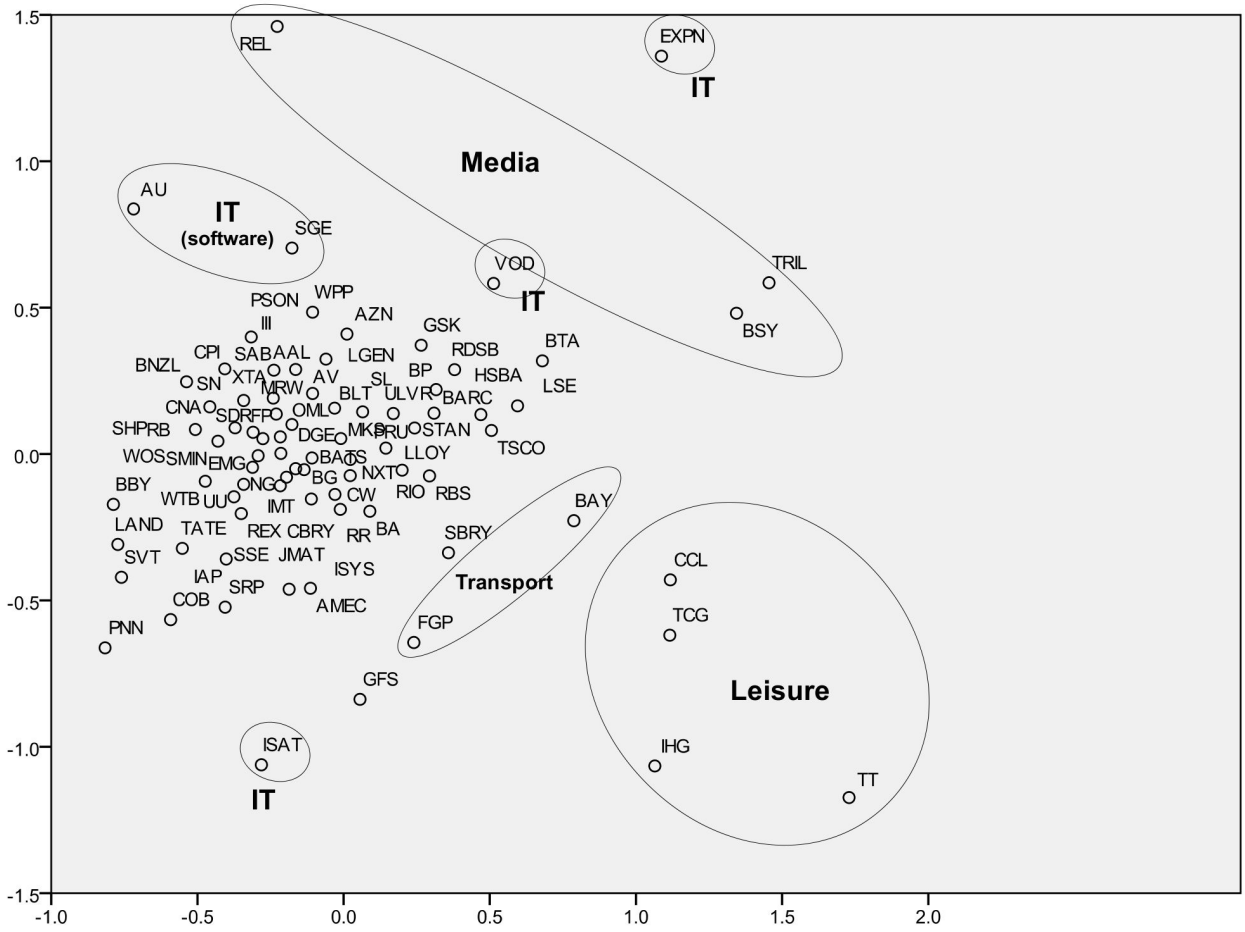


FTSE 100

Las empresas representadas en el FTSE 100 reúnen más del 80% de la capitalización de mercado de la *London Stock Exchange*, constituyendo el principal índice de referencia del mercado británico. La Figura 4.3 muestra el mapa con las 80 principales empresas del índice en función del número de enlaces recibidos. A pesar de que la alta concentración de puntos en el centro del mapa hace la interpretación prácticamente inviable, este retrato es especialmente útil para mostrar claramente cómo determinados grupos están situados en las partes más

externas del mismo. Estos grupos son los correspondientes a tecnologías de la Información (*IT*), medios (*Media*) y Ocio (*Leisure*). Dado que las empresas de tecnologías de la información incluidas en este índice no son tan homogéneas como las incluidas en el Dow Jones, no se encuentran situadas en un único espacio; sin embargo, todas ellas se localizan en una posición periférica del mapa, como ocurre también en la Figura 4.2. En este caso, tres de las cuatro empresas más importantes de medios se hallan localizadas en una zona común. Solamente Pearson permanece fuera de este grupo aunque muy cercana; en todo caso, su afiliación al mismo aparece de forma más clara en la Figura 4.4. El grupo de empresas de ocio está conformado por empresas orientadas fundamentalmente al usuario final, tales como: Carnival, Thomas Cook y Tui Travel (agencias de viajes), e Intercontinental Hotels (cádena de hoteles). El sector de los viajes es una de las actividades que también ha conseguido implementar un modelo de negocio basado de forma intensiva en el comercio electrónico. Aunque no se aprecia en el mapa, las empresas del sector financiero se encuentran relativamente bien agrupadas en una posición intermedia en la parte derecha de la nube de puntos.

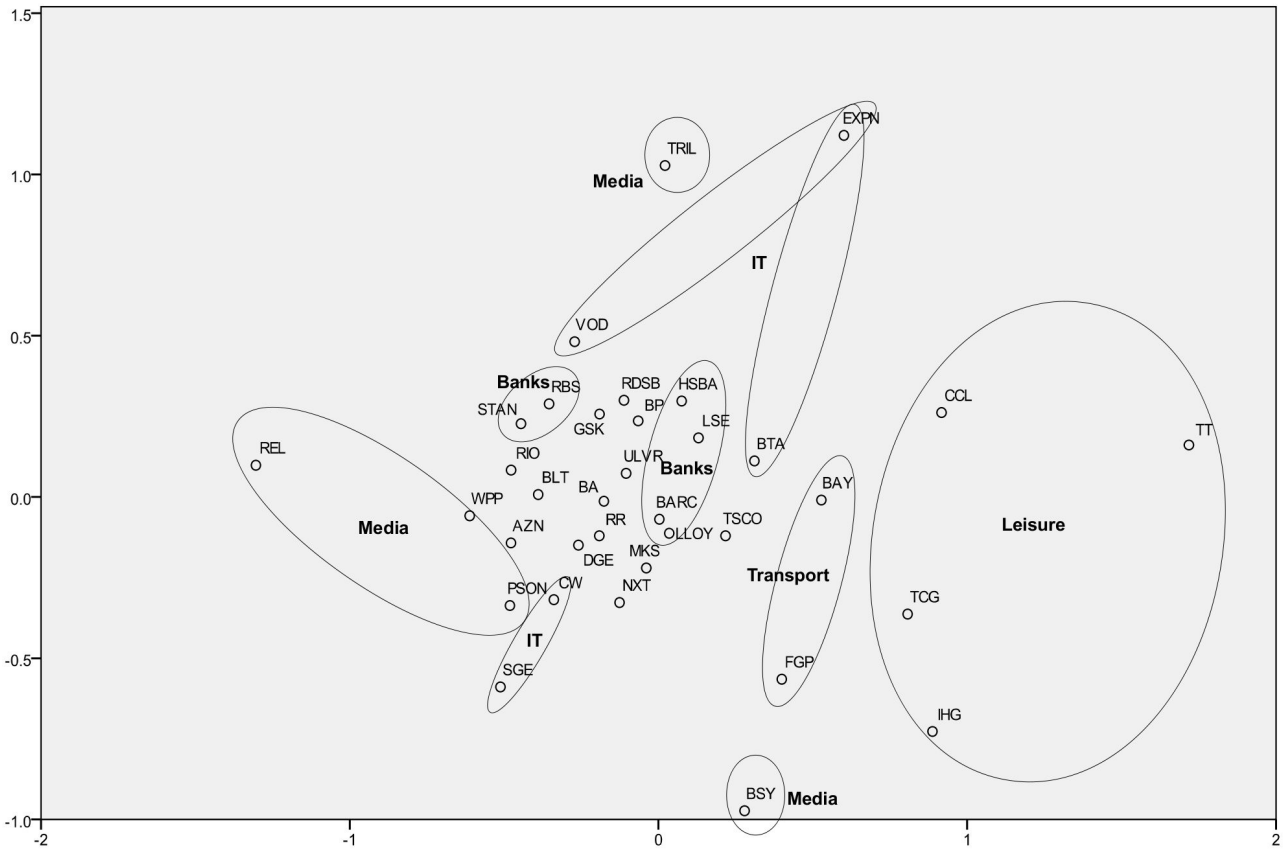
Figura 4.3. Mapa MDS del FTSE 100 (80 empresas con mayor número de enlaces recibidos)



La Figura 4.4 incluye a las 35 empresas del FTSE 100 que reciben un mayor número de enlaces. Esta selección probablemente sobrepondera la presencia de determinados sectores que, por su naturaleza, son proclives a recibir más enlaces (tales como el sector de medios o de tecnologías de la información). Esta selección proporciona una oportunidad para comprobar si podemos observar *clusters* mejor definidos utilizando una perspectiva con mayor detalle. El mapa MDS resultante constituye una representación diversa de empresas, mostrando unos patrones de agrupación consistentes que se han identificado también en otros índices. Los resultados sugieren que las empresas pertenecientes a tecnologías de la

información, medios y ocio se sitúan en la parte más externa del mapa, con los bancos situados en una posición próxima a ellos. Los *clusters* en este caso no se distinguen con tanta claridad, debido probablemente a la heterogeneidad de las actividades y al elevado número de empresas incluidas en los grupos de tecnologías de la información y de medios. Por ejemplo, las empresas de medios incluyen actividades tan distintas como las basadas en el negocio de la televisión, editorial, prensa o servicios de marketing (esta actividad, si bien no es estrictamente de medios, sí se encuentra vinculada).

Figura 4.4. Mapa MDS del FTSE 100 (35 empresas con mayor número de enlaces recibidos)

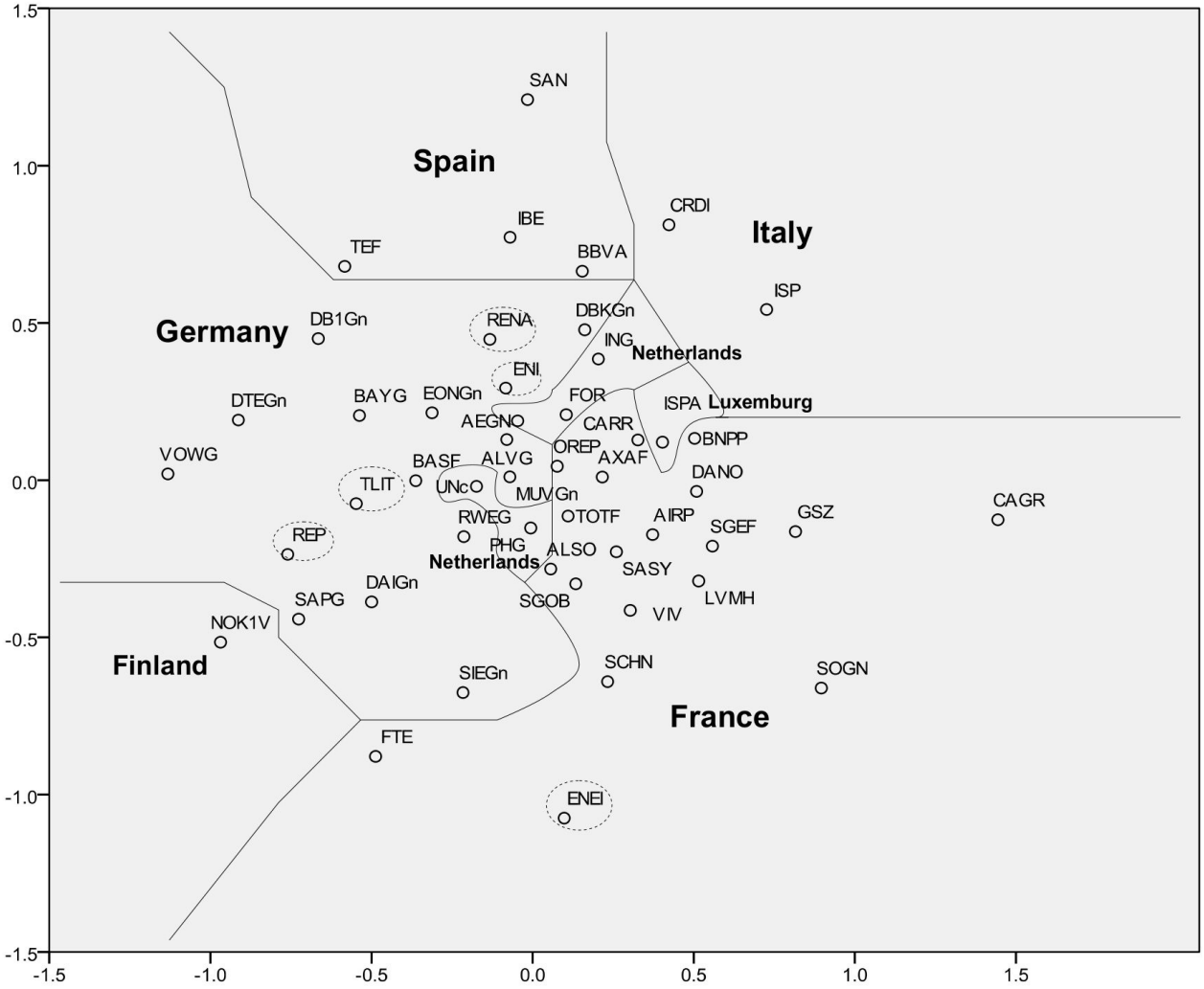


Euro Stoxx 50

El índice Dow Jones Euro Stoxx 50 proporciona una representación de las empresas más grandes de la Eurozona. Se encuentra compuesto por 50 empresas de siete países: Francia (19), Alemania (13), Holanda (5), Italia (6), España (5), Finlandia (1) y Luxemburgo (1).

El índice es heterogéneo en cuanto a países y sectores, por lo que se incluyen dos interpretaciones diferentes del mapa, basadas en estos criterios. En primer lugar, la Figura 4.5 muestra que el mapa MDS de acuerdo con la procedencia nacional de las empresas. Las empresas pertenecientes a un mismo país aparecen juntas. Esto se observa especialmente en el caso de las empresas francesas. En la zona correspondiente a las empresas de Alemania podemos encontrar también algunas pertenecientes a otros países (se señalan mediante la línea de puntos). Sin embargo, veremos posteriormente que este hecho puede explicarse en algunos casos por la fortaleza de la variable "sector" al posicionar las empresas. La mayoría de empresas españolas aparecen agrupadas también muy cercanas entre sí. Las empresas holandesas se sitúan en una posición central en el mapa entre las francesas y las alemanas. Esta posición refleja la situación real tanto desde un punto de vista geográfico como económico en la Eurozona, donde las empresas holandesas, debido al pequeño tamaño de su mercado nacional, tienden a ser más internacionales, especialmente en sus mercados próximos. La única empresa luxemburguesa es ArcelorMittal (ISPA), que se encuentra situada dentro del espacio francés. Esto se entiende mejor si consideramos que se trata de una empresa que es resultado de la fusión de tres corporaciones en 2002: Aceralia (España), Usinor (Francia) y Arbed (Luxemburgo). La proximidad geográfica también refuerza la posición mostrada en el mapa.

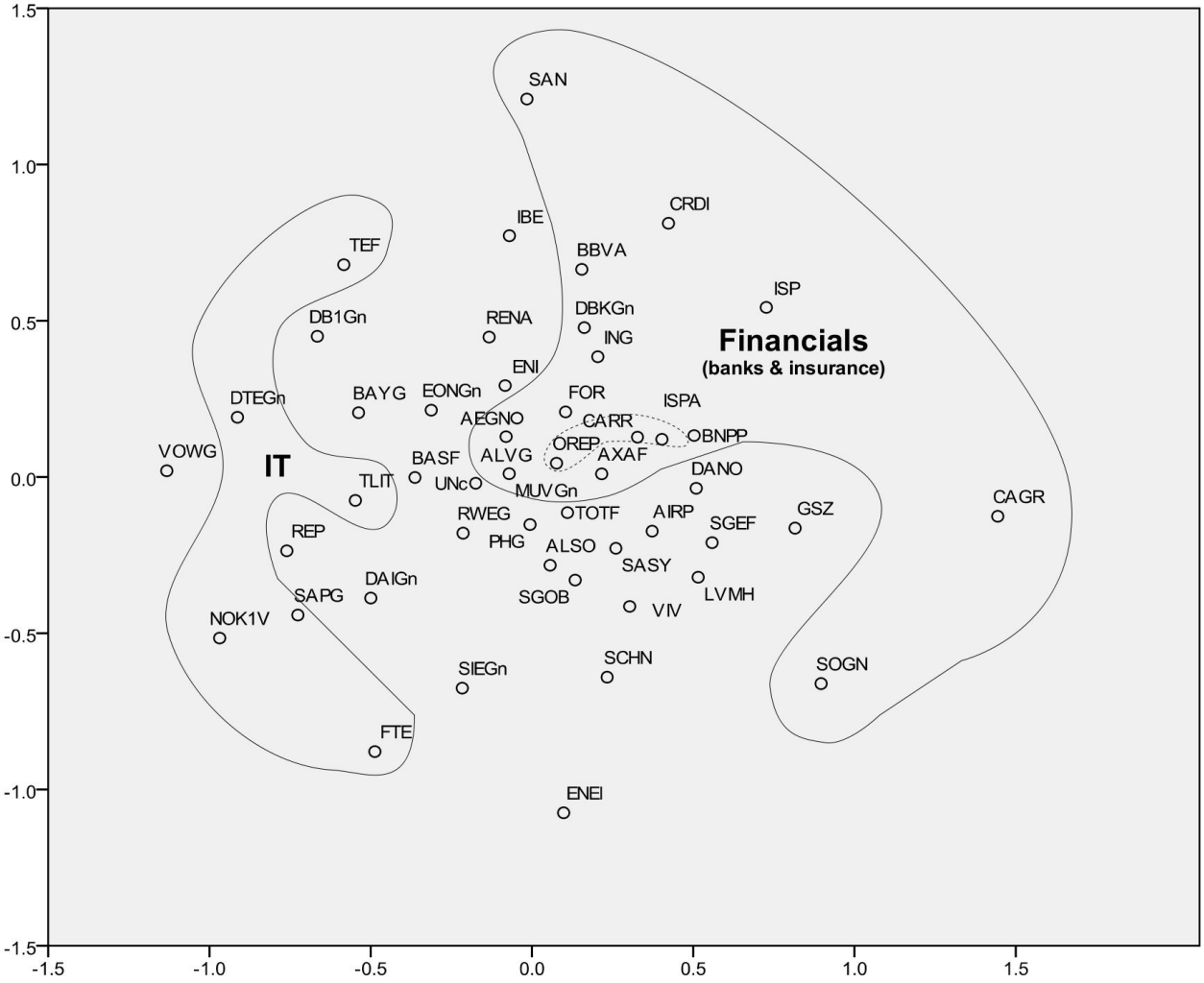
Figura 4.5. Mapa MDS del Euro Stoxx 50, con interpretación basada en países.



La Figura 4.6 contiene el mismo mapa interpretado en función de los sectores de actividad. Únicamente se han identificado claramente dos grupos: tecnologías de la información y financiero. Cabe notar que ninguna empresa de medios o de ocio está incluida en el índice. El grupo de empresas financieras está principalmente compuesto por bancos y empresas de seguros. Todos ellos, con la excepción de Deutsche Boerse (DB1Gn), que está próxima al grupo de tecnologías de la información, se encuentran agrupados en una amplia área con independencia de país al que pertenecen. Hay también tres empresas (marcadas con una línea de

puntos) de otros sectores que aparecen dentro de este *cluster*. Las empresas de tecnologías de la información se sitúan en la parte izquierda del mapa, incluyendo seis empresas de cinco países distintos (dos de ellas son alemanas). La existencia de este *cluster* de empresas de tecnologías de la información explica por qué la zona alemana del mapa (Figura 4.5) incluye algunas empresas procedentes de otros países. Las empresas de tecnologías de la información han emergido de forma consistente como un grupo claramente definido en todos los índices. Esto se puede explicar por la naturaleza de su actividad, basada de manera intensiva en contenidos e infraestructuras vinculadas a la economía de la información y el conocimiento. Las empresas financieras en el mapa también forman un grupo relativamente claro, debido, entre otros factores, a la naturaleza de su actividad cada vez más basada en el procesamiento de información. La integración económica en la Eurozona es probable que sea más profunda en estos sectores, ya que para la provisión de bienes y servicios las limitaciones geográficas pueden ser menores.

Figura 4.6. Mapa MDS del Euro Stoxx 50, con interpretación basada en sectores de actividad

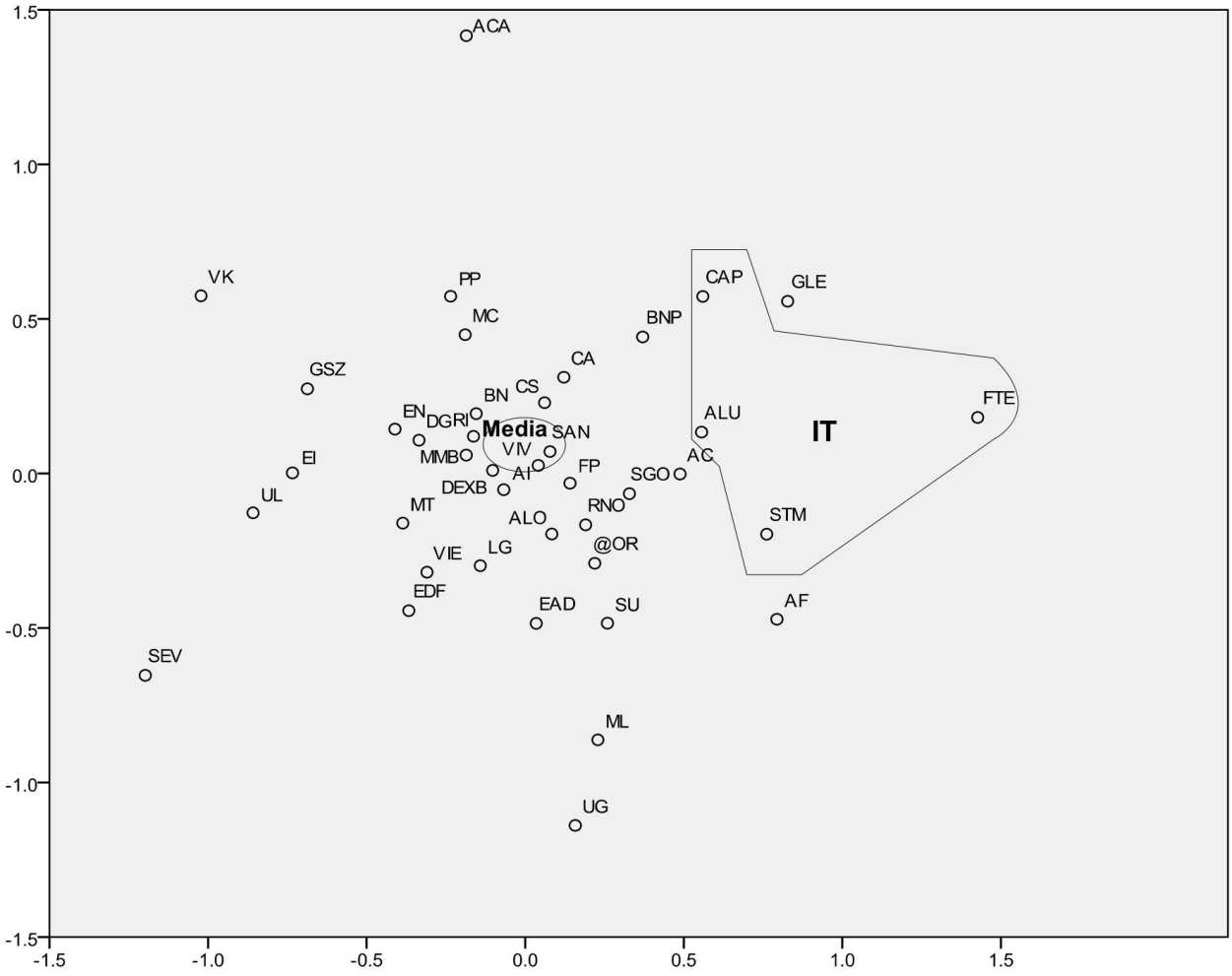


CAC 40

El CAC 40 es el índice de referencia del mercado de valores francés y representa a los 40 valores con mayor capitalización bursátil de la *Euronext Paris Bourse*. La Figura 4.7 muestra el mapa para este índice. Se observa que las empresas de tecnologías de la información están en un *cluster* bien diferenciado y separado del

resto de empresas. Tanto la posición como el hecho de que estas empresas reciban un mayor número de enlaces son coherentes con la evidencia encontrada en otros índices. Accor (AC), una cadena de hoteles (industria del ocio, *Leisure*) recibe un gran número de enlaces y está situada próxima a las empresas de tecnologías de la información. Los bancos están más dispersos en el mapa. Sin embargo, dos de ellos están localizados cerca de las empresas de tecnologías de la información: BNP Paribas (BNP) y Société Générale (GLE). Otro banco, Crédit Agricole (ACA) se encuentra en uno de los límites del mapa. Sin embargo, unas pocas empresas en otros sectores presentan posiciones no esperadas en zonas más externas del mapa: Peugeot (UG, automóviles), Suez (SEV, servicios), Vallourec (VK, maquinaria industrial). Una posición que es claramente inconsistente con resultados anteriores es la de Vivendi (VIV), una empresa de medios que se encuentra situada en el centro del mapa. También Lagardere (MMB), un grupo con actividades muy diversas, pero con un segmento editorial muy importante, se halla en el centro del mapa próximo a Vivendi. Además, el número de enlaces recibidos de estas empresas es significativamente bajo en relación con otras empresas de medios en otros índices. Estos resultados podrían explicarse parcialmente por un problema de sesgo basado en la lengua. En todo caso, es necesaria investigación adicional en este mercado en concreto para entender mejor el caso de algunas empresas en particular.

Figura 4.7. Mapa MDS del CAC 40

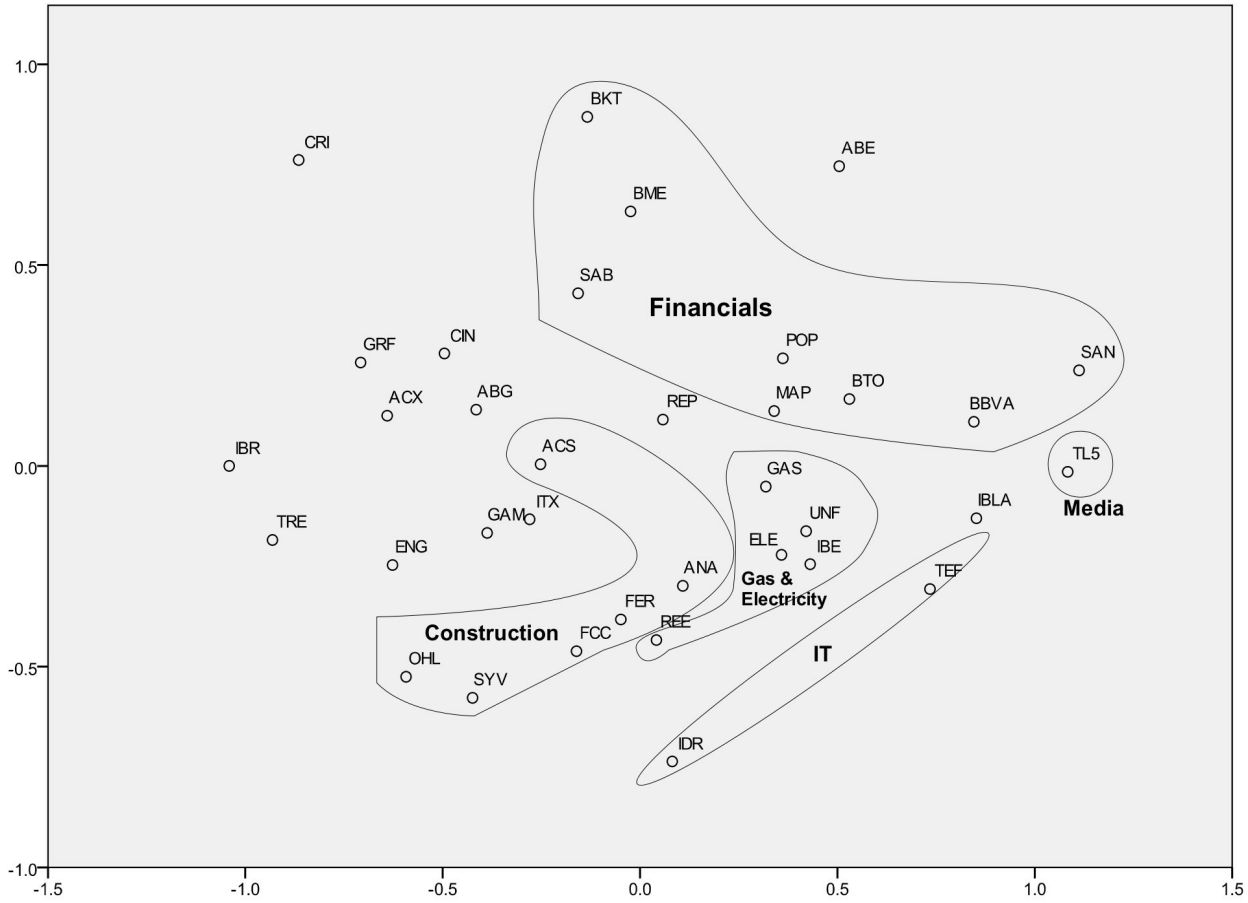


Ibex 35

El IBEX 35 es el índice de referencia de la Bolsa de Madrid e incluye los 35 valores españoles con mayor liquidez del mercado continuo. La Figura 4.8 confirma algunos de los patrones detectados en otros índices. La única empresa de medios, TL5, se encuentra situada en la parte externa del mapa. El conjunto con las dos empresas de tecnologías de la información se sitúa próximo. El resto de empresas en el centro del mapa pertenecen a distintos sectores, sin mostrar ningún patrón claro o

bien sin un número suficiente de empresas para formar *clusters*. Con todo, dos de los sectores más importantes de la economía española aparecen claramente delimitados: construcción y gas y electricidad. El sector de la construcción es uno de los sectores líderes en el país si bien hoy en día es el que está sufriendo la crisis económica de manera más profunda.

Figura 4.8. Mapa MDS del IBEX 35



4.2.1.3. Conclusiones

La originalidad de este trabajo se basa en la aplicación del análisis de co-enlaces a empresas de naturaleza heterogénea incluidas en cinco de los principales índices bursátiles del mundo. Estudios previos basados en co-enlaces se centraron en el examen de un único sector con resultados positivos a la hora de localizar geográficamente sus posiciones competitivas en un mapa. Sin embargo, debido a la diversa adscripción sectorial de las empresas en este trabajo, se espera que las empresas pertenecientes a un mismo sector aparezcan reunidas en *clusters*, de acuerdo con el principio de similitud en el que se basan las investigaciones basadas en co-enlaces. Se puede llevar a cabo un análisis competitivo, pero únicamente dentro de un *cluster* previamente identificado de empresas homogéneas. En cualquier caso, esto no es posible realizarlo en este trabajo debido al escaso número de empresas en los índices que pertenezcan a un mismo sector. Hasta ahora, los estudios basados en tipos de organizaciones heterogéneos se han desarrollado en el marco de la teoría de la triple hélice (Stuart y Thelwall, 2006; García-Santiago y Moya-Anegón, 2009). Sin embargo, su perspectiva no es puramente comercial o económica, sino que se centran en el examen de la transferencia de conocimiento entre distintos ámbitos de la sociedad (universidad, empresa, administración pública). En el apartado anterior nos centramos en los resultados, considerando los mapas de manera individualizada. Sin embargo, hay algunas conclusiones generales que pueden generar nuevas ideas para desarrollar la investigación en el futuro y para avanzar en la aplicación de las técnicas webmétricas al ámbito de la empresa.

Los mapas individuales muestran que la expectativa de localizar a las empresas de un mismo sector agrupadas en *clusters* bien diferenciados no se cumple plenamente. Los sectores cuyos modelos de negocio se encuentran más centrados en la información forman *clusters* en posiciones en los extremos de los mapas, alejados del resto de empresas que ocupan posiciones más centrales. Los sectores centrados en información, de acuerdo con nuestra interpretación, comprenden principalmente cuatro grandes grupos: empresas basadas en la producción de contenidos (identificados como *Media* en el estudio), empresas que proporcionan

servicios e infraestructuras basados en la información (tecnologías de la información, *IT*), empresas que aplican de manera muy intensiva procedimientos de comercio electrónico (especialmente empresas vinculadas al turismo y al ocio) y empresas de actividades financieras. Las empresas financieras, que incluyen bancos, aseguradoras y otros servicios financieros, son intensivas en información en el sentido de que en muchos casos no proporcionan servicios físicamente tangibles o de que sólo operan en ámbitos virtuales (por ejemplo, empresas de bolsas de valores, bancos que en muchos casos están presentes principalmente en Internet, etc.).

Además de sus posiciones en los mapas, las empresas de medios y de tecnologías de la información reciben más enlaces que cualquiera de las otras empresas que pertenecen a sectores más tradicionales (en el sentido de basados en modelos de negocios no tan centrados en los flujos de información y en la digitalización que ello conlleva). El patrón de enlaces recibidos refleja un modelo de negocio que está en gran medida expuesto a Internet. Esto también es cierto para las empresas financieras, en cierta medida. En términos generales, las empresas financieras están en segunda posición tras las empresas de tecnologías de información en el número de enlaces recibidos. Esto explica el porqué, en la mayoría de los índices, las empresas financieras están localizadas en una posición más intermedia entre el centro del mapa, donde se localizan las empresas más tradicionales, y la parte más exterior con las empresas de medios y tecnologías de la información. De este modo, se puede afirmar que las empresas que se localizan en posiciones más alejadas del centro del mapa son aquellas que se separan, bien por la propia naturaleza de la actividad o por decisión de la dirección, de un modelo de negocio más tradicional hacia otro que se desarrolla cada vez más en un contexto digital en la Web. Esta interpretación se ve reforzada por el empleo del índice Jaccard en el análisis de los datos. Debido a que las empresas más centradas en información reciben un mayor número de enlaces, de los cuáles sólo un número más reducido son co-enlaces, los valores obtenidos al aplicar el índice Jaccard tienden a ser más pequeños. Esto hace que estas empresas se posicionen lejos del centro, en los extremos del mapa.

¿Por qué se identifican diferentes patrones en función del empleo de la información en el sector o en una empresa en particular? Las denominadas empresas centradas en la información han sufrido un proceso de digitalización con el resultado de un incremento de su presencia en la Web. Los cambios más agudos afectan a la industria de producción de contenidos (el sector de medios, por ejemplo, la prensa, música, cine, etc.). Estas empresas necesitan reinventar sus modelos de negocio con el objeto de hacer frente al proceso de digitalización de las últimas décadas. En la mayoría de los casos, estas transformaciones conllevan graves problemas relacionados con la propiedad intelectual de los contenidos. Empresas que en el pasado parecían gozar de posiciones estables basadas en el dominio del mercado se han encontrado compitiendo en un contexto social y económico totalmente nuevo y plagado de retos.

El Euro Stoxx 50 es el único índice en el estudio que está formado por empresas que pertenecen a diferentes países. Este hecho nos permite observar la integración económica en la Eurozona. Los resultados muestran que la interpretación por países del mapa de dicho índice se ajusta bien a la dimensión geográfica de las economías incluidas. Si nos fijamos en sectores de actividad, excepto en el caso de los sectores basados en información (tecnologías de la información y financiero), no se identifican *clusters* bien definidos en el mapa. Como contraste, el mapa correspondiente al Dow Jones está más claramente definido en términos de sectores de actividad, pudiendo subrayar una mayor integración económica en el mercado de los Estados Unidos. A pesar de los avances realizados en el contexto europeo, la existencia de diferencias culturales, lingüísticas y geográficas complica lograr un grado de integración similar. Es posible que próximos estudios revelen una menor división por países conforme la integración de los mercados continúe desarrollándose en Europa. Consideramos que una efectiva integración económica se puede lograr más fácilmente en sectores que están basados en tecnologías de la información. Dicha integración se ve potenciada por el rápido proceso de digitalización de los países de la Unión Europea. En cualquier caso, podrían existir algunas limitaciones en el análisis de co-enlaces del Euro Stoxx 50 derivado de diferencias lingüísticas.

Este trabajo pone de relieve que la información de co-enlaces extraída de la Web constituye una fuente de información fácilmente accesible para obtener nuevas perspectivas sobre el mundo empresarial. El estudio de empresas heterogéneas pertenecientes a diferentes sectores permite a gerentes, analistas financieros y a otras partes interesadas situar la actividad de sus empresas en relación con otras e identificar modelos de negocio diferentes y su evolución a lo largo del tiempo.

Si una empresa intenta potenciar su presencia en la red, el análisis de co-enlaces podría emplearse por los directivos para monitorizar el progreso en sus esfuerzos, mediante la observación de cómo periódicamente se modifica la posición en los mapas en relación con otras empresas que sean referentes para la misma. De este modo, el análisis de co-enlaces podría integrarse como una herramienta de gestión adicional útil para empresas que especialmente se encuentren en procesos de modificación de sus estrategias en relación con las tecnologías de la información. Finalmente, el trabajo emplea datos obtenidos directamente de la Web con el fin de proporcionar evidencia que confirme los cambios en los modelos de negocio de sectores centrados en la información debido al proceso de digitalización de determinadas actividades producido a lo largo de las últimas décadas.

Las conclusiones que alcanzamos permiten abrir la puerta a futuros estudios que confirmen o rechacen algunos de los resultados obtenidos en el trabajo, sugiriendo nuevas direcciones en el estudio webmétrico de la realidad económica y empresarial. Otro importante ámbito de investigación es la detección de cambios en mapas MDS a lo largo del tiempo, bien a nivel de empresas determinadas, de sectores específicos o de múltiples sectores. Debido al rápido desarrollo de la Web durante la última década, es posible que las posiciones que ocuparan las empresas de tecnologías de la información y de medios años atrás no fueran tan externas en el mapa, existiendo una mayor cohesión. La evolución en los próximos años podría proporcionar confirmación acerca del cambio en los modelos de negocio. También se podría aplicar una perspectiva longitudinal con el objeto de monitorizar los cambios en la Eurozona para descubrir si las fronteras entre países en los mapas se hacen menos visibles y si los *clusters* por sector de actividad toman un mayor protagonismo. Finalmente, un análisis de contenido sería útil para examinar las

motivaciones que se esconden detrás de la creación de enlaces y, muy especialmente, para comprobar si existen diferencias significativas entre el tipo de páginas y las razones existentes al crear enlaces a empresas basadas en información y empresas en industrias más tradicionales.

Aunque el trabajo explora el empleo del análisis de co-enlaces para grupos heterogéneos de empresas, hay una serie de limitaciones que en muchos casos se encuentran fuera de nuestro control. En primer lugar, la interpretación de los resultados podría diferir debido a factores tales como las distintas clasificaciones de actividades, al empleo de criterios alternativos para la creación de *clusters*, o al conocimiento específico del investigador acerca de un determinado contexto económico. La composición de los índices, en cuanto al tipo de empresas que los integran, difiere en función de las características de cada país o región. También cabe señalar que el número de empresas en cada índice no es homogéneo. Por ejemplo, el FTSE 100 incluye varias empresas de ocio, mientras que el Dow Jones no. En ocasiones sólo hay una empresa perteneciente a un sector específico, lo cual impide determinar la existencia de *clusters*. En estos casos, se podría interpretar la posición de estás empresas en particular en relación con las posiciones relativas que ocupan empresas similares en otros índices. Finalmente existen diferencias lingüísticas que podrían sesgar los resultados, debido a que los buscadores en ocasiones priman la indexación de determinados contenidos. Este problema podría ser más significativo en el caso del Euro Stoxx 50, en el que se incluyen en el mismo mapa empresas de distintos países.

4.2.2. Análisis de las dificultades financieras en el sector bancario de los Estados Unidos a través del análisis de co-enlaces

El presente trabajo constituye un estudio fallido sobre la posibilidad de emplear datos de hiperenlaces para obtener información sobre la crisis financiera de 2009 en el sector bancario de los Estados Unidos. De manera más específica, el objetivo es valorar si el empleo de co-enlaces filtrados mediante el empleo de palabras clave refleja el grado de la crisis en los bancos estadounidenses. Para visualizar la información contenida en los co-enlaces, se empleará el escalamiento multidimensional con el propósito de observar si los bancos con problemas financieros forman *clusters*. También se recoge información financiera y económica de los bancos al objeto de verificar los resultados obtenidos a partir de los enlaces. Si bien en una primera medición en enero de 2009 existían expectativas de la viabilidad del proyecto, una segunda recogida de datos en agosto del mismo año no ofreció evidencia confirmatoria, por lo que se consideró apropiado no continuar con el proyecto. En la sección de conclusiones se examinan algunas de las razones por las que se cree que el proyecto no resultó viable.

4.2.2.1. Método

En estudios anteriores se han empleado co-enlaces para el análisis competitivo de sectores comerciales o para estudiar diversos sectores de forma conjunta, como en el caso del estudio incluido en el Apartado 4.2.1. Vaughan y You (2008) desarrollaron un procedimiento para combinar el análisis basado en el recuento de enlaces de la Web con un componente cualitativo mediante la inclusión de una palabra clave que permitiera filtrar los enlaces recuperados de acuerdo con determinados objetivos de investigación. En nuestro caso, aplicamos este análisis de co-enlaces filtrados mediante palabras clave para estudiar el impacto de la crisis económica y financiera. Los términos empleados para refinar la búsqueda son "crisis", "bailout" y "subprime", los cuales son usados con relativa frecuencia para

describir los problemas financieros a nivel internacional. Estas palabras clave tienen por objetivo excluir aquellas páginas que enlacen a los sitios de los bancos, tanto si son enlaces recibidos como si son co-enlaces, por temas distintos a los de la crisis financiera. Desde un punto de vista metodológico, este método proporciona nuevas posibilidades para emplear técnicas webmétricas en el ámbito empresarial. El análisis de co-enlaces viene apoyado por correlaciones entre el número de enlaces recibidos y los datos financieros disponibles sobre la gravedad de la crisis.

El sector bancario ha sido seleccionado porque se encuentra en el mismo origen de la crisis financiera a nivel mundial y, especialmente, en los Estados Unidos. La crisis en el sector muestra características particulares, por ejemplo, el caso de las hipotecas "subprime" o los planes de rescate del gobierno para salvar la industria. Otra de las razones por las que se ha elegido este sector es porque estadísticas sobre el empleo comercial de la Web suelen situarlo en los puestos más altos en cuanto a nivel de utilización, sólo por debajo del sector de tecnología de la información, que ya ha sido estudiado en anteriores trabajos. Estas características hacen que se trate de un sector apropiado y relevante para la presente investigación.

Es conveniente señalar que nuestro primer intento de estudio de la crisis se centró en el estudio de un conjunto de bancos internacionales; sin embargo, este ensayo contaba con algunas limitaciones importantes debido especialmente al sesgo introducido por el empleo de términos de distintas lenguas. El diferente desarrollo de las economías de los distintos países ante la crisis hace que también sea más complicada la comparación debido a la existencia de un contexto no homogéneo. Muchos países, por ejemplo, han establecido planes de rescate para ayudar a los bancos. Sin embargo, las condiciones que se aplican para la distribución de los recursos y los importes entregados difieren considerablemente y no siempre constituyen información pública.

Siendo esto así, decidimos centrarnos en el sector bancario de los Estados Unidos y, más específicamente, en los bancos cotizados en la New York Stock Exchange, NYSE (http://www.nyse.com). El Gobierno federal diseñó un plan de rescate sin

precedentes en cuanto a su alcance y cantidad de fondos. Con él se pretendía comprar activos financieros considerados como "tóxicos" y que no tenían liquidez en el mercado a fin de proteger el sistema financiero y evitar una quiebra generalizada. En el trabajo se han incluido 46 empresas cotizadas en la NYSE (www.nyse.com). Esto permite contar con un conjunto de los mayores bancos de Estados Unidos, así como de cierto grado de homogeneidad. Estas empresas aparecen incluidas en la página web de la NYSE bajo la siguiente etiqueta: "*Industry-Financials, Supersector-Banks, Sector-Banks, and Subsector-Banks*". Solamente se tienen en cuenta empresas nacionales de los Estados Unidos con acciones ordinarias cotizadas en el mercado. La lista de empresas se obtiene el 16 de enero de 2009 y se puede consultar en el Anexo 4, junto con la información de los bancos que han recibido fondos del Gobierno federal y su cuantía. Al margen de dichos bancos, Wachovia Corporation, que no se encontraba en la lista debido a su reciente fusión con Wells Fargo, también se incluye en la muestra.

En un principio, se consideró el empleo de distintas variables para medir el impacto de la crisis financiera, por ejemplo, el cambio en el precio de las acciones de los bancos a lo largo del último año. Tras probar distintas variables para las que se dispone de información pública, decidimos que un indicador razonable de las dificultades financieras de un banco es la cantidad de fondos recibida bajo el programa de rescate federal, el conocido como *Troubled Asset Relief Program* (TARP). Los datos correspondientes al mismo fueron obtenidos de un informe especial realizado por CNN Money (2009). En enero de 2009, sólo 22 de los 47 bancos en el estudio habían recibido dinero a través del TARP.

Yahoo! es el buscador seleccionado para la obtención de los datos de enlaces ya que ni Google ni Live Search (MSN) permiten el tipo de búsquedas que se indican en la Tabla 4.20.

Tabla 4.20. Términos de búsqueda empleados en Yahoo! para la obtención de enlaces recibidos y de co-enlaces

Tipo de información	Términos de búsqueda
Enlaces recibidos sin palabras clave	linkdomain:boh.com –site:boh.com
Enlaces recibidos con palabras clave	linkdomain:boh.com –site:boh.com (crisis OR bailout OR subprime)
Co-enlaces sin palabras clave	(linkdomain:boh.com -site:boh.com) AND (linkdomain:cnb.com -site:cnb.com)
Co-enlaces con palabras clave	(linkdomain:boh.com -site:boh.com) AND (linkdomain:cnb.com -site:cnb.com) (crisis OR bailout OR subprime)

Tras efectuar pruebas empleando los operadores *link* y *linkdomain*, los resultados mostraron que el empleo de este último permitía generar un mapa MDS que agrupaba los bancos con problemas financieros más claramente. Cuatro bancos tuvieron que ser excluidos del análisis de co-enlaces porque no presentaban ningún co-enlace con otros bancos en el estudio. En el caso de emplear el operador *link* en lugar de *linkdomain*, el número de bancos excluido habría aumentado.

4.2.2.2. Resultados

Debido a la no normalidad de las variables consideradas, se emplea el test de correlación de Spearman para determinar si existen relaciones entre ellas. Los resultados muestran la existencia de correlaciones significativas ($p<0,01$) entre el número de enlaces que apunta a la página web de los bancos y la cantidad de fondos recibidos para el rescate. Los coeficientes de correlación son 0,72 y 0,78, haciendo referencia el primero al número de enlaces recuperado sin emplear las

palabras clave y en segundo lugar, empleando las palabras clave. Ambos coeficientes de correlación son altos, aunque el coeficiente con las palabras clave es ligeramente superior, apuntando a que el empleo de las palabras clave como modo de filtrar la información podría mejorar los resultados obtenidos. En cualquier caso, la diferencia entre ambos coeficientes no es muy grande y no se puede alcanzar ninguna conclusión más definitiva. Una investigación adicional sería necesaria para profundizar en ello, contando probablemente con más datos sobre la evolución financiera de la crisis.

Los mapas obtenidos mediante escalamiento multidimensional a partir del primer conjunto de datos obtenidos en enero de 2009 pueden observarse en las Figuras 4.9 y 4.10. La Figura 4.9 ha sido generada a partir de los datos sin palabras clave, mientras que la Figura 4.10 muestra el resultado tras incluir las palabras clave. La intensidad del color negro que cubre los puntos indica aquellas empresas que han recibido más dinero (color más oscuro), mientras que los bancos que no han recibido fondos no aparecen redondeadas.

Figura 4.9. Mapa MDS sin palabras clave (enero 2009)

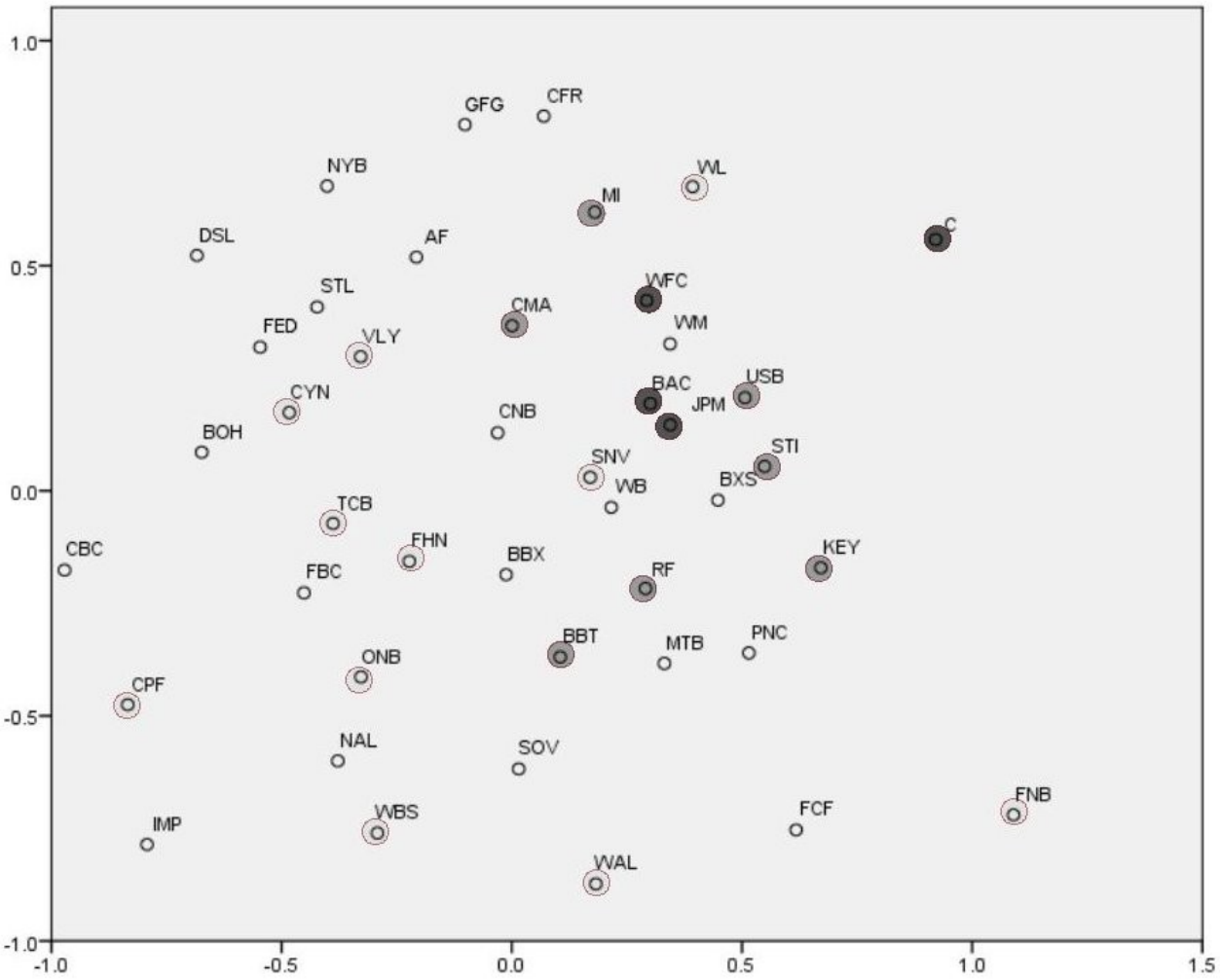
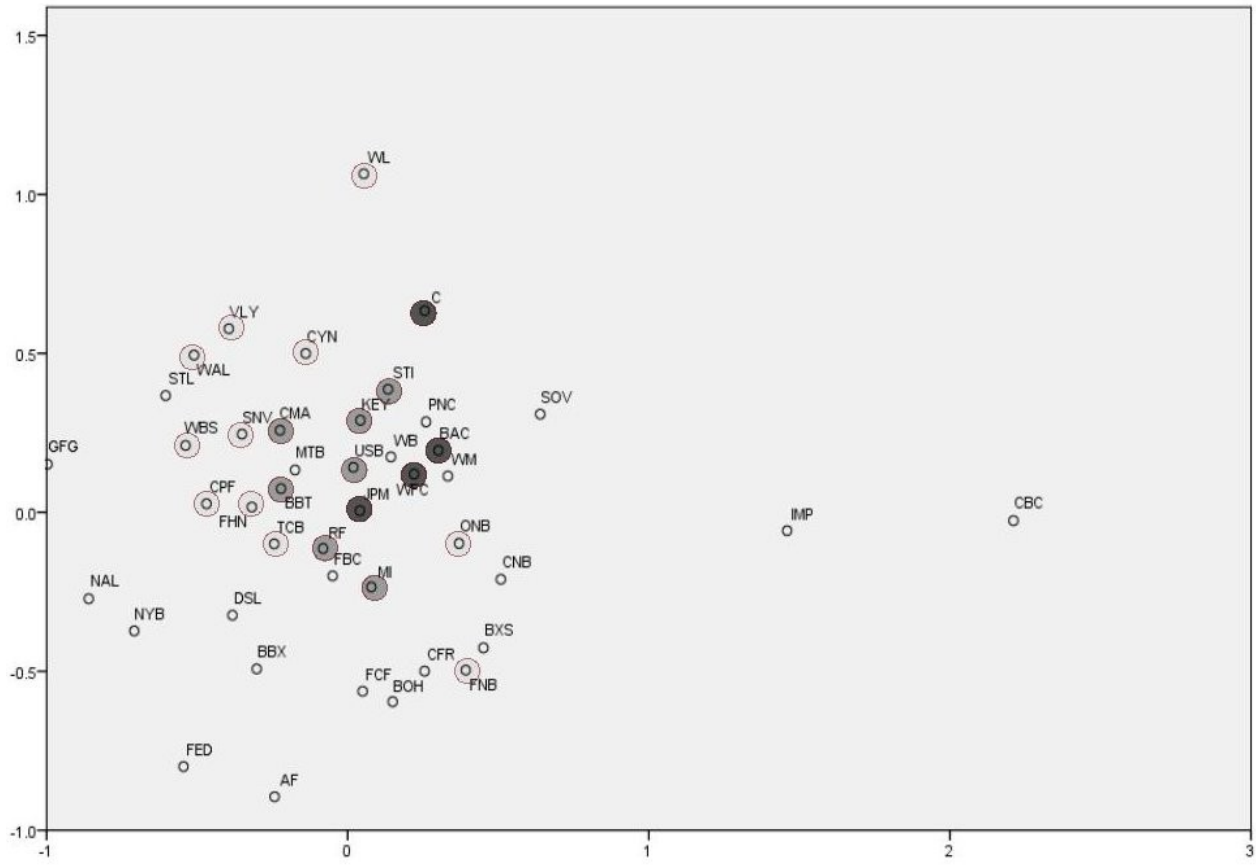


Figura 4.10. Mapa MDS con palabras clave (enero 2009)



Los cuatro bancos que aparecen en color negro más oscuro recibieron más de 10.000 millones de dólares, los siete bancos en tono gris recibieron entre 1.000 y 10.000 millones de dólares, mientras que, finalmente, los bancos con color gris claro fueron ayudados con un importe inferior a 1.000 millones de dólares. Comparando los dos mapas en las Figuras 4.9 y 4.10, podemos observar que el mapa filtrado con las palabras clave agrupa los bancos aproximadamente en tres franjas: con los bancos que recibieron mayor volumen de fondos en el centro, seguidos progresivamente por aquellos que fueron ayudados con cantidades menores, hasta finalmente localizar en el resto del mapa a los bancos que no necesitaron acudir a los fondos federales. Si asumimos que el dinero federal destinado al rescate se distribuye de manera proporcional a la gravedad de la crisis

de las entidades financieras, el mapa permite posicionar los bancos de acuerdo con el alcance de sus problemas financieros, situando a los más afectados en el centro del mapa y a los menos afectados fuera.

Por contra el mapa que no emplea las palabras clave (Figura 4.9) no presenta un patrón claro que refleje la gravedad de la crisis (los bancos más afectados por la crisis aparecen distribuidos por distintas partes del mapa). Este hecho sugiere que la incorporación de las palabras clave en la investigación de enlaces permite obtener información más específica sobre el evento o la circunstancia en estudio. En este sentido, es lógico pensar que aparecerán más co-enlaces con palabras clave entre bancos que se encuentren en una situación de crisis financiera análoga, ya que cuanto más en crisis esté un banco mayor será el número de páginas, abordando dicha crisis, que lo enlacen; es decir, acaparará más atención en la Web por parte de páginas que aborden este tema.

En agosto de 2009 se volvió a recoger información de co-enlaces con el fin de comprobar la evolución de los mapas y verificar su efectividad de cara a los objetivos propuestos. El análisis llevado a cabo es igual que el anterior, indicando con distinta gradación de grises el importe recibido para rescate de activos. Los resultados, incluidos en las Figuras 4.11 y 4.12 no reflejan la evolución en la crisis financiera, no detectándose ningún patrón significativo.

Figura 4.11. Mapa MDS sin palabras clave (agosto 2009)

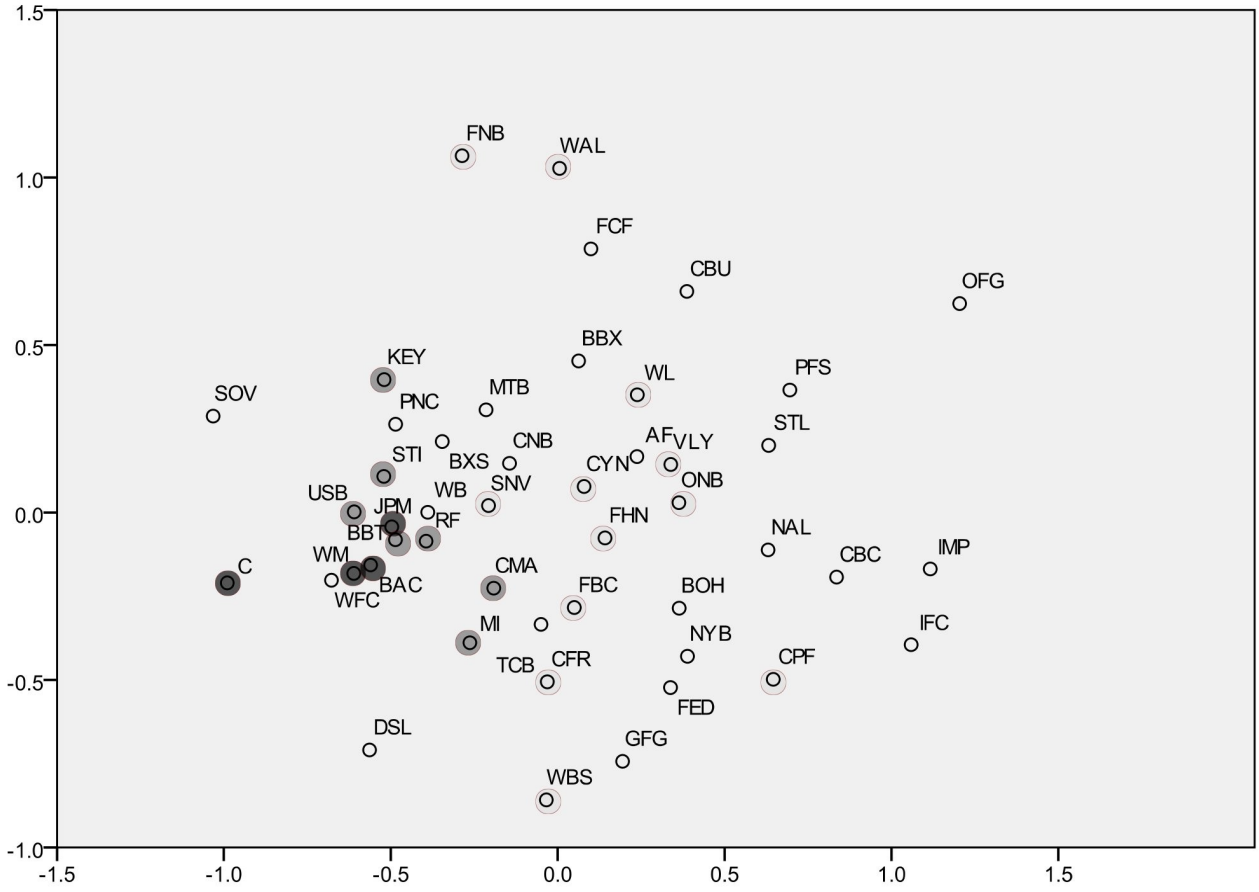
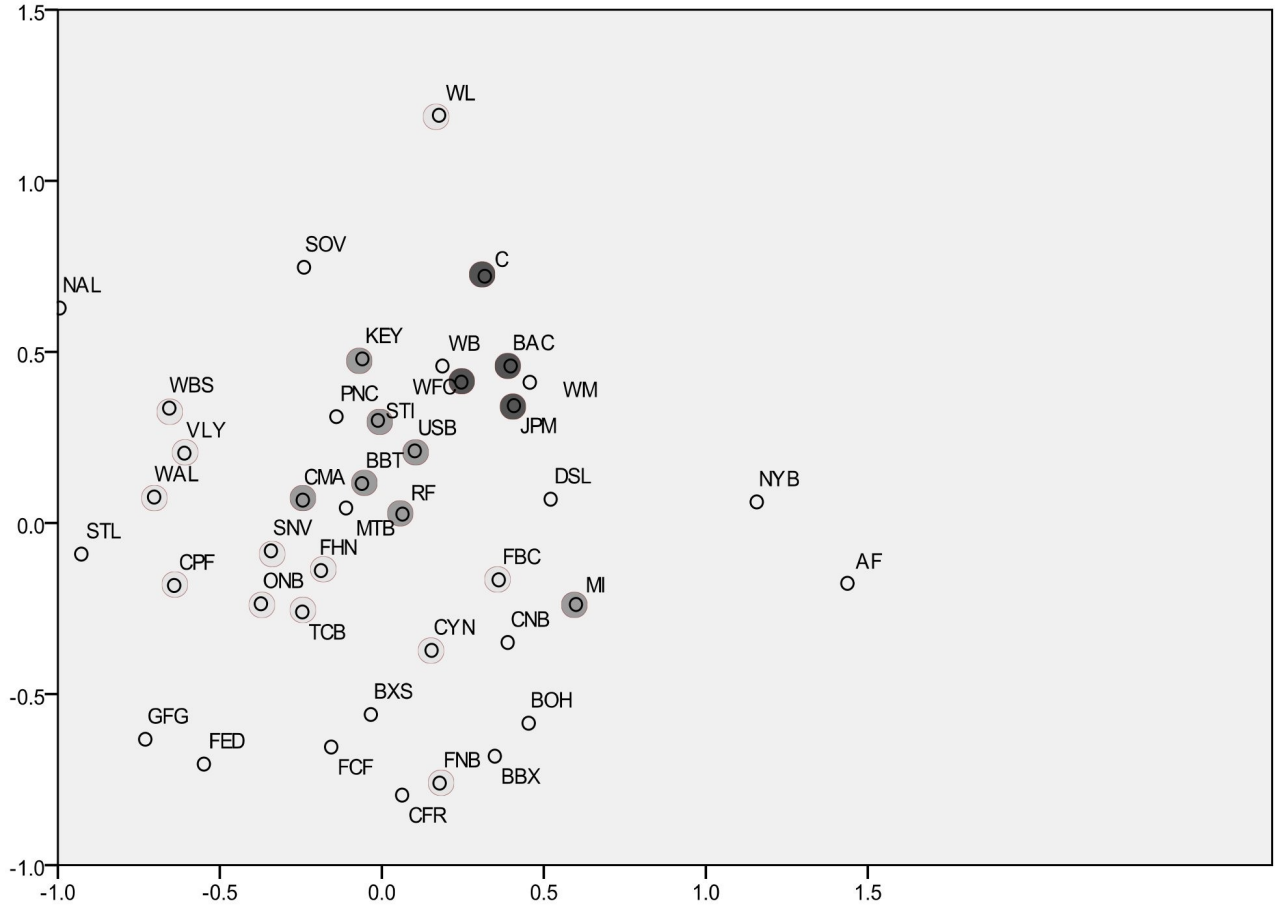


Figura 4.12. Mapa MDS con palabras clave (agosto 2009)



4.2.2.3. Conclusiones

La idea de emplear el análisis de co-enlaces filtrados por palabras clave con el fin de identificar aquellos bancos que estaban experimentando dificultades financieras no ha proporcionado los resultados esperados. Son varias las razones que pueden explicar esta situación. En primer lugar, podemos poner en duda si el empleo de co-enlaces es una medida adecuada de la incidencia de la crisis en los bancos. Con esto queremos decir que si una página web establece un enlace al sitio web de un banco con problemas financieros, mencionando una de las palabras clave que

estamos considerando, no necesariamente debemos esperar que enlace al mismo tiempo a otro banco con dificultades financieras. Es más, podría incluso darse el caso de que ese segundo enlace se realizara a un banco con fortaleza financiera con el fin de contraponerlo a la situación de crisis que sufre el primero. En este sentido, el análisis de co-enlaces no parece ser la mejor opción, debiendo primar un análisis de enlaces directos a dichos sitios web.

Existe además un efecto tamaño considerable que hace que, en terminos absolutos, los bancos mayores sean los que más dinero del TARP reciben, con lo que es probable que el posicionamiento de los bancos en el mapa no se deba tanto al nivel de problemas financieros sino al tamaño de los mismos. Es por ello también que habría que considerar algún tipo de medida relativa de las dificultades financieras experimentadas por las empresas.

4.3. Análisis combinado

En este último apartado se realiza un intento de combinar los dos tipos de análisis webmétricos anteriormente trabajados para proporcionar una perspectiva más completa de un sector empresarial determinado. La investigación webmétrica en el ámbito de empresas debe tender a desarrollar instrumentos de análisis que puedan integrarse como herramientas de gestión empresarial.

El trabajo que se incluye en la Apartado 4.3.1 pretende dar respuesta a la quinta pregunta de investigación de la tesis (Apartado 1.2.5): *¿Cómo se pueden combinar distintos tipos de análisis webmétrico para ofrecer una visión global de un sector empresarial?* A continuación se realiza un análisis de los principales bancos mundiales combinando los dos métodos webmétricos.

4.3.1. Análisis webmétrico del sector bancario internacional

El último trabajo incluido en el capítulo dedicado a la investigación empírica pretende aunar las dos principales técnicas webmétricas exploradas en los estudios anteriores con el objeto de avanzar en una futura metodología de análisis que permita sistematizar la obtención de conocimiento a partir de la información extraída de la estructura de enlaces de los sitios web de empresas. En concreto, el presente estudio analiza el sector bancario internacional. En primer lugar, pretendemos averiguar si existe algún tipo de correlación entre el número de hiperenlaces que el sitio web de un banco atrae y las variables financieras propias de dicho banco. En segundo lugar, empleamos el análisis de co-enlaces para elaborar un mapa con las posiciones globales que ocupan en el mercado los distintos bancos. Tanto los datos de enlaces recibidos como los de co-enlaces son obtenidos en dos momentos del tiempo, diciembre de 2008 y junio de 2009, con el objeto de determinar si se han producido cambios a lo largo de los seis meses, periodo de especial interés por hallarse el sector inmerso en una importante crisis financiera. La comparación efectuada viene guiada fundamentalmente por dos elementos, las diferencias

culturales y el rápido desarrollo de la economía asiática.

4.3.1.1. Método

Se han seleccionado para el trabajo a los 50 principales bancos del mundo en función de sus activos totales (teniendo en cuenta cifras consolidadas), tomados en la mayoría de los casos a fecha de cierre del ejercicio 2007. La lista de empresas fue obtenida el 4 de diciembre de 2008 de la siguiente página web: http://www.bankersalmanac.com/addcon/infobank/wldrank.aspx.

Los bancos proceden de un total de 15 países diferentes, como se puede observar en la Tabla 4.21. La lista completa de URLs, siglas empleadas para el análisis en los mapas y número de enlaces recibidos por los sitios web de los bancos aparecen en el Anexo 5.

Tabla 4.21. Número de bancos por país

País	Australia	Bélgica	Canadá	China	Dina-marca	Francia	Alemania	Italia
Nº de bancos	1	2	2	4	1	6	8	2
País	Japón	Holanda	España	Suecia	Suiza	Reino Unido	Estados Unidos	
Nº de bancos	5	3	2	1	2	6	5	

Las direcciones de los sitios web de los bancos se obtuvieron empleando Google y posteriormente comprobando cada una de ellas para asegurar su corrección. La gran mayoría de las empresas en el estudio disponían únicamente de un sitio web; sin embargo, para las que tenían varias URL alternativas, se examinó el número de enlaces que recibía cada una de ellas y se seleccionó aquella con un mayor

número de enlaces. Una de las opciones que se barajó fue la de emplear en la consulta las distintas URLs alternativas, sin embargo, Yahoo!, el buscador empleado para la obtención de los datos, no permitía este tipo de consultas complejas a la hora de obtener datos de co-enlaces.

Los datos financieros de los bancos se recogieron en noviembre de 2009, empleando la base de datos Mergent (http://www.mergent.com). Los últimos datos financieros disponibles correspondían al ejercicio 2008, si bien también se recabaron los datos de 2007 con el objeto de efectuar comparaciones entre las correlaciones con uno y otro grupo de datos de enlaces recibidos. Se tuvieron en consideración tres grupos de variables financieras:

- Variables de posición financiera: activos totales.
- Variables de desempeño financiero, en términos absolutos: ingresos totales y resultado neto.
- Variables de desempeño financiero, en términos relativos: *Return on Assets* (ROA).

Los datos financieros para algunos bancos no se encontraban disponibles en la base de datos, por lo que se omitieron al efectuar las correlaciones. Como se muestra en la Tabla 4.24, el número de bancos en cada una de las correlaciones oscila entre 38 y 41. Dado que sí existen datos de enlaces recibidos para todos los bancos en el estudio, los 50 bancos están presentes en el análisis de co-enlaces.

La obtención de información de co-enlaces se lleva a cabo empleando Yahoo!, ya que es el único gran buscador comercial que permite combinar operadores para realizar las consultas necesarias para nuestro trabajo. Tanto Google (Google, 2006; 2009) como MSN Live Search (Live Search, 2007) presentaban limitaciones en el momento de la recolección de los datos a principios de 2009. Estas limitaciones han sido abordadas previamente en la parte metodológica de las anteriores investigaciones expuestas. En el momento de la obtención de los datos, diciembre

de 2008 y junio de 2009, Yahoo! representaba la única alternativa viable para conseguir la información necesaria.

Dado que los buscadores comerciales cuentan con versiones en distintos países es posible que las páginas de un mismo país sean tratadas de forma más extensa y pormenorizada que las de otros, proporcionando resultados sesgados (Vaughan y Thelwall, 2004). Esto ocurre por ejemplo con China, país para el que Yahoo! opera con una base de datos propia (www.yahoo.cn). Para este trabajo se hicieron las oportunas comprobaciones de resultados empleando Yahoo! España y Yahoo! France, sin embargo los resultados arrojados fueron los mismos, por lo que se puede afirmar que pese a emplear distintas interfaces en cada país, la base de datos de búsquedas es común con la de Yahoo! Global, por lo que es esta versión (www.yahoo.com) la que se ha utilizado para realizar las consultas.

De los dos operadores con los que cuenta Yahoo!, *linkdomain* y *link*, explicados en apartados anteriores, hemos usado en el trabajo *linkdomain* ya que teóricamente nos interesa conocer todos los enlaces que apuntan al dominio de la empresa y no únicamente a su página de inicio. La sintaxis de los términos de búsqueda empleados aparece en la Tabla 4.22 utilizando como referencia las siguientes URLs ficiticias: www.abc.com y www.xyz.com. La parte www se elimina al objeto de incluir los subdominios en la búsqueda. Por otra parte, el componente *-site:abc.com* permite filtrar los enlaces internos que provienen del propio dominio en estudio. Dado que los co-enlaces hacen siempre referencia a dos sitios web, para el análisis estadístico se procesan en forma de una matriz simétrica, en la que el dato de cada cruce representa el número de co-enlaces entre el sitio web *x* y el sitio web *y*.

Tabla 4.22. Términos de consulta empleados en Yahoo!

Tipos de enlaces	Términos de búsqueda
Enlaces (inlinks) que apuntan a www.abc.com	linkdomain:abc.com –site:abc.com
Co-enlaces entre www.abc.com y www.xyz.com	(linkdomain:abc.com –site:abc.com) (linkdomain:xyz.com –site:xyz.com)

La primera ronda de datos se recogió en diciembre de 2008. En ese momento, la crisis financiera del sector bancario y la recesión económica mundial se encontraban en su punto álgido. Posteriormente se recogió el mismo de tipo de información de enlaces recibidos en junio de 2009, cuando la economía mundial empezaba a recuperarse.

El procedimiento de análisis empleado para estudiar la matriz de co-enlaces de cada grupo de datos es el escalamiento multidimensional (MDS), que permite generar un mapa con las posiciones relativas de los bancos. El escalamiento multidimensional emplea un método heurístico para situar los bancos con mayor número de co-enlaces entre ellos más próximos entre sí en un mapa de posiciones relativas. La lógica del análisis responde a la idea de que, desde una perspectiva global, los co-enlaces son creados con un propósito determinado y, por tanto, pueden ofrecer, de modo agregado, un patrón que nos proporcione información útil para conocer el sector. Los bancos que son similares o vinculados entre sí tienen más posibilidades de estar co-enlazados. En otras palabras, el número de co-enlaces entre dos bancos podría considerarse como una medida del grado de similitud entre ambas entidades. Cuanto mayor sea el número de co-enlaces, más similares o vinculados estarán los negocios o los servicios de los dos bancos. Dos empresas de un mismo sector que sean similares en cuanto a la actividad que realizan o a los servicios que prestan serán con toda probabilidad empresas competidoras. De este modo el mapa MDS se puede interpretar como un mapa de posiciones competitivas entre las empresas. Nuestro objetivo es emplear los mapas para establecer grupos de bancos competidores de modo que podamos observar

las posiciones de mercado de los mismos en relación con los demás. Igualmente se pretende examinar la viabilidad del análisis cuando se comparan mapas correspondientes a distintos momentos del tiempo. Se espera que de la observación de sucesivos mapas se pueda obtener información que refleje los cambios en el sector o en entidades en particular.

El número de co-enlaces obtenido ha sido normalizado mediante la aplicación del índice Jaccard para obtener una medida relativa del grado de similitud entre los bancos. Dicho medida ha sido ya examinada con propósito de los anteriores trabajos expuestos. Las matrices de co-enlaces normalizadas se han analizado mediante el programa estadístico SPSS. Los valores de ajuste (*stress value*) del análisis de escalamiento multidimensional son: 0,075, para la primer conjunto de datos, y 0,073, para el segundo. Consideramos que estos valores son lo suficientemente bajos como para afirmar que existe un buen ajuste entre los datos y las posiciones de los bancos en el mapa MDS.

4.3.1.2. Resultados

Correlaciones entre el número de enlaces y los datos financieros

Antes de llevar a cabo el análisis de correlaciones, se realizó un análisis estadístico de tipo descriptivo de ambos conjuntos de datos disponibles. Para los datos correspondientes a diciembre de 2008, la media y la mediana son respectivamente 133.396 y 52.100, mientras que para los datos de junio de 2009 son 131.564 y 42.700. El número de enlaces no varía mucho a lo largo del periodo de seis meses, indicando que el número total de enlaces recibidos es una medida relativamente estable a lo largo del tiempo. En cualquier caso si analizamos los cambios al nivel de los bancos individuales, encontramos que el número de enlaces para algunas empresas cambia de manera significativa, como se puede observar en la Tabla 4.23.

Tabla 4.23. Principales cambios en el número de enlaces recibidos entre diciembre de 2008 y junio de 2009

	5 principales bancos por incremento del número de enlaces recibidos		5 principales bancos por disminución del número de enlaces recibidos	
1	China Construction Bank Corporation	77,30%	The Norinchukin Bank	-54,06%
2	DZ Bank AG	48,19%	Wachovia Bank NA	-41,33%
3	Kreditanstalt für Wiederaufbau (KfW)	41,53%	Calyon	-39,54%
4	Agricultural Bank of China	41,45%	Fortis Bank SA/NV	-39,12%
5	Deutsche Bank AG	33,28%	Dresdner Bank Group	-39,02%

Los bancos alemanes se encuentran entre los principales en cuanto a incremento en el número de enlaces. Esto puede explicarse a partir de algunos datos económicos. Por ejemplo, el banco KfW (con un incremento del 41,53%) ha regresado a los beneficios en el primer trimestre de 2009, con un resultado consolidado de 80 millones de euros, después de dos años de fuertes pérdidas fruto de la crisis financiera. Por su parte, Deutsche Bank, el mayor banco alemán, tuvo un incremento del 33,2% en el número de enlaces recibidos. Durante el primer trimestre de 2009 tuvo un resultado neto de 1.200 millones de euros, en contraste con la pérdida de 141 millones para el mismo periodo del ejercicio anterior. Adicionalmente, se puede indicar que, en el primer trimestre de 2009, Deutsche Bank aceleró sus planes para hacerse con el control del mayor banco minorista del país Deutsche Postbank (Mergent, 2009b). Entre los cinco bancos con mayor incremento en el número de enlaces, dos son chinos. Los bancos chinos, como también se subrayará más adelante, han sido capaces de sortear la crisis de una mejor manera debido, entre otros factores, a una mayor regulación pública (Shaw y Parussini, 2009).

Por el contrario, bancos, como el Wachovia, Fortis y Dresdner, muestran un importante descenso en el número de enlaces recibidos para el mismo periodo. Este hecho puede interpretarse como una pérdida de interés por parte de inversores y otras partes interesadas. De nuevo, algunos datos económicos pueden ayudarnos a interpretar estos descensos. Wachovia Corporation (con un descenso del 41,33% en el número de enlaces) fue adquirida por Wells Fargo en diciembre de 2008 y, por lo tanto, dejó de ser una empresa independiente en esa fecha. La marca Wachovia será absorbida por Wells Fargo a lo largo de los próximos tres años. Fortis (con un descenso del 39,12%) ha sufrido serios problemas financieros a lo largo de los dos últimos años y la mayor parte de la compañía fue vendida en distintos paquetes en 2008. Finalmente, Commerzbank anunció la adquisición de Dresdner Bank (39.02% de descenso en el número de enlaces) el 31 de agosto de 2008. Esta adquisición por valor de 9.000 millones de euros se concluyó en el primer trimestre de 2009 (Mergent, 2009b). A pesar de que esta unión creará el segundo mayor grupo alemán junto con Deutsche Bank, la disminución en el número de enlaces recibidos por el Dresdner Bank indica una pérdida de atención hacia la entidad adquirida. Como se señala en el sitio web del Dresdner Bank, ésta es una marca de Commerzbank desde mayo de 2009. Por último, la situación del Norinchukin Bank podría entenderse como resultado de la mala situación que atraviesa el sector bancario japonés en su conjunto, especialmente en contraste con otros bancos asiáticos, como es el caso de los bancos chinos (Mergent, 2009a).

Dado que las distribuciones de frecuencias de ambos conjuntos de datos (correspondientes a diciembre de 2008 y a junio de 2009) son muy asimétricas, se emplea en el análisis el test de correlación de Spearman en lugar del de Pearson. Los coeficientes de correlación entre el número de enlaces y los datos de desempeño financiero se muestran en la Tabla 4.24. Todos los coeficientes de correlación son estadísticamente significativos salvo los correspondientes a la variable ROA (*return on assets*).

Table 4.24. Correlación entre el número de enlaces recibidos y los datos financieros.

Correlación rho de Spearman	Ejercicio	N	Número de enlaces (diciembre 2008)	Número de enlaces (junio 2009)
Activos totales	2008	39	0,58**	0,54**
	2007	39	0,44**	0,43**
Ingresos totales	2008	38	0,62**	0,60**
	2007	39	0,44**	0,40*
Resultado neto	2008	39	0,56**	0,52**
	2007	39	0,57**	0,55**
ROA (*Return on assets*)	2008	39	0,23	0,19
	2007	39	0,02	0,06
Número de empleados	2008	41	0,80**	0,80**
	2007	39	Sin datos	Sin datos
* Coeficientes de correlación significativos al nivel 0,05. ** Coeficientes de correlación significativos al nivel 0,01.				

Los resultados indican que el número de enlaces podría emplearse como un indicador de la posición y el desempeño financieros de los bancos. Sin embargo, no existe una relación significativa entre el número de enlaces y el ROA, una medida relativa del desempeño financiero. Esto se puede explicar a partir de las características de la variable "número de enlaces recibidos", obtenida a partir de la información ofrecida por el buscador Yahoo!, la cual incluye todos los enlaces que apuntan al sitio web de un determinado banco desde su creación. Esto significa que se trata de una variable que es acumulativa por su propia naturaleza. Lo mismo se puede decir de las variables financieras en valores absolutos, especialmente en el caso de la variable activos totales. Se espera que una empresa crezca con el tiempo y que, por tanto, sus activos se incrementen. Aunque las variables de

desempeño financiero, tales como los ingresos totales o el resultado neto se calculan para un periodo determinado de tiempo, por lo general se incrementan año tras año; por ejemplo, se espera que los ingresos totales sean mayores en el quinto año de vida de una empresa que en su primer año. Por el contrario, la variable ROA (*return on assets*) no depende tanto del tamaño de la empresa, por lo que parece lógico que no haya ninguna correlación significativa. Sería interesante analizar los cambios en el número de enlaces a lo largo de los años, con el fin de calcular también una medida relativa para esta variable.

Las correlaciones son relativamente estables en el tiempo, como sugiere la similitud entre los dos conjuntos de coeficientes de correlación. Los coeficientes de correlación más altos se producen entre los enlaces obtenidos en diciembre de 2008 y los datos financieros correspondientes al ejercicio 2008, esto es, en el caso en el que la información de enlaces en la Web coincide con el periodo al que pertenecen los datos financieros. Si comparamos los coeficientes de correlación entre el número de enlaces recibidos y los datos financieros para los ejercicios 2007 y 2008, observamos que la relación es más fuerte en el caso de los datos financieros del año 2008. La única excepción es el resultado neto, que presenta un coeficiente de correlación muy similar en ambos casos, aunque es ligeramente superior para los datos financieros del 2007. Esto indica que el ajuste entre el periodo de observación de los enlaces con el periodo de los datos financieros puede constituir un factor significativo de cara a obtener una posible predicción más ajustada de una variable basada en la otra.

Los coeficientes de correlación son consistentes con los coeficientes obtenidos en el trabajo incluido en el Apartado 4.1.2 referidos al sector bancario en los Estados Unidos (Tabla 4.25).

Tabla 4.25. Correlación entre el número de enlacs y los datos financieros para el sector bancario en los Estados Unidos (Apartado 4.1.2)

Nº enlaces recibidos	Activos totales	Pasivos totales	Ingresos totales	Resultado neto	ROA
Enero 2009	0.74**	0.73**	0.75**	0.63**	0.13
Mayo 2009	0.70**	0.69**	0.71*	0.62**	0.18
* Coeficientes de correlación significativos al nivel 0,05. ** Coeficientes de correlación significativos al nivel 0,01.					

Los coeficientes de correlación para el sector bancario en Estados Unidos son mayores que los referidos al sector bancario internacional. Esto se explica por las condiciones competitivas homogéneas que existen en los Estados Unidos frente a la diversidad de mercados en la que operan las empresas incluidas en este trabajo. Distintos países cuentan con condiciones económicas y financieras diversas, lo cual podría explicar las bajas correlaciones existentes cuando bancos de distintos países son analizados conjuntamente. El grado de empleo de Internet con fines comerciales en los distintos países puede representar también un factor relevante.

Es importante interpretar de forma apropiada las correlaciones observadas en el estudio. El que una correlación sea estadísticamente significativa no prueba que exista una relación de causalidad. El gran número de enlaces recibidos por el sitio web de una empresa no causa un mejor desempeño financiero, aunque si es cierto que una imagen positiva en la Web, como sugiere el que exista un gran número de enlaces, puede contribuir positivamente al éxito del banco. Una explicación de la correlación existente es que un banco que tiene un buen desempeño empresarial tendrá unas mejores variables financieras y también será capaz de mantener un perfil elevado en la Web atrayendo un número considerable de enlaces. En otras palabras, podríamos decir que el gran número de enlaces recibidos puede constituir un síntoma y no la causa del buen desempeño económico de un banco. Desde una perspectiva de minería de datos, el establecimiento de una relación causal no es indispensable para que las correlaciones observadas constituyan una información

muy útil. Desde esta lógica, si sabemos que dos variables están vinculadas, se puede intentar predecir una de ellas basada en la otra. Dado que la información sobre enlaces recibidos es accesible públicamente y puede ser recogida fácilmente, se podría intentar buscar tendencias en el desempeño financiero de un banco a largo plazo, basado en el número de enlaces recibidos en su sitio web. Este intento sería especialmente útil en situaciones en las que la información financiera disponible es escasa o directamente no es accesible. Otra posible aplicación sería la identificación de empresas cuya presencia en la Web no se corresponde con sus datos financieros, lo cual puede indicar ciertos riesgos vinculados a un menor desarrollo de sus activos intangibles vinculados a la presencia en la red. Al margen, se podrían llevar a cabo investigaciones adicionales con el fin de averiguar el porqué y cómo mejorar dicha situación.

Comparación entre el número de enlaces recibidos por bancos asiáticos frente al resto de bancos

Debido a las características particulares de la economía asiática y a las posiciones de los bancos chinos y japoneses puestas de relieve en el análisis de co-enlaces incluido en el siguiente epígrafe, se ha considerado conveniente examinar si existen diferencias significativas entre el número de enlaces que reciben los sitios web de los bancos asiáticos frente a los del resto del mundo. Para ello se ha llevado a cabo una prueba de Mann-Whitney. La prueba indica que el número de enlaces es significativamente diferente, con p=0,042 para los datos de diciembre de 2008 y p=0,024 para los datos de junio de 2009. Tal y como aparece en la Tabla 4.26, los sitios web de los bancos asiáticos atraen mayor número de enlaces que los bancos en otros lugares del mundo. Esto debe ser valorado teniendo en cuenta que el buscador empleado para la obtención de los datos es Yahoo! global, el cual es probable que sobreestime el número de enlaces a sitios web de Estados Unidos (Vaughan y Zhang, 2007). Ello indica que el mayor número de enlaces recibidos por los bancos asiáticos constituye una evidencia muy importante, ya que si se empleara la versión china del buscador el número sería aún mayor. Es interesante

destacar cómo la diferencia entre los enlaces de los bancos asiáticos y los otros bancos se ha incrementado entre diciembre de 2008 y junio de 2009. Se trata de un hecho que pone de relevancia el poder financiero de los bancos asiáticos, en particular de los bancos chinos, como ha quedado de manifiesto a lo largo de la reciente crisis financiera. El informe Mergent de la industria bancaria en la región de Asia Pacífico (Mergent, 2009a) subraya el buen rendimiento de los bancos chinos. Algunos de los factores que contribuyen a ello son: las reformas bancarias, el incremento en la inversión extranjera, un sistema de supervisión más estricto y un incremento en los niveles de competencia. En este momento, China ha emergido como un actor destacado en la escena internacional y sus bancos presentan una mayor fortaleza financiera en comparación con muchas otras empresas de Estados Unidos y Europa. La fortaleza de los bancos chinos compensa las debilidades del sector bancario japonés, que también se incluye dentro del grupo asiático, de cara a la realización del análisis.

Tabla 4.26. Comparación entre el número de enlaces recibidos por los bancos asiáticos y por resto de bancos

	Número de bancos	Mediana del nº de enlaces (diciembre 2008)	Mediana del nº de enlaces (junio 2009)
Bancos asiáticos	9	193.000	273.000
Otros bancos	41	40.600	35.900

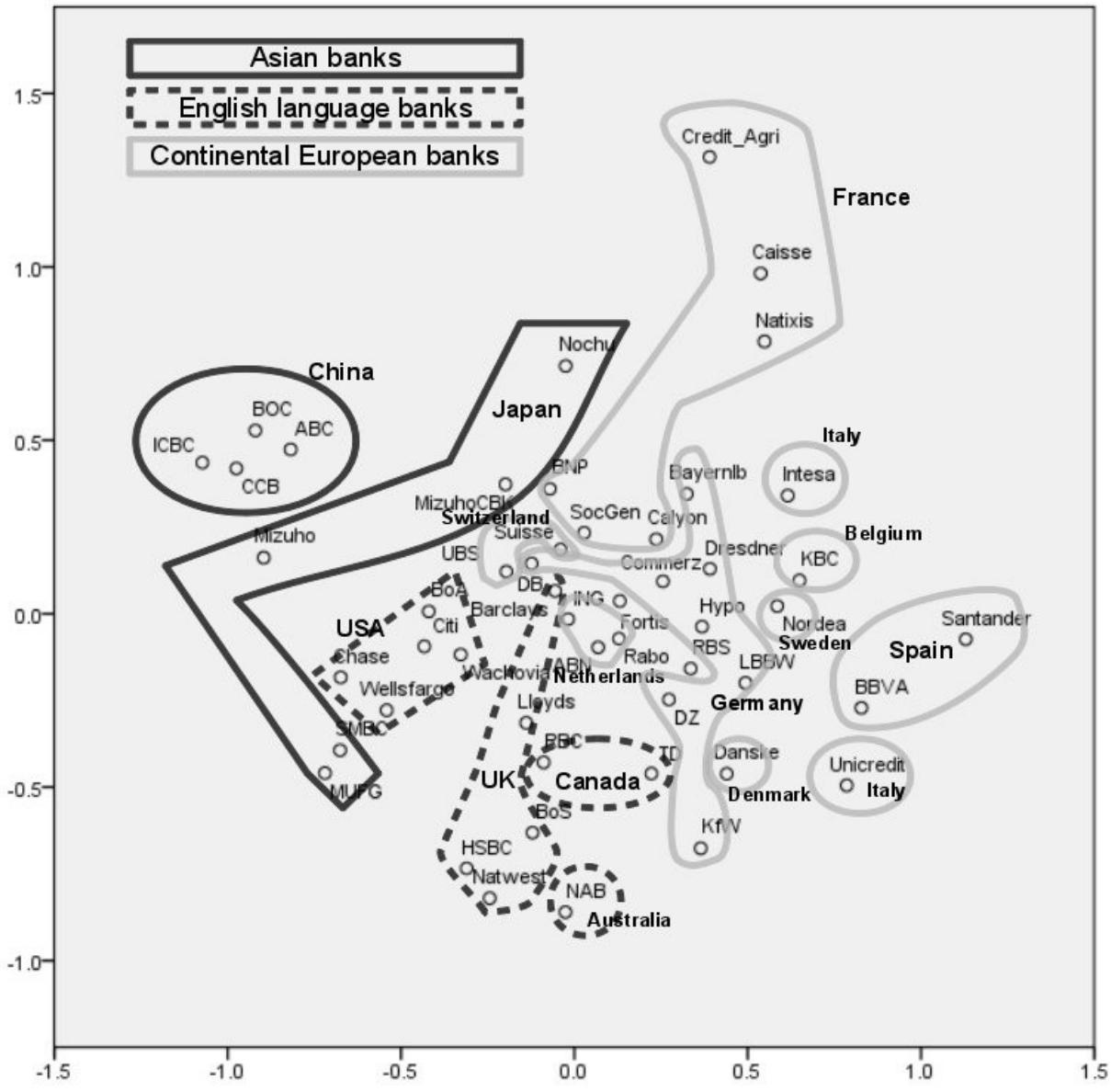
Análisis de co-enlaces

Los mapas MDS (Figuras 4.13 y 4.14), basados en el número de co-enlaces entre bancos, muestran sus posiciones competitivas relativas. El mapa basado en la información de diciembre de 2008 (Figura 4.13) permite observar una serie de *clusters* basados en criterios nacionales, regionales y lingüísticos. Tres son las

principales áreas identificadas: bancos asiáticos, bancos de países de lengua inglesa y bancos de la Europa continental (excluyendo, por tanto, al Reino Unido). Los bancos chinos y japoneses se encuentran agrupados juntos en una misma área del mapa. Los bancos chinos aparecen localizados en un extremo del mapa, aislados del resto de empresas. Esto indica que compiten principalmente en el mercado local. Son bancos que reciben muchos enlaces, pero de los cuales muy pocos son co-enlaces con otros bancos. Los bancos japoneses se encuentran en una posición intermedia, con MUFG y SMBC próximos a los bancos estadounidenses, Mizuho a los bancos chinos, y Nochu y MizuhoCBK a las entidades europeas.

Los bancos de países de lengua inglesa (se identifican como tales, Estados Unidos, Reino Unido, Canadá y Australia) aparecen agrupados, aunque también se observan dentro de ese *cluster* grupos nacionales. Los bancos de la Europa continental (incluyendo a aquellos de los siguientes países: Bélgica, Dinamarca, Francia, Alemania, Italia, Holanda, España, Suecia y Suiza) ocupan el resto del mapa. Los dos bancos suizos se encuentran una posición intermedia con respecto al resto de bancos en el estudio, lo cual refleja su posición competitiva en el mercado bancario internacional. Los bancos holandeses y algunos otros bancos (Fortis y Royal Bank of Scotland), que no aparecen agrupados junto a sus otros bancos nacionales, se concentran en el centro del mapa. El *cluster* de los bancos franceses se identifica claramente, al igual que el de los bancos alemanes, aunque éstos se encuentran algo más dispersos en el mapa.

Figura 4.13. Mapa obtenido a partir de los datos de diciembre de 2008

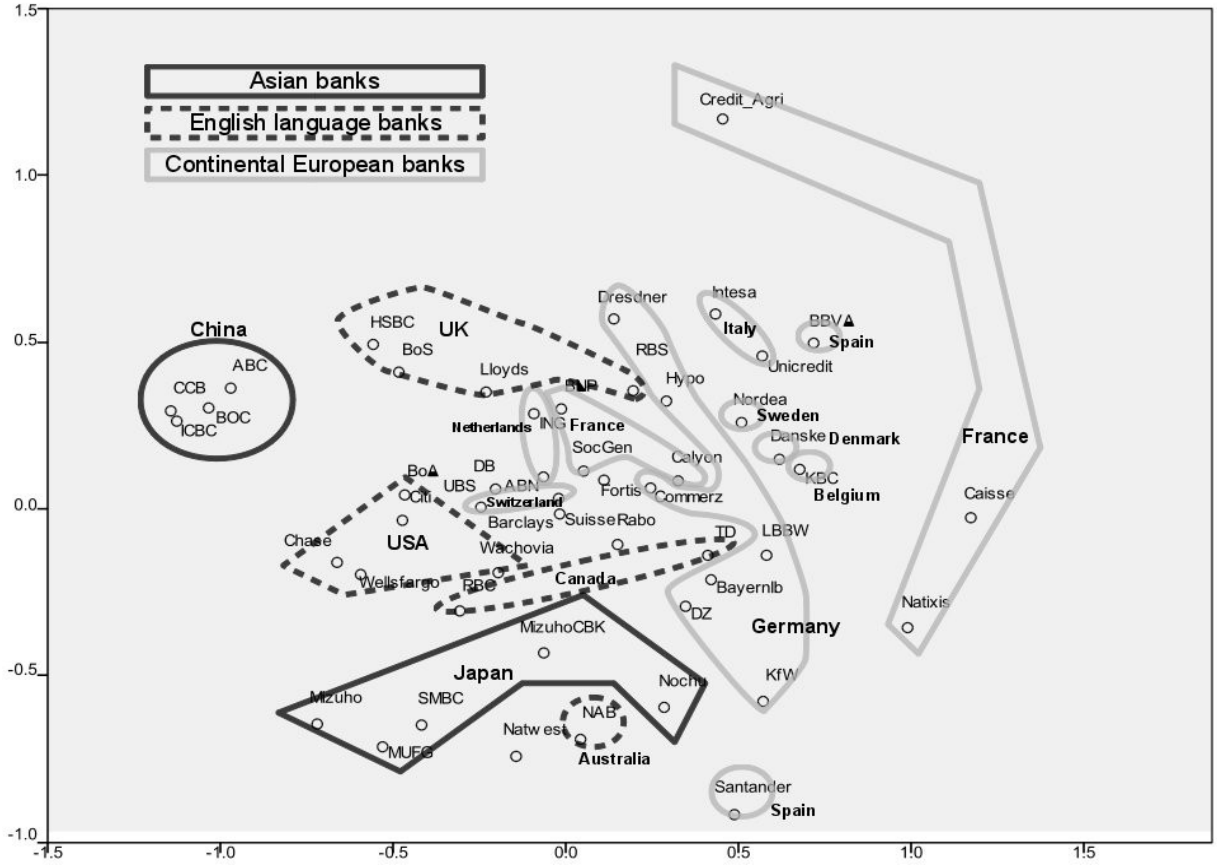


El mapa basado en los datos de junio de 2009 aparece en la Figura 4.14. Para facilitar la comparación con la Figura 4.13, los bancos en la Figura 4.14 se han analizado en función de los mismos criterios. A pesar de que el factor país se mantiene como predominante, hay algunos cambios significativos entre las Figuras 4.13 y la 4.14, que pueden estar vinculados a la evolución de la crisis económica y

financiera a lo largo del periodo de seis meses.

Cabe destacar cómo los bancos chinos y japoneses se han separado, situándose los bancos estadounidenses entre ellos. Los bancos chinos se han movido también más cerca de los bancos británicos. Aunque sería preciso un seguimiento más continuado y análisis adicionales para averiguar con certeza si estos cambios reflejan la percepción cambiante del mundo en relación con los bancos chinos como resultado de la crisis financiera, es razonable especular que son indicativos del hecho de que los principales bancos de la zona Asia-Pacífico, excepto los bancos japoneses, se mantienen fuertes y bien posicionados en el contexto internacional (Mergent, 2009a). Esto es especialmente cierto en el caso de los bancos chinos, los cuales se encuentran entre los mayores del mundo. Los bancos suizos y los principales bancos europeos (Deutsche Bank y Barclays) mantienen sus posiciones centrales. Esto refleja la situación de estabilidad de que estos bancos disfrutan en contexto global del sector financiero. Los bancos franceses aparecen ahora divididos en dos grupos, acentuando diferencias que ya se apuntaban en el mapa de la Figura 4.13.

Figura 4.14. Mapa MDS obtenido a partir de los datos de junio de 2009



4.3.1.3. Conclusiones

El presente trabajo alcanza los objetivos propuestos al combinar las dos técnicas de análisis webmétrico más empleadas, el análisis de impacto de enlaces y el análisis de co-enlaces, para desarrollar una aproximación global al sector bancario internacional. Se han detectado correlaciones significativas entre el número de enlaces recibidos y varias variables financieras, algunas de las cuáles no se habían empleado anteriormente. Los resultados son consistentes con los que se muestran en el trabajo incluido en el Apartado 4.1.1 para el caso del sector bancario en

Estados Unidos. Esto confirma que la información sobre la estructura de enlaces de la Web puede proporcionar información empresarial útil, incluso cuando empresas de diversos países son tenidas en cuenta. Una comparación del número de enlaces entre bancos asiáticos y otros bancos indica una posición de dominio de los primeros. Esto es consistente con el hecho de que los bancos asiáticos, especialmente los chinos, han soportado la crisis financiera relativamente mejor que el resto. Este hecho permite también poner de relevancia la utilidad de los datos de enlaces para obtener información empresarial. El análisis de escalamiento multidimensional basado en los co-enlaces indica que el sector bancario global se encuentra organizado parcialmente en mercados regionales a pesar de la existencia de importantes corporaciones fruto del proceso de internacionalización de las actividades financieras. Algunas regiones parecen estar más aisladas del resto, como es el caso de los bancos chinos, especialmente en el mapa correspondiente al primer periodo. Este aislamiento puede ser una de las razones por las que los bancos chinos no han sufrido tan seriamente la crisis financiera mundial. Los bancos suizos, junto con otras grandes entidades europeas, ocupan una posición central en el mapa.

Pueden existir algunas limitaciones en el análisis de enlaces y de co-enlaces derivadas del empleo de motores de búsqueda para la obtención de los datos. Yahoo! mantiene diferentes bases de datos en algunos países, por ejemplo, en China. Aunque esto no prueba que exista un problema en este sentido, ya que a pesar de haber empleado Yahoo! Global los resultados para los bancos chinos eran mayores, sí que podría generar potenciales problemas en otros trabajos. Por otra parte, cabe también destacar que el número de bancos en cada país no es homogéneo y que esto podría afectar a los resultados en tanto que la principal variable empleada en el análisis es el país de origen. La política de gestión del sitio web por parte de los bancos puede ser también un factor significativo, ya que cuestiones relativas al diseño o al empleo de dominios alternativos pueden alterar el resultado del recuento de enlaces.

Investigaciones futuras deberán centrarse en observar los cambios en los mapas MDS a través de la recogida y análisis de datos de forma periódica. También se

podría ampliar el alcance del estudio con el fin de incluir más bancos de un mayor número de países con el objeto de determinar si las conclusiones alcanzadas pueden generalizarse. Aunque la Webmetría es fundamentalmente una disciplina de carácter cuantitativo, los trabajos podrían complementarse con análisis cualitativos, llevando a cabo un análisis de contenidos de las páginas que enlazan a los sitios en estudio con el propósito de entender el fenómeno y las motivaciones que subyacen al hecho de crear un hipervínculo. Finalmente, sería útil consolidar la evidencia alcanzada en anteriores trabajos al objeto de desarrollar una metodología sistemática para obtener conocimiento empresarial a partir de técnicas webmétricas.

5. FINAL OVERVIEW AND DISCUSSION

> *«Si cada uno de nosotros extiende su atención de manera igual entre grupos de diferentes tamaños, desde lo personal a lo global, ayudamos a evitar esos extremos. Vínculo a vínculo construimos sendas de entendimiento a través de la red de la humanidad. Somos los hilos que cohesionan el mundo. Cuando hacemos esto, acabamos teniendo de manera natural unos cuantos sitios web muy demandados, y una continua disminución de la enorme cantidad de sitios con muy pocos visitantes. En otras palabras, por muy atractiva que pueda ser la igualdad entre pares, semejante estructura no es óptima por su uniformidad. No presta la suficiente atención a la coordinación global, y puede requerir muchos clics para ir desde el problema a la solución.»*
>
> *Tim Berners-Lee*
>
> Tejiendo la red *(1999:189)*

5.1. Background

During its 20 years of life, the Web has become a laboratory for social sciences (Section 2.3). Since its creation, scientists have struggled to study and analyze the diverse phenomena that takes place in the Web, which is characterized by the large amounts of data available and its continuous transformation and growth. The massive collection of information makes it possible to describe the Web as an

enormous unstructured and heterogeneous database that, despite its appearance, is not randomly built. Therefore, Web data can be exploited from different perspectives (based on content, structure and user behavior) in order to study the unique online phenomena or offline phenomena reflected in the Web.

The combination of science and the Internet is known by various terms, such as e-Science or e-Research (Section 2.4.1). e-Science refers to large scale science that is carried out through distributed global collaborations enabled by the Internet. Many e-Science initiatives have underlined the importance of distributed computational power and grid computing, although most of the time, especially when referred to Social Science, the collaboration between researchers through the Internet does not require this type of resources. For instance, the investigation based on information about the content and structure of the Web does not necessarily require large computational resources. Many databases are available online, for example, search engines could be used to collect data in order to carry out research on different aspects of the Web (Section 3.3.3). Search engines crawl large parts of the Web and provide information about the content of the Web pages and the links that form the structure of the Web. This data could be used to reveal patterns that otherwise would remain hidden from us. A potential source of information on the Web derives from visualizing hyperlink networks. The exploitation of this type of information has been done by applying data mining techniques to the Web. In this line, Webometrics has emerged as a new discipline that applies to the Web concepts and methods derived from bibliometrics and information science (see chapter 3) .

Webometrics is defined by Björneborn (2004; in Björneborn and Ingwersen, 2004: 1217) as "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches". However, the study of the Web differs significantly from the study of scientific production and academy. The differences could be understood through the analysis of some of the main features of the Web (Section 2.4.2). The characteristics of the Web could also help us to grasp the implications and limitations of the findings obtained in this type of research. For instance, the dynamic nature of the Web (Ke et al., 2005) implies that any research

based on Web data is always a snapshot of that particular moment and state of the Web, subject to the limitations of the tools used to gather the information (for example, search engines) and impossible to be reproduced in future times.

As mentioned, commercial search engines are one of the most important sources of Web data, fundamentally when we seek to analyze the entire Web. However there are some significant limitations derived from the use of commercial search engines (Section 3.3.3.5). Here we summarize some of those limitations that can be useful to highlight some of the specificities of the Web as a matter of research:

- Search engines do not index the entire Web (Lawrence and Giles, 1998; 1999b; Sherman and Price, 2001; Bar-Ilan, 2004; Thelwall, Vaughan and Björneborn, 2005). Sometimes this is not due to technical reasons but due to the design of the Web sites, which could ban Web crawlers to access some of their parts (Koster, 2009a; 2009b). Nevertheless, Vaughan and Zhang (2007) found the percent of coverage to be much higher.

- Ranking systems eliminate similar or identical pages in their results, in order to avoid providing useless information (Gomes and Smith, 2003; Thelwall, 2008a).

- Crawling and reporting algorithms are commercial secrets and, therefore the exact criteria used to rank the information are unknown (Thelwall, Vaughan and Björneborn, 2005).

- The total numbers of results offered by search engines are estimates as they use algorithms that prioritize response time rather than exhaustiveness (Björneborn and Ingwersen, 2001).

- Results can be subject to national or language biases (Vaughan and Thelwall, 2004).

- The results can fluctuate and change over the time (Bar-Ilan, 2000; Mettrop and Nieuwenhuysen, 2001). In addition, only a few number of pages are

accessible (usually just a maximum of 1,000), creating problems when random sampling is needed, for instance. Nevertheless, some research also indicates that search engines have become more stable (e.g. Thelwall, 2001a; 2001b; Vaughan and Thelwall, 2003; Vaughan, 2004a). Uyar (2009) has recent findings about accuracy of search engine hit count.

Even with such limitations, earlier studies in different matters have discovered patterns in linking networks and significant correlations between online and offline phenomenon validating the results through link classifications. Nevertheless, it is essential to keep in mind these limitations when analyzing and interpreting Web data.

This thesis focuses on the analysis of hyperlinks (Section 3.4), this is, in the analysis of the structure of the Web. A hyperlink can be defined as a reference or navigational element in a document to another section of the same document or to another document. Somehow, the links constitute the hidden structure of the Web connecting different sites and Web pages that would stay isolated unless the specific URL is known (Berners-Lee, 1999). They can be regarded as an endorsement of a target page, especially if the creator has placed that link because it points to a useful or relevant resource. This idea resembles clearly the Garfield's proposal (1979) to use bibliographic citations as the basis to rank scientific production, a technique that is still used to evaluate the quality of academic research (Section 3.1). The creation of hyperlinks is not an irrelevant phenomenon, but implies significant social repercussions (Turow and Lokman, 2008). Exploitation of hyperlinks is well illustrated by the functioning of commercial search engines (Batelle, 2006); for instance, Google’s search engine dominance (Section 3.3.3.4) is derived from the exploitation of *PageRank* system (Brin and Page, 1998).

The application of Webometric techniques to commercial Web sites is not as developed as in the academic field (Thelwall, Vaughan and Björneborn, 2005). Competitive Intelligence (Kahaner, 1996) is one of the areas where this research has been more successful and promising (Tan, Foo y Hui, 2002; Reid, 2003). We

believe that the application of Webometric techniques and other Internet research methodologies to companies could provide new resources for companies to generate and maintain competitive advantages (Porter, 1980; 1985). For example, a recent study (Choi and Varian, 2009), carried out by Google researchers, used data from the Google Trends service to anticipate consumer behavior in several industries.

The empirical research developed in this thesis was based on two approaches: link impact analysis and co-link analysis. On one hand, link impact analysis is based on the comparison of the number of Web pages or Web sites that are linked to a set of Web pages or Web sites under research. The purpose of this type of research is, according to Thelwall (2009: 28), "to evaluate whether a given website has a high link-based web impact compared to its peers". Inlink counts can also be an indirect gauge of other attributes of the organization represented by the Web site. For instance, this has been traditionally used, within the academic field, as a potential estimator of research performance (Smith and Thelwall, 2002). Concerning business Web sites, Vaughan (2004a; 2004b) and Vaughan and Wu (2004) found significant evidence of the positive relationships between the number of links received by a business Web site and financial variables. The studies included in the Sections 4.1.1 and 4.1.2 extend this research by finding evidence in different countries and industries. The study in Section 4.1.3 represents the first attempt to determine the explanatory variables of the number of inlinks received by a business Web site.

On the other hand, co-link analysis could be considered as a type of link relationship mapping technique (Thelwall, 2009). These are based on the link data that interconnect a set of Web sites in different ways in order to draw a diagram that illustrates the relationships between them. In particular, co-link analysis is based upon the number of Web pages that link at the same time two Web pages or sites belonging to the group of entities under study. Co-links are analogous to the bibliometric concept of co-citation (Small, 1973). Co-link analysis has also been demonstrated to be a useful tool to reveal the cognitive or intellectual structure of a particular field of study (Zuccala, 2006a). This method is particularly useful when Web sites interlink each other which happens very rarely. It is the case of

commercial Web sites that scarcely link the Web site of a competing company, especially when they are in the same industry (Vaughan, Gao and Kipp, 2006). The explanation to this could be that companies seem to avoid diverting Web traffic to competitors (Shaw, 2001). Moreover, as Vaughan (2006) points out, co-link data are more robust than inlink data as the former are less easily manipulated. Web co-link analysis for business information started by focusing on a single industry (Vaughan and You, 2006) or on a specific sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009). The studies included in the Sections 4.2.1 and 4.2.2 explored the use of co-link analysis for investigating companies belonging to different industries and to analyze the evolution of financial crisis in the American banks.

Finally, the research in Section 4.3.1 combines both methods to provide an overview of the international banking industry.

5.2. Research questions: findings and contributions

Several aspects of business Web sites were studied in this thesis, some results were expected while others were not. Webometric methods were used to analyze the linking patterns to business Web sites, by extending the evidence found by previous research. As far as we know, this is the first approach to Webometric research from the area of business studies. Therefore one of the main contributions of this thesis is to present and develop new methods of research on business issues by profiting from a interdisciplinary approach.

From a broad perspective, the goal of this research was to establish a transversal link, using the jargon presented in the paper, between business studies and information science studies in the field of Web research. Moreover, extending previous research on this area, we tried to improve the understanding of linking patterns on commercial Web sites. To achieve this goal, we worked in three main lines:

- to verify whether the correlations found between inlink counts and financial variables existed in different industries, regions and types of companies;
- to investigate whether co-link analysis could provide us with knowledge about economies in general; and,
- to combine previous methods to provide an overall approach to a single industry.

The goals of this thesis were summarized in five research questions that we tried to answer throughout different studies.

5.2.1. Research question 1: To what extent financial variables and business web presence, measured by inlink counts, are related?

The first research question was answered by studying the relationships between inlink counts and financial variables in different world regions and in multiple industries. We found that in general terms the inlink data collected correlated well with offline financial variables as shown by previous research. The correlations between the online and the offline data validated the conclusions that were already found. Previous studies (Vaughan, 2004a; 2004b; Vaughan and Wu, 2004) were limited to the Information Technology industry and to three countries: China, USA and Canada. It was necessary to extend the investigation:

- to a wider variety of industries; and,
- to other regions, especially Europe, one of the most important markets in the world.

In addition, some companies in the studies were private entities as we included all the companies that have financial data in the database we used. The majority of

European companies in the Amadeus database were private.

The study in Section 4.1.1 extended previous research by finding evidence to significant correlations in several industries in the United States. The study analyzed five different industries (commercial banks, construction of buildings, general merchandise store, utilities and mining), as well as the companies in the Dow Jones Industrial. The results revealed that there are significant correlations in industries that not only pertain to Information Technology, indicating that hyperlink data could be used as a meaningful variable in multiple industries, not exclusively to information-centered ones.

In Section 4.3.1 we analyzed the top 50 international banks. Spearman's test indicated that the majority of the correlation coefficients between inlinks and a set of financial variables (i.e. total assets total revenue, net income and earnings before tax) were significant at the .01 level. Only return on assets (ROA) was not found to be statistically significant. Correlation coefficients for U.S. banking industry are higher than those for the global banking industry. This is explained by the homogeneous competitive conditions that exist in the U.S. market compared to the heterogeneous markets where the banks included in this study operate. Different countries have different economical and financial conditions, which may explain the lower correlations when banks from many numerous countries are analyzed together. The extent of which the Internet is used for commercial purposes in different countries could also be a significant factor.

Finally, the study in Section 4.1.2 extends the research into the European context. Spain and the United Kingdom were selected as they represent two big economies in the European Union and their two languages are among the most commonly used on the Web. The study confirmed, in the European context, findings from the previous studies. It has also been found that the Web hyperlink data reflected some unique features of the EU economy. For example, the correlations between the inlink data and the financial data do not change significantly when data from the two countries are merged, reflecting that two countries share a common market. When comparing results from the current study with that from previous studies of other

countries such as the U.S., the study also found issues that need further exploration to gain a deeper understanding of the relationship between Web data and financial variables. The majority of studies so far indicate that only when companies belong to the same industry the correlations are significant. However, there are contradictory evidence as the study in Spain and United Kingdom shows significant correlations when different industries in the same country are put together.

From these studies included in the thesis we could say that the number of links received by a business Web site is strongly correlated with financial position measures (total assets) and, to a lower extent, with financial performance measures (total revenue). Profit (and loss) and some relative financial performance ratios, such as Return on Equity (ROE) and Return on Assets (ROA), are rarely found significant. The findings prove that inlink counts could be used as an indicator of the business financial position and financial performance measures in absolute terms. However, there is no evidence to this theory when we refer to relative financial measures, such as ROA. This could be explained by the nature of the variable "inlink" that represents the number of links pointing to a particular Web page or Web site since its creation. The functions offered by search engines, at this moment, do not allow us to retrieve inlinks that were created during a specific period of time. Somehow, this is similar to the nature of financial position variables that accumulates over time. Financial performance measures as revenue or total income also tend to increase over time based on previous performance. According to the evidence found, inlinks seem to reflect mainly the size of a company and to a lower extent, some measures of business performance.

Finally, these studies allow us to say that the significant correlations found in the previous studies are also verified in different industries, out from the Information Technology industry, and in Europe.

5.2.2. Research question 2: Which financial variables explain the inlink count received by a business Web site?

The second research question was answered by designing and testing an explanatory model of the number of inlinks based on financial variables of companies in Spain and the United Kingdom (Section 4.1.3). This is an attempt to explain the inlink variable in commercial Web sites using a multiple regression analysis. Vaughan and Thelwall (2005) applied a multiple regression analysis to explain the linking pattern in Canadian Universities. The model was tested in five different industries and some general conclusions were found to answer this research question:

- The variable "Total assets" (transformed using natural logarithm) is the most relevant variable to explain the number of inlinks received by the business Web sites. Only in the publishing industry this variable is not significant. We consider that this variable should consistently appear in a explanatory model of the number of inlinks received by business Web sites in most industries.

- The variable "Intangible assets" is not significant in any industry. There could be several explanations for this: inconsistencies in the way that companies report this information, difficulties by accounting standards to recognize and measure intangible assets and lack of relationship between the intangible assets recognized in the books and the intangible assets of the company related to the Web.

- "Turnover" is included as a significant variable for the Construction and the Publishing industries; whereas "Profit after tax" only in the Telecommunications industry. Also the country origin of the company (dummy variable) is significant in 3 out of 5 industries.

To conclude, "Total assets" (variable of financial position) and "Turnover" (variable of performance position) are the most relevant variables that explain the number of

inlinks received by commercial Web sites. Also, country origin seems to be a variable to be taken into account, although it is not clear in this study if this variable represents a geographic or a linguistic component, or a mix of both.

5.2.3. Research question 3: How are companies belonging to different industries related when observed through hyperlink structure in the Web?

The third research question was answered by studying several stock exchange indexes including companies belonging to different industries (Section 4.2.1). This research is based on co-link analysis. Web co-link analysis for business information started with a single industry (Vaughan and You, 2006) or on a detailed picture of the competitive landscape of a sector within an industry (Vaughan and You, 2008; 2009). The methodology developed in these papers has also been tested and verified in different countries and industries, e.g. China's chemical industry and electronics industry (Vaughan, Tang and Du, 2009). Parallel to these studies on commercial Websites, research on heterogeneous Web sites has been carried out to test the triple helix theory on the Web (Stuart and Thelwall, 2006; García-Santiago and Moya-Anegón, 2009). This theory analyzes the transference of knowledge between business, university and government.

This study extended co-link analysis to Web sites of heterogeneous companies belonging to five stock exchange indexes. Multidimensional scaling was used to map the relative positions of the companies in the indexes. We compared results from the five different stock exchange indexes that represent different economic, geographical and cultural backgrounds. There is one main variable that could determine a generalized pattern between economic activities, this is, the degree in which the activity is information centered. Industries whose business model is more information centered form distinct clusters located at a position that is distant from other companies located in more central position. Information centered industries, according to our interpretation, comprise mainly four groups: companies based in

the production of information contents (so called Media in the study), companies providing information infrastructures and services (IT), companies very intensive in applying e-commerce techniques (Leisure companies in the tourism industry) and Financial companies. These industries are in an ongoing process of business model transformation, e.g. the deep crisis in the media sector due to digitalization of information.

In addition to their location in the maps, Media and IT companies receive more inlinks than any other companies that belong to more traditional industries. This particular inlink pattern reflects a business model that is highly exposed to Internet. Also, individual maps show that the expectation of finding same industry companies grouped together is only partially fulfilled.

Euro Stoxx 50 is the only index in the study that is made of companies belonging to different countries. This fact allowed us to observe economic integration in the Eurozone. Results show that the cluster of companies in this map is better explained by country origin of the company rather than by industries. As a contrast, Dow Jones map is more clearly defined in terms of industries, underlining a deeper economic integration of the U.S. market.

This study reveals that Web co-link data constitute a readily available source of information to obtain new insights into the business environment. The study of heterogeneous companies belonging to different industries allows managers, financial analysts and other stakeholders to locate their company's activity in relation to others and to identify differing business models and to follow their evolution over time. If a company is trying to increase its presence online, inlink count constitutes a more direct measure of the success of the company’s effort in achieving this goal. In addition, co-link analysis could be used by managers to monitor the progresses of the company in relation to other companies in the same industry or in different industries by observing how the co-link MDS maps change. Therefore, co-link analysis could become a managerial tool useful for companies changing their IT strategies in the Web. Conclusions are mainly exploratory but relevant to answer this research question and to open a new direction in the Webometric research of

commercial and economic issues.

5.2.4. Research question 4: Can co-link analysis be used to investigate particular economic events?

The forth research question was answered by performing a co-link analysis to study the evolution of the 2009 financial crisis in the U.S. banks listed in the New York Stock Exchange. This research was inspired by a recent study (Vaughan and You, 2008) that proposed a method that combines page content with co-link data to achieve a more detailed picture of the competitive landscape of a sector within an industry. We apply this content assisted method to the banking industry in order to study the impact of the current financial and economic crisis. The keywords selected to refine the search were "crisis", "bailout" and "subprime", which are used frequently to describe the current worldwide financial problems. These keywords were intended to filter out pages that link the banks, whether they are inlinks or co-inlinks, for reasons other than the financial crisis. With this study (Section 4.2.2), we tried to find out if the inlink and keyword data could provide information on banking financial crisis. Therefore we expected that the data could be used to visualize clusters of banks with more distress. Preliminary results seemed to offer promising results in January 2009. However, a second round of data in August 2009 did not confirm these findings. Some of the reasons that can explain this situation could be the following:

- Co-link analysis could not be an appropriate method to achieve the intended goals. Not all the Web pages linking to a bank with financial distress link simultaneously to another bank in trouble. It seems that a direct link analysis could be a better approach.

- The variable used to measure the financial crisis (the money received by the federal government) is affected by a size effect because largest banks are the banks that logically received more money and therefore it is possible that

we are measuring the size of the entity instead of the degree of financial crisis.

Regarding this research question, further research needs to be done in other issues to evaluate the usefulness of this approach. This study can provide guidelines to avoid future problems.

5.2.5. Research question 5: How could different Webometric methods be combined to provide an overall approach to particular industries?

The fifth and final research question was answered by combining the link impact analysis and the co-link analysis within the same paper to analyze a single industry. This study (Section 4.3.1) analyzed the international banking industry through various Webometric techniques. First, we wanted to find out if there is any correlation between the number of Web hyperlinks that a bank attracts and the bank's financial variables. Second, we used co-link analysis to map these banks' global market positions. The study achieved its goal of using a combination of inlink and co-link analysis by providing an overview of the global banking industry. We found significant correlations between inlink counts and various financial variables. A comparison of the inlink count between Asian banks and other banks showed that Asian banks received more inlinks. This is consistent with the fact that Asian banks weathered the recent world financial crisis relatively better.

The multidimensional scaling maps based on co-links indicate that the global banking industry is partially organized in regional markets despite the existence of main global players due to the internationalization process of financial activities. Many of them seem to be more isolated, such as the Chinese banks. This is possibly one of the reasons why the Chinese banks did not suffer from the recent world financial crisis as much as other banks. However, data collected from the second time period revealed changing positions of the Chinese banks. They moved

closer to major players such as U.S. and U.K. banks. This reflected the changing perception of the global financial industry toward the Chinese banks as the result of their relative resilience in the world financial crisis.

Although more sophisticated and integrated approaches need to be developed to study companies from different Webometric perspectives, this study shows how different methods could provide complementary views on a particular economic issue.

5.3. Limitations

Although the research provided additional evidence on phenomena previously reported and suggested new lines for future research in business studies using Webometric techniques, there are some limitations in the present research.

Its dynamic nature makes the Web a challenging source of data and this has to be taken into account when interpreting any link based research results. Web research is always a snapshot of that particular time and situation on the Web and this evolves every second. This also makes it impossible to do a research based on linking patterns on the past and to reproduce the same results. Some of the limitations were underlined when we addressed the search engines limitations as a source of data for Webometric research. Therefore, it is also unclear to what extent the results could be generalized when discussing the Web.

The usage of search engines to collect hyperlink information is also affected by the characteristics of the search engine market, which is an oligopoly of three operators Google, Yahoo! and Bing (Microsoft). From a global perspective, they share the majority of the search engine market, although in specific areas of the Web there are other players, for instance, Technorati for searching blogs. Search engine industry is under a constant process of change and innovation. This issue has been treated by many papers in recent years (Lewandowski, Wahlig and Meyer-Bautor, 2006; Evans, 2007). In 2009 Yahoo! and Microsoft sign an agreement that could

affect the inlink search functions that so far Yahoo! is providing. This could imply major changes in the development of Webometric research based on search engines data.

Regarding co-link analysis, the studies could present some limitations derived from the interpretation of these results, which could differ depending on the criteria used on the analysis or in the knowledge of the researcher about a specific issue in the study.

5.4. Future research

Although this research answered some questions about commercial Web sites and their inlink patterns, it also creates new questions. The discovery of new evidence on the relationships between inlink counts to business Web sites and financial variables reinforces the evidence previously found (Vaughan, 2004a; 2004b; Vaughan and Wu, 2004). It is necessary to develop models to display this information for business purposes. For example, the elaboration of a model in line with the multiple regression analysis in Section 4.1.3 could be used to predict the number of inlinks a Web site is expected to receive based on financial variables and industrial affiliation. Then we could determine which companies are under-performing or over-performing in the Web. Inlink counts, as a measure of the business presence in the Web, could also be used to quantify intangible assets for the company related to the Web.

The use of co-link analysis to study heterogeneous companies in terms of industry provides some promising results, but more research is needed to test the usefulness of the method in order to identify industries that are evolving to business models more intensive in information. In addition, the use of new visualization techniques need to be explored to extract more information from the data.

The use of link impact analysis and co-link analysis have shown promising results for the analysis of business Web sites, however a more systematic approach need

to be develop to interpret the results and to combine the methodologies in a way that an overview of the industry is provided. New Webometric measures need also to be explored, such as outlinks or co-outlinks, in the study of commercial Web sites.

Another promising approach is to carry out longitudinal studies to analyze the evolution of industries, stock exchange indexes or single companies at one time. This approach was demonstrated in the studies in Sections 4.2.2 and 4.3.1.

Qualitative research is a necessary complement to quantitative research because it provides confirming evidence about the relevant nature of the links being analyzed (Vaughan, Gao and Kipp, 2006). This can also reveal the nature of the Web pages, especially to know if they belong to Web 2.0 tools or not. The strong development of the Web 2.0 (blogs, social networking sites, wikis, etc.) in the last five years has changed significantly the Web scenario and its impact can be measured by using a Webometric approach. One of the advantages of studying the Web 2.0 is the possibility of using alternative sources to collect web data, for instance, *delicious* (a social marking service), *Flickr* (a photography based community), *Youtube* (a video based community), *Technoraty* (a search engine specialized in blogs), etc. The development of the Semantic Web and the use of APIs can widen the research options.

Finally, another suggestion for future research is to apply link impact analysis and co-link analysis to investigate the competitive and cooperative relationships in other areas, such as the public sector or political parties.

Epílogo

"La esperanza en la vida procede de las interconexiones entre todas las personas del mundo. Creemos que si todos trabajamos por aquello en lo que creemos individualmente que es bueno, entonces como conjunto conseguiremos más fuerza, más comprensión, más armonía a medida que seguimos el viaje. No encontramos al individuo subyugado por el todo. No encontramos las necesidades del todo subyugadas por el creciente poder de un individuo. Pero podemos ver más entendimiento en las luchas entre esos extremos. No esperamos que el sistema llegue a ser perfecto. Pero nos sentimos cada vez mejor acerca de ello. El viaje nos parece cada vez más excitante, pero no esperamos que acabe.

¿Deberíamos entonces sentir que nos estamos volviendo más listos, que cada vez controlamos mejor la naturaleza, a medida que evolucionamos? En realidad, no. Sólo estamos mejor conectados, conectados en mejor forma. La experiencia de ver el despegue del Web gracias al esfuerzo fundamental de miles de personas me da la enorme esperanza de que si tenemos la voluntad individual suficiente, podemos hacer colectivamente de nuestro mundo lo que queramos."

Tim Berners-Lee

Tejiendo la red *(1999: 194-195)*

Internet y la Web han cambiado nuestra vida para siempre. No se trata de una cuestión principalmente tecnológica, sino de una transformación social y personal

tan profunda que difícilmente somos capaces de vislumbrar sus consecuencias y repercusiones. La juventud actual ha vivido la mayor parte de sus años de formación rodeados de medios digitales, ordenadores, teléfonos móviles, reproductores de música, cámaras fotográficas. Los niños y niñas que hoy acuden a una escuela de primaria disponen de herramientas digitales en el aula y en casa. No conocen un mundo en el que estas formas de mediación no existen. Llegan a casa y junto a los libros y libretas encienden su ordenador y se conectan con sus amigos a Tuenti o Facebook. Intercambian fotografías. Leen páginas web, tal vez un blog o un wiki; sin embargo, desconocen la teoría y los conceptos que elaboramos tras ellos. No han escuchando hablar de Web 2.0 ó de la Web Semántica, simplemente la viven. Para ellos no se trata de una nueva tecnología sino de una de las herramientas más útiles que conocen para la interacción social.

El mundo se ha hecho pequeño e inmenso a la vez, al ritmo de cada conexión que se establece, de cada hipervínculo que se crea. No existe Internet y el mundo físico como entes separados. Son sólo constructos para ayudarnos a entender dos contextos que aún percibimos como distintos. Cualquier intento de oponer lo virtual a lo real carece de fundamento. Toda realidad no puede ser sino una realidad virtual, un simulacro posmoderno de discursos fragmentarios. Investigar Internet y la Web desde las ciencias sociales no consiste en estudiar una tecnología, al igual que estudiar el funcionamiento organizacional de una empresa que fabrica componentes aeronáuticos no consiste en estudiar la física que encierran unos motores.

Nosotros, la gente, interactuamos socialmente para sentirnos parte de un grupo, para conocer a otras personas, para recibir y otorgar reconocimiento. Nos comunicamos para ser y somos en buena medida en tanto que nos comunicamos. Por ello, cada vez más, nos desarrollamos en redes, nos conectamos los unos a los otros. Investigar en Internet no es investigar en “nuevas tecnologías”. La ciencia social que presta atención a este contexto no debería pretender otro objetivo último que el de abordar las grandes preguntas genuinas del ser humano: qué quieren, qué sienten, qué hacen y qué experimentan los otros. Y a través de ello entender qué quiero, qué siento, qué hago y qué experimento yo mismo, cada uno de

nosotros. Nuestras preguntas marcan el destino de nuestro viaje. Nuestra curiosidad, nuestro amor, nuestro miedo, son los mismos desde el principio de la Humanidad. Volvemos ahora al inicio de esta tesis, a la compañía de Heródoto y de Kapuściński, un diálogo sin tiempo sobre el descubrimiento del otro, de lo otro y de nosotros mismos. El viaje puede transcurrir igual en la antigua Persia, en la moderna China, en la inabarcable África, en las agitadas redes de Internet. Podemos surcar el mar Mediterráneo en una nave griega o navegar la Web saltando de enlace a enlace.

Recordamos las palabras de Kapuściński: «Por eso, después de hacer este descubrimiento -otras culturas como espejo en que mirarnos para comprendernos mejor a nosotros mismos-, cada mañana a la salida del sol, incansablemente, Heródoto reanuda su viaje.»

A fin de cuentas cada punto es un Aleph,

y todos habitamos en la misma Babel.

BIBLIOGRAFÍA

Abraham, R.H. (1996). *Webometry: Measuring the complexity of the World Wide Web*. Santa Cruz, CA Visual Math Institute, University of California at Santa Cruz. Disponible en: http://www.ralph-abraham.org/vita/redwood/vienna.html (consultado el 7 de marzo de 2010).

ACLS (2006). Our cultural commonwealth: The Report of the American Council of Learned Societies Commision on cyberinfrastructure for the humanities & social sciences. American Council of Learned Societies (ACLS). Disponible en: http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf (consultado el 18 de noviembre de 2009).

Adamic, L.A. (1999). The small world web. *Lecture Notes in Computer Science*, 1696: 443-452.

Adamic, L.A. & Huberman, B.A. (2001). The Web‘s hidden order. *Communications of the ACM*, 44(9): 55-59.

Albert, R., Jeong, H. & Barabási, A.L. (1999). Diameter of the World Wide Web. *Nature*, 401: 130-131.

Almind, T.C. & Ingwersen, P. (1996). *Informetric analysis on the World Wide Web: A methodological approach to "internetometrics"* (CIS Report 2). Copenhagen, Denmark: Centre for Informetric Studies, Royal School of Library and Information Science.

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the World Wide Web Methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4): 404-426.

Alvarez, R.M. & Hall, T.E. (2006). Controlling democracy: the principal agent problems in election administration. *Policy Studies Journal*, 34(4): 491-510.

Aminpour, F., Kabiri, P., Otroj, Z. & Keshtkar, A.A. (2009). Webometric analysis of Iranian universities of medical sciences. *Scientometrics*, 80(1): 255–266.

Anderson, C. (2006). *The Long Tail. How Endless Choice is Creating Unlimited Demand*. London, UK: Random House Business Books.

Anderson, P. (2007). What is Web 2.0? Ideas, technologies and implications for Education. *Joint Information Systems Committee (JISC).* Disponible en: http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf (consultado el 18 de febrero de 2010).

Anstead, N. & Chadwick, A. (2009). Parties, election campaigning, and the internet. Towards a comparative institutional approach. En A. Chadwick & P.N. Howard (Eds.) *Routledge Handbook of Internet Politics* (pp. 56-71). New York: Routledge.

Antweiler, W. & Frank, M.Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, 59(3): 1259-1294.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology*, 1(1): 2-43.

Arrow, K.J. (1979). The economics of information. En M.L. Dertouzos & J. Moses (Eds.) *The computer age: A twenty-year view* (pp. 306-317). Cambridge, MA: MIT Press.

Arroyo, N. (2004). What is the Invisible Web? A Crawler Perspective. *Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods*, September 19, 2004, Brighton, UK. Disponible en: http://cybermetrics.wlv.ac.uk/AoIRASIST/arroyo.html (consultado el 9 de marzo de 2010).

Ashbaugh, H., Johnstone, K.M. & Warfield, T.D. (1999). Corporate reporting on the internet. *Accounting Horizons*, 13(3): 241-258.

Atkins Report (2003). Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Disponible en: http://www.nsf.gov/od/oci/reports/atkins.pdf (consultado el 18 de noviembre de 2009).

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley Longman Publishing Co.

Baeza-Yates, R., Castillo, C. & López, V. (2005). Characteristics of the Web of Spain. *Cybermetrics*, 9(1), paper 3. Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html (consultado el 7 de marzo de 2010).

Bailey, K.E., Bylinsky, J.H. & Shields, M.D. (1983). Effects of audit report wording changes on the perceived message. *Journal of Accounting Research*, 21: 355-70.

Baker, S. (2009). *They've got your number... Data, Digits and Destiny – how the Numerati are changing our lives*. London: Vintage.

Bar-Ilan, J. (2000). Evaluating the stability of the search tools Hotbot and Snap: a case study. *Online Information Review*, 24 (6): 439-449.

Bar-Ilan, J. (2002). How much information do search engines disclose on the links to a web page? A longitudinal case study of the 'cybermetrics' home page. *Journal of Information Science*, 28(6): 455–466.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. En B. Cronin (Ed.) *Annual review of information science and technology* (pp. 231–288). Medford, NJ: Information Today.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing and Management*, 41(4): 973-986.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century – A review. *Journal of Informetrics*, 2: 1-52.

Bar-Ilan, J. & Peritz, B.C. (1999). The life span of a specific topic on the Web: The case of 'Informetrics' a quantitative analysis. *Scientometrics*, 46(3): 371-382.

Bar-Ilan, J. & Peritz, B.C. (2009). The lifespan of "informetrics" on the Web: An eight year study (1998-2006). *Scientometrics*, 79(1): 7-25.

Barabási, A.L. (2002). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. London: Penguin Books limited.

Barabasi, A.L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439): 509-512.

Barabási, A.L., Albert, R. & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, 281: 69-77.

Barber, B.M. & Odean, T. (2001). The Internet and the Investor. *Journal of Economic Perspectives*, 15(1): 41-53.

Barjak, F. & Thelwall, M. (2008). A statistical analysis of the web presences of European life sciences research teams. *Journal of the American Society for Information Science and Technology*, 59(4): 638-643.

Barlow, J.P. (1996). Declaración de independencia del Ciberespacio. Disponible en: http://es.wikisource.org/wiki/Declaraci%C3%B3n_de_independencia_del_ciberespacio (consultado el 5 de febrero de 2010).

Batagelj, V. & Mrvar, A. (1998). Pajek – A Program for Large Network Analysis. *Connections*, 21(2): 47-57.

Batelle, J. (2006). *Buscar. Cómo Google y sus rivales han revolucionado los mercados y transformado nuestra cultura*. Barcelona: Ediciones Urano.

Bauer, C. & Scharl, A. (2000). Quantitative evaluation of Web site content and structure. *Internet Research: Electronic Networking Applications and Policy*, 10: 31-43.

Beaulieu, A. & Wouters, P. (2009). e-Research as Intervention. En N.W. Jankowski (Ed.) *e-Research. Transformation in Scholarly Practice* (pp. 54-69). New York, NY: Routledge.

Becker, S.A. (2004). E-government visual accessibility for older adult users. *Social Science Computer Review*, 22(1): 11-23.

Belkaoui, A. (1980). The interprofessional linguistic communication of accounting concepts: an experiment in sociolinguistics. *Journal of Accounting Research*, 18: 362-374.

Bell, D. (1973). *The coming of the post-industrial society: A venture in social forecasting*. New York: Basic Books.

Benoît, G. (2002). Data Mining. En B. Cronin (Ed.) *Annual Review of Information Science and Technology* (pp. 265-310). Medford, NJ: Information Today.

Bentley, T. (1997). Webs Are for Catching Flies. *Management Accounting*, 75 (4): 52.

Bergman, M.K. (2001). The deep web: Surfacing hidden value. *Journal of Electronic Publishing*, 7(1) August. Disponible en: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104 (consultado el 7 de marzo de 2010).

Berners-Lee, T. (1989/1990). Information management: a proposal. Disponible en: http://www.w3.org/History/1989/Proposal.html (consultado el 19 de febrero de 2010).

Berners-Lee, T. (1997). Realising the full potential of the Web. World Wide Web Consortium. Disponible en: http://www.w3.org/1998/02/Potential.html (consultado el 7 de marzo de 2010).

Berners-Lee, T. (1999). *Tejiendo la red*. Madrid: Siglo XXI.

Berners-Lee, T. & Cailliau, R. (1990). WorldWideWeb: proposal for a hypertext project. Disponible en: http://www.w3.org/Proposal.html (consultado el 18 de febrero de 2010).

Berners-Lee, T. & Hendler, J. (2001). Publishing on the Semantic Web. *Nature*, 410: 1023-1024.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5): 29-37.

Björneborn, L. (2001). Small-world linkage and co-linkage. *Proceedings of the 12th ACM conference on hypertext and hypermedia*. ACM: 33-134.

Björneborn, L. (2004). *Small-world link structures across an academic Web space: A library and information science approach*. Doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark. Disponible en: http://vip.db.dk/lb/phd/phd-thesis.pdf (consultado el 21 de julio de 2008).

Björneborn, L. & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1): 65–82.

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14): 1216–1227.

Boateng, R., Heeks, R., Molla, A. & Hinson, R. (2008). E-commerce and socio-economic development: conceptualizing the link. *Internet Research*, 18(5): 562-594.

Bonsón-Ponte, E., Escobar-Rodríguez, T. & Flores-Muñoz, F. (2006). Online transparency of the banking sector. *Online Information Review*, 30(6): 714-730.

Bonsón-Ponte, E., Escobar-Rodríguez, T. & Flores-Muñoz, F. (2009). Towards an ontology-based network for banking supervision. *Online Information Review*, 33(5): 943-955.

Bonsón, E., Cortijo, V. & Escobar, T. (2009). Towards the global adoption of XBRL using International Financial Reporting Standards (IFRS). *International Journal of Accounting Information Systems*, 10(1): 46–60

Bordons, M. & Gómez, I. (2000). Collaboration networks in science. En B. Cronin & H.B. Atkins (Eds.) *The web of knowledge – A festschrift in honor of Eugene Garfield* (pp. 197-213). Medford, N.J.: Information Today, Inc. & American Society for Information Science.

Borges, J.L. (1944/2005a). "La Biblioteca de Babel". En *Obras Completas I* (pp. 465-471). Barcelona: RBA.

Borges, J.L. (1944/2005b). "Tlön, Uqbar, Orbis Tertius". En *Obras Completas I* (pp. 431-443). Barcelona: RBA.

Borges, J.L. (1944/2005c). "Funes el Memorioso". En *Obras Completas I* (pp. 485-490). Barcelona: RBA.

Börner, K., Chen, C. & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37: 179-255.

Bossy, M.J. (1995). The last of the litter: "Netometrics". Disponible en: http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html (consultado el 7 de marzo de 2010).

Boulding, K.E. (1966). *The economics of knowledge and the knowledge of economics*. *American Economic Review*, 56: 1-13.

Boyack, K.W., Klavans, R. & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3): 351-374.

Brennan, N. & Kelly, S. (2000). Use of the Internet by Irish companies for investor relations purposes. *Irish Business and Administrative Research*, 21(2): 107–135.

Brin S. & Page L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30: 1-7.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the Web. *Journal of Computer Networks*, 33 (1-6): 309-320.

Brynjolfsson, E., Hu, Y. & Smith, M.D. (2003). Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Book Sellers. Management *Science*, 49(11), November 2003: 1580-1596.

Brynjolfsson, E., Hu, Y. & Smith, M.D. (2006). From Niches to Riches: Anatomy of the Long Tail. *MIT Sloan Management Review*, 47(4): 67-71.

Buchanan, M. (2002). *Nexus: Small worlds and the groundbreaking theory of networks*. New York: W.W. Norton & Company.

Burn, J. & Robins, G. (2003). Moving towards eGovernment: a case study of organizational change processes. *Logistics Information Management*, 16(1): 25-35.

Burt, E. & Taylor, J. (2001). When 'virtual' meets values: insights from the voluntary sector. *Information, Communication and Society*, 4(1): 54-73.

Burwell, H.P. (1999). *Online competitive intelligence: Increase your profits using cyberintelligence*. Tempe, AZ: Facts on Demand Press.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176 (July): 641-649. Disponible en: http://www.theatlantic.com/doc/194507/bush (consultado el 18 de febrero de 2010).

Caba Pérez, M.C., López Hernández, A.M. & Rodríguez Bolívar, M.P. (2005). Citizens' access to on-line governmental financial information: practices in the European Union countries. *Government Information Quarterly*, 22(2): 258-276.

Caba Pérez, M.C., Rodríguez Bolivar, M.P & López Hernández, A.M. (2008) E-Government process and incentives for online public financial information. *Online Information Review*, 32(3): 379-400.

Cahlik, T. (2000). Comparison of the maps of science. *Scientometrics*, 49(3): 373-387.

Cailliau, R. (1995). A little history of the World Wide Web: from 1945 to 1995. *World Wide Web Consortium*. Disponible en: http://www.w3.org/History.html (consultado el 18 de febrero de 2010).

Calero, C., Ruiz, J. & Piattini, M. (2005). Classifying web metrics using the web quality model. *Online Information Review*, 29(3): 227-248.

Capurro, R. & Hjørland, B. (2003). The Concept of Information. En B. Cronin (Ed.) *Annual Review of Information Science and Technology* (pp. 343-411). Medford, NJ: Information Today.

Card, S.K., Mackinlay, J.D. & Shneiderman, B. (Eds.) (1999). *Readings in information visualization - using vision to think*. San Francisco: Morgan

Kaufman Publishers.

Castells, M. (1989). *The informational city: Information technology, economic restructuring and the urban-regional process*. Oxford, UK: Blackwell.

Castells, M. (1996-1998). *The information age: Economy, society and culture*. Oxford: Blackwell.

Castells, M. (2001). *The Internet Galaxy. Reflections on the Internet, Business and Society*. Oxford: Oxford University Press.

Castells, M. (2009). *Communication Power*. Oxford: Oxford University Press.

Castells, M. & Kiselyova, E. (1995). *The Collapse of Soviet Communism: the View from the Information Society*. Berkeley, CA: University of California International Area Studies Book Series.

Chadwick, A. & Howard, P.N (2009). Introduction. New directions in internet politics research. En A. Chadwick & P.N. Howard (Eds.) *Routledge Handbook of Internet Politics* (pp. 1-9). New York: Routledge.

Chakrabarti, S., Joshi, M.M., Punera, K. & Pennock, D.M. (2002). The structure of broad topics on the Web. *Proceedings of the WWW 2002 Conference*. Disponible en: http://www2002.org/CDROM/refereed/338 (consultado el 7 de marzo de 2010).

Chakrabati, S., Dom, B., Kumar, R.S., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J.M. & Gibson, D. (1999). Hypersearching the Web. *Scientific American*, 280(6): 54-60.

Chau, M., Shiu, B., Chan, I. & Chen, H. (2007). Redips: Backlink search and analysis on the Web for business intelligence analysis. *Journal of the American Society for Information Science and Technology*, 58(3): 351-365.

Chen, C., Newman, J., Newman, R. & Rada, R. (1998). How did university departments interweave the web: A study of connectivity and underlying factors. *Interacting with Computers*, 10(4): 353-373.

Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R. & Card, S.K. (1998). Visualizing the evolution of Web ecologies. *Proceedings of Human Factors in Computing Systems* (pp. 400–407). ACM Press.

Choi, H. & Varian, H. (2009). Predicting the Present with Google Trends. Disponible en: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf (consultado el 10 de mayo de 2009).

Chu, H., He, S. & Thelwall, M. (2002). Library and information science schools in Canada and USA: A Webometric perspective. *Journal of Education for Library and Information Science,* 43: 110-125.

Chua, A. & Yang, C. (2008). The shift towards multi-disciplinarity in information science. *Journal of the American Society for Information Science and Technology,* 59(13): 2156-2170.

CNN Money (2009). Economy rescue: Adding up the dollars. Disponible en: http://money.cnn.com/news/specials/storysupplement/bailout_scorecard/index.html (consultado el 21 de enero de 2009).

Cooley, R., Mobasher, B. & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. *Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI '97)*. Disponible en: http://maya.cs.depaul.edu/~mobasher/papers/webminer-tai97.pdf (consultado el 8 de marzo de 2010).

Cornella, A. (2000). *Infonomia.com. La empresa es informacion*. Bilbao: Deusto.

Cortazar, J. (1963/1997). *Rayuela*. Madrid: Cátedra.

Craven, B.M. & Marston, C.L. (1999). Financial reporting on the Internet by leading UK companies. *European Accounting Review*, 8(2): 321–333.

Cuil (2010). Seach engine Cuil. Disponible en: http://www.cuil.com (consultado el 30 de marzo de 2009).

D'Alessio, D. (2007). A preliminary evaluation of the impact of unsolicited commercial email promoting stocks on the price of the stock. *New Media &*

Society, 9(6): 1011-1027.

Dans, E. (2007). La empresa y la "Web 2.0". *Harvard Deusto Marketing & Ventas*, 80 (Mayo/Junio): 2-9.

Das, S.R. & Sisk, J. (2005). Financial communities. *Journal of Portfolio Management*, 31(4): 112-123.

Davenport, E. & Cronin, B. (2000). The citation network as a prototype for representing trust in virtual environments. En B. Cronin & H. Atkins (Eds.) *The Web of knowledge—A Festschrift in honor of Eugene Garfield* (pp. 517–534). Medford, N.J.: Information Today, Inc. & American Society for Information Science.

Davis, I. (2005). Talis, Web 2.0 and All That. Disponible en: http://internetalchemy.org/2005/07/talis-web-20-and-all-that (consultado el 26 de marzo de 2010).

Davison, R.M., Wagner, C. & Ma, L.C.K. (2005). From government to e-government: a transition model. *Information Technology & People*, 18(3): 280-299.

Dean, J. & Henzinger, M.R. (1999). Finding related pages in the world wide web. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 31(11–16): 1467–1479.

Delclós, T. (2010, 15 de marzo). "El dominio ".com" cumple 25 años". *El País*. Disponible en: http://www.elpais.com/articulo/tecnologia/dominio/com/cumple/25/anos/elpeputec/20100315elpeputec_2/Tes (consultado el 16 de marzo de 2010).

Deller, D., Stubenrath, M. & Weber, C. (1999). A survey on the use of the Internet for investor relations in the USA, the UK and Germany. *European Accounting Review*, 8(2): 351–364.

Dhyani, D., Keong W. & Bhowmick, S.S. (2002). A survey of Web metrics. *ACM Computing Surveys*, 34(4): 469-503.

Dill, S., Kumar, S. R., McCurley, K., Rajagopalan, S., Sivakumar, D. & Tomkins, A. (2001). Self-similarity in the Web. *Proceedings of the 27th International Conference on Very Large Data Bases* (pp. 69-78).

Ding, Y., Fensel, D., Klein, M. & Omelayenko, B. (2002). The Semantic Web: Yet another hip? *Data & Knowledge Engineering*, 41: 205-227.

Dodge, M. & Kitchin, R. (2001). *Mapping cyberspace*. London: Routledge.

Dodge, M. & Kitchin, R. (2002). New cartographies to chart Cyberspace. *GeoInformatics*, 5(April/May): 38–41. Disponible en: http://www.casa.ucl.ac.uk/martin/geoinformatics_article.pdf (consultado el 8 de marzo de 2010).

Ebrahim, Z. & Irani, Z. (2005). E-government adoption: architecture and barriers. *Business Process Management Journal*, 11(5): 589-611.

Eco, U. (1962/1984). *Obra abierta*. Editorial Planeta-Agostini.

Ellison, G. & Ellison, S.F. (2005). Lessons About Markets from the Internet. *Journal of Economic Perspectives*, 19(2): 139-158.

Engelbart, D.C. (1961). Special considerations of the individual as a user, generator, and retriever of information. *American Documentation*, 12: 121-125.

Espadas, J., Calero, C. & Piattini, M. (2008). Web site visibility evaluation. *Journal of the American Society for Information Science and Technology*, 59(11): 1727-1742.

Essential Science Indicators (2005). Online help. Disponible en: http://esi.isiknowledge.com/help/h_datres.htm (consultado el 7 de marzo de 2010).

Etzioni, O. (1996). The World-Wide Web: Quagmire or gold mine? *Communications of the ACM*, 39(11): 65–68.

Etzkowitz, H. & Leydesdorff, L. (1995). The Triple Helix - University-Industry-Government Relations: A Laboratory for Knowledge Based Economic Development. *EASST Review*, 14, (14-19).

Etzkowitz, H. & Leydesdorff, L. (1997). *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relations*. London: Pinter.

Evans, M.P. (2007). Analysing Google rankings through search engine optimization data. *Internet Research*, 17(1): 21-37.

Faba-Pérez, C., Guerrero-Bote, V.P. & Moya-Anegón, F. (2003). Data mining in a closed Web environment. *Scientometrics*, 58(3): 623-640.

Faba Pérez, C., Guerrero Bote, V. & Moya Anegón, F. (2004a). *Fundamentos y técnicas cibermétricas*. Mérida: Junta de Extremadura. En http://www.juntaex.es/consejerias/economia-comercio-innovacion/dg-telecomunicaciones-sociedad-informacion/Publicaciones/common/tecnicascibermetricas.pdf (consultado el 21 de julio de 2008).

Faba-Pérez, C., Guerrero-Bote, V.P. & Moya-Anegón, F. (2004b). Methods for analysing web citations: a study of web-coupling in a closed environment. *Libri*, 54: 43-53.

Faloutsos, M., Faloutsos, P. & Faloutsos, C. (1999). On power–law relationships of the internet topology. *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, 29(4): 251-262. Disponible en: http://www.caip.rutgers.edu/TASSL/Fall05Reading/Vincent_Onpower-lawrelation.pdf (consultado el 7 de marzo de 2010).

Fisher, R., Laswad, F. & Oyelere, P. (2005). Determinants of voluntary internet financial reporting by local government authorities. *Journal of Accounting & Public Policy*: 24(2): 101-121.

Fleisher, C.S. & Blenkhorn, D.L. (Eds.) (2001). *Managing frontiers in competitive intelligence*. Wesport, CT: Quorum.

Foucault, M. (1972). *The Archaeology of Knowledge*. London: Tavistock Publications.

Friedman, T. (2005). *La tierra es plana. Breve historia del mundo globalizado del siglo XXI*. Barcelona: Ediciones Martínez Roca.

Fumero, A. & Roca, G. (2007). *Web 2.0*. Fundación Orange. Disponible en: http://www.fundacionorange.es/areas/25_publicaciones/publi_253_11.asp (consultado el 18 de febrero de 2010).

Gandía, J.L. & Archidona, M.C. (2008). Determinants of web site information by Spanish city councils. *Online Information Review*, 32(1): 35-57.

García-Borbolla, A., Larrán, M. & López, R. (2005). Empirical evidence concerning SMEs' corporate websites: explaining factors, strategies and reporting. *International Journal of Digital Accounting Research*, 5(10). Disponible en: http://www.uhu.es/ijdar/10.4192/1577-8517-v5_5.pdf (consultado el 30 de marzo de 2010).

García-Santiago, L. & Moya-Anegón, F. (2009). Using co-outlinks to mine heterogeneous networks. *Scientometrics ,* 79(3): 681-702.

Garfield, E. (1979). *Citation indexing: Its theory and applications in science, technology and the humanities*. New York: Wiley, Interscience.

Gascó, M. (2003). New technologies and institutional change in public administration. *Social Science Computer Review*, 21(1): 6-14.

Geerings, J., Bollen, L.H.H. & Hassink, H.F.D. (2003). Investor relations on the Internet: a survey of the Euronext zone. *European Accounting Review*, 12(3): 567-579.

Gershon, N., Eick, S.G. & Card, S. (1998). Information visualization. *Interactions*, 5(2): 9-15.

Gibson, W. (1989). *Neuromante*. Buenos Aires: Minotauro.

Gibson, R. & Ward, S. (2000). A proposed methodology for studying the function and effectiveness of party and candidate web sites. *Social Science Computer Review*, 18(3): 301-319.

Gibson, D., Kleinberg, J. & Raghavan, P. (1998a). Inferring web communities from link topology. Disponible en: http://www.cs.cornell.edu/home/kleinber/ht98.pdf (consultado el 21 de julio de 2008).

Gibson, D., Kleinberg, J. & Raghavan, P. (1998b). Structural analysis of the World Wide Web. Disponible en: http://www.w3.org/1998/11/05/wc-workshop/Papers/kleinber1.html (consultado el 21 de julio de 2008).

Gibson, R.K., Lusoli, W. & Ward, S.J. (2003) The internet and political campaigning: the new medium comes of age? *Representation*, 39(3), 166-80.

Gibson, R.K., Lusoli, W. & Ward, S.J. (2005). Online participation in the UK: testing a contextualised model of internet effects. *British Journal of Politics and International Relations*, 7(4): 561-583.

Gillies, J. & Calliau, R. (2000). *How the Web Was Born*. Oxford: Oxford University Press.

Giner, B. y Larrán, M. (2002). Use of the Internet for Corporate Reporting by Spanish Companies. *The International Journal of Digital Accounting Research*, 2(1): 53-83.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457: 1012-1014.

Girardin, L. (1995). *Cyberspace geography visualization: Mapping the World-Wide Web to help people find their way in cyberspace*. The Graduate Institute of International Studies, Geneva. Disponible en: http://www.girardin.org/luc/cgv/report/report.pdf (consultado el 8 de marzo de 2010).

Girardin, L. (1996). Mapping the virtual geography of the World-Wide Web. *Proceedings of the Fifth WWW Conference*. Disponible en: http://www.girardin.org/luc/cgv/www5/index.html (consultado el 8 de marzo de 2010).

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges—A methodological evaluation. *Scientometrics*, 57(1): 27–57.

Gomes, B. & Smith, B.T. (2003). Detecting query-specific duplicate documents. U.S. Patent 6,615,209. Disponible en: http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/ (consultado el 15 de noviembre de 2009).

Google (2006). Google SOAP Search API Reference. Disponible en: http://www.google.com/apis/reference.html#2_2 (consultado el 3 de junio de 2009).

Google (2008). 'We knew the web was big...'. Disponible en: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (consultado 14 de enero de 2009).

Google (2009). Links to your site. Disponible en: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=55281 (consultado el 3 de junio de 2009).

Google Librarian Central (2007). Back from WebSearch University. Disponible en: http://librariancentral.blogspot.com/2007/05/back-from-websearch-university.html (consultado el 9 de marzo de 2010).

Graef, J.L. (1997). Using the Internet for competitive intelligence: A survey report. *Competitive Intelligence Review,* 8: 41-47.

Granovetter, M.S. (1973). The strength of weak ties. *American Journal of Sociology*, 78: 1360-1380.

Gulli, A. & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. *Proceedings of the 14th international World Wide Web conference (WWW2005)*. Disponible en: http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf (consultado el 7 de marzo de 2010).

Gutiérrez-Nieto, B., Fuertes-Callén, Y. & Serrano-Cinca, C. (2008). Internet reporting in microfinance institutions. *Online Information Review*, 32(3): 415-436.

Ha, L. & James, E.L. (1998). "Interactivity Reexamined: a Baseline Analysis of Early Business Web Sites". *Journal of Broadcasting & Electronic Media*, 42(4): 457-474.

Halavais, A. (2000). National borders on the world wide web. *New Media and Society*, 2(1): 7-28.

Han, J. & Chang, K. (2002). Data mining for web intelligence. *IEEE Computer*, 35(11): 54-60.

Hanke, M. & Hauser, F. (2008). On the effects of stock spam e-mails. *Journal of Financial Markets*, 11: 57–83

Hanneman, R.A. & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside. Disponible en: http://faculty.ucr.edu/~hanneman (consultado el 7 de marzo de 2010).

Harries, G., Wilkinson, D., Price, L., Fairclough, R. & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5): 236-447.

Healy, P.M. & Palepu, K. (2001). Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1/3): 405-440.

Hedlin, P. (1999). The Internet as a vehicle for investor relations: the Swedish case. *European Accounting Review*, 8(2): 373–381.

Heimeriks, G. & van den Besselaar, P. (2006). Analyzing hyperlinks networks: The mean of hyperlink based indicators of knowledge production. *Cybermetrics*, 10(1). Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.html (consultado el 9 de marzo de 2010).

Heimeriks, G., Horlesberger, M. & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2): 391-413.

Heppes, D. & du Toit, A. (2009). Level of maturity of the competitive intelligence function. Case study of a retail bank in South Africa. *Aslib Proceedings: New Information Perspectives*, 61(1): 48-66.

Himanen, P. (2003). *La ética del hacker y el espíritu de la era de la información*. Barcelona: Destino.

Hine, C. (2005) Virtual Methods and the Sociology of Cyber-Social-Scientific Knowledge. En C. Hine (Ed.) *Virtual Methods. Issues in Social Research on the Internet* (pp. 1-13). Oxford: Berg.

Holmberg, K. (2009). *Webometric Network Analysis. Mapping Cooperation and Geopolitical Connections between Local Government Administration on the Web*. Doctoral dissertation. Åbo: Åbo Akademi University Press. Disponible en: http://kimholmberg.fi/phd/ThesisFinal.pdf (consultado el 30 de noviembre de 2009).

Hou, J. & Zhang, Y. (2003). Effectively finding relevant web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering*, 15(4): 1–12.

Houston, R.D. & Harmon, G. (2007). Vannevar Bush and Memex. En B. Cronin (Ed.) *Annual Review of Information Science and Technology* (pp. 55-92). Medford, NJ: Information Today.

Huberman, B.A. (2001). *The laws of the Web: Patterns in the ecology of information*. Cambridge, MA: MIT Press.

Huberman, B.A. & Adamic, L.A. (1999). Growth dynamics of the World-Wide Web. *Nature*, 401: 131.

Huberman, B.A., Pirolli, P. L.T., Pitkow, J.E. & Lukose, R.M. (1998). Strong regularities in World Wide Web surfing. *Science*, 280: 95-97.

Ingwersen, P. (1998). The calculation of Web Impact factors. *Journal of Documentation*, 54(2): 236-243.

Introna, L.D. & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The Information Society*, 16: 169-180.

Jacsó, P. (2005). Visualizing overlap and rank differences among web-wide search engines. *Online Information Review*, 29(5): 554-560.

Jankowski, N.W. (2009). The Contours and Chanllenges of e-Research. En N.W. Jankowski (Ed.) *e-Research. Transformation in Scholarly Practice* (pp. 3-31). New York, NY: Routledge.

Jankowski, N.W. & Van Selm, M. (2007). Research ethics in a virtual world: Guidelines and illustrations. En N. Carpentier, P. Pruulmann-Vengerfeldt, K. Nordenstreng, M. Hartmann, P. Vihalemm, B. Cammaerts & H. Nieminen (Eds.) *Media technologies and democracy in an enlarged Europe* (pp. 275-284). Tartu: Tartu University Press. Disponible en: http://www.researchingcommunication.eu/reco_book3.pdf (consultado el 18 de noviembre de 2009).

Jansen, B.J., Spink, A. & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society of Information Science and Technology*, 56(6): 559-570.

Jenkins, H. (2006). *Convergence Culture: Where Old and New Media Collide*. Cambridge, MA: MIT Press.

Jin, S. & Bestavros, A. (2002). *Small-world internet topologies*. Technical report. Disponible en: http://www.cs.bu.edu/techreports/pdf/2002-004-internet-topology-smallworld-sources.pdf (consultado el 7 de marzo de 2010).

Jin, Y., Matsuo, Y. & Ishizuka, M. (2009). Ranking companies on the Web using Social Network Mining. En I.H. Ting & H.J. Wu (Eds.) *Web Mining Application in E-commerce & E-services* (pp. 137-151). Berlín: Springer-Verlag.

Jones, M.J. & Xiao, J.Z. (2004). Financial reporting on the internet by 2010: a consensus view. Accounting Forum, 28(3): 237-263.

Kahaner, L. (1996). *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed. New York: Touchstone.

Kapuściński, R. (2004/2008). *Viajes con Heródoto*. Barcelona: Compactos Anagrama.

Ke, Y., Deng, L., Ng, W. & Lee, D.L. (2006). Web dynamics and their ramifications for the development of Web search engines. *Computer networks*, 50(10): 1430-1447.

Keen, A. (2007). *The Cult of the Amateur: How the Democratization of the Digital World is Assaulting Our Economy, Our Culture, and Our Values*. New York: Doubleday Currency.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1): 10-25.

Kessler, M.M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16(3): 223-233.

Kim, H.J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. *Journal of the American Society for Information Science*, 51(10): 887-899.

Kim, H., Park, H.W. & Thelwall, M. (2006). Comparing academic hyperlink structures with journal publishing in Korea. *Science Communication*, 27(4): 540-564.

King, J. (2006). Democracy in the information age. *Australian Journal of Public Administration*, 65(2): 16-32.

Klavans, R. & Boyack, K.W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2): 251-263.

Kochen, M. (Ed.) (1989). *The small world*. Norwood, N.J.: Ablex Publishing Corporation.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2): 162-180.

Koehler, W. (2002). Web page change and persistence – A four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2): 162-171.

Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence. *Information Research*, 9(2). Disponible en: http://informationr.net/ir/9-2/paper174.html (consultado el 30 de marzo de 2010).

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, 2(1): 1-15.

Koster, M. (2009a). A standard for robot exclusion. Disponible en: http://www.robotstxt.org/orig.html (consultado el 26 de febrero de 2010).

Koster, M. (2009b). About /robots.txt. Disponible en: http://www.robotstxt.org/robotstxt.html (consultado el 26 de febrero de 2010).

Kralisch, A. & Köppen, V. (2005). The impact of language on website use and user satisfaction. *Proceedings of the 13th European Conference on Information Systems: Information Systems in a Rapidly Changing Economy*. Disponible en: http://is2.lse.ac.uk/asp/aspecis/20050081.pdf (consultado el 26 de febrero de 2010).

Kralisch, A. & Mandl, T. (2006). Barriers of information access across languages on the Internet: network and language effects. *Proceedings of Hawaii International Conference on System Sciences (HICSS-39)*. Disponible en: http://www2.computer.org/portal/web/csdl/doi/10.1109/HICSS.2006.71 (consultado el 14 de abril de 2009).

Kretschmer, H. & Aguillo, I.F. (2004). Visibility of collaboration on the web. *Scientometrics*, 61(3): 405-426.

Kretschmer, H., Kretschmer, U. & Kretschmer, T. (2007). Reflection of co-authorship networks in the Web: Web hyperlinks versus Web visibility rates. *Scientometrics*, 70(2): 519-540.

Kristeva, J. (1980). *Desire in Language: A Semiotic Approach to Literature and Art.* New York, NY: Columbia University Press.

Kruskal, J.B. & M. Wish (1978). *Multidimensional Scaling*. London, UK: Sage Publications.

Kuhn, T.S. (1970/1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

La Porte, T.M., Demchak, C.C. & de Jong, M. (2002). Democracy and bureaucracy in the age of the web. Empirical findings and theoretical speculations. *Administration and Society*, 34(4): 411-446.

Landow, G.P. (1992). *Hypertext: the Convergence of Contemporary Critical Theory and Technology*. Baltimore and London: John Hopkins UP.

Larson, R.R. (1996). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of Cyberspace. *Information Today*: 71-78.

Lavoie, B.R. & O'Neill, E.T. (2001). How world wide is the Web? Trends in the internationalization of Web sites. *Journal of Library Administration*, 34(3-4): 407-419.

Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. *Science*, 280: 98-100. Disponible en: http://clgiles.ist.psu.edu/papers/Science-98.pdf (consultado el 3 de marzo de 2010).

Lawrence, S. & Giles, C.L. (1999a). Accessibility and distribution of information on the Web. *Nature*, 400(6740): 107-109.

Lawrence, S. & Giles, C.L. (1999b). Searching the Web: General and scientific information access. *IEEE Communications*, 37(1): 116-122.

Lessig, L. (2004). *Cultura libre*. Disponible en: http://cyber.law.harvard.edu/blogs/gems/ion/Culturalibre.pdf (consultado el 28 de febrero de 2008).

Levene, M. & Poulovassilis, A. (2001). Web dynamics. *Software Focus*, 2(2): 60–67.

Lévy, P. (2008). "El tiempo real, una velocidad trascendental", entrevista por Patrick Javault, *Antroposmoderno*. Disponible en: http://www.antroposmoderno.com/word/eltiemporeal.doc (consultado el 5 de febrero de 2010).

Lewandowski, D., Wahlig, H. & Meyer-Bautor, G. (2006). The freshness of web search engine databases. *Journal of Information Science*, 32(2): 131–148.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, 11: 291-320.

Leydesdorff, L. & Curran, M. (2000). Mapping university-industry-government relations on the Internet: The construction of indicators for a knowledge-based economy. *Cybermetrics*, 4(1), paper 2. Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.pdf (consultado el 14 de abril de 2009).

Leydesdorff, L. & Meyer, M. (2007). The scientometrics of a Triple Helix of university-industry-government relations (Introduction to the topical issue). *Scientometrics*, 70(2): 207-222.

Li, X. (2003). A review of the development and application of the Web impact factor. *Online Information Review*, 27(6): 407-417.

Li, X. (2005). *National and International University Departmental Web Site Interlinking: A Webometric Analysis*. Doctoral Dissertation. University of Wolverhampton.

Libby, R. (1979). Banker's and auditor's perceptions of the message communicated by the audit report. *Journal of Accounting Research*, 17: 99-122.

Lipsman, A. (2007). 61 billion searches conducted worldwide in August. ComScore: Measuring the Digital World, October 10, 2007. Disponible en: http://www.comscore.com/Press_Events/Press_Releases/2007/10/Worldwide_Searches_Reach_61_Billion (consultado el 9 de marzo de 2010).

Live Search (2007). We are flattered, but... Disponible en: http://www.bing.com/community/blogs/search/archive/2007/03/28/we-are-flattered-but.aspx (consultado el 3 de junio de 2009).

Llosa Sanz, A. (2006). De Tlön a Wikipedia: Borges, la World Wide Web, el libro-orbe y el conocimiento contenido del universo. *Divergencias. Revista de estudios lingüísticos y literarios*, 4(2): 13-20.

Lyman, P. & Varian, H.R. (2000). How big is the information explosion. *iMP, Information Impacts Magazine*, Nov. 2000. Disponible en: http://web.archive.org/web/20020220072409/http://cisp.org/imp/november_2000/11_00lyman.htm (consultado el 18 de febrero de 2010).

Lymer, A. (1999). The internet and the future of corporate reporting in Europe. *The European Accounting Review*, 8: 289-302.

Machill, M., Beiler, M. & Zenker, M. (2008). Search-engine research: a European-American overview and systematization of an interdisciplinary and international research field. *Media, Culture & Society*, 30(5): 591–608.

Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton, NJ: Princeton University Press.

Madnick, S. & Siegel, M. (2002). Seizing the opportunity: Exploiting Web aggregation. *MIS Quarterly Executive*, 1: 35-46.

Madria, S., Bhowmick, S.S., Ng, W.K. & Lim, E.P. (1999). Research issues in web data mining. *Lecture Notes in Computer Science*, 1676: 303–312. Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, Florence, Italy, Aug. 30 – Sept. 1, 1999.

Mar Molinero, C. & Serrano-Cinca, C. (2001). Bank failure: a multidimensional scaling approach. *European Journal of Finance*, 7(2): 165-83.

Mar Molinero, C., Bishop, H. & Turner, M. (2005). The Distress of Marks and Spencers PLC in 2001: A Multidimensional Scaling Analysis. *Cuadernos de Estudios Empresariales*, 15 : 107-126.

March, L. (2006). Virtual parties in a virtual world: Russian parties and the political internet. En S. Oates, D.M. Owen & R.K. Gibson (Eds.) *The Internet and Politics: citizens, voters, and activists*. London: Routledge.

Marshakova, V. (1973). System of document connections based on references. *Nauchno-Tekhnichescaya Informatisya*, Series II (6): 3-8.

Marston, C. (2003). Financial reporting on the Internet by leading Japanese companies. *Corporate Communications: An International Journal*, 8(1): 23-34.

Marston, C.L. & Polei, A. (2004). Corporate reporting on the internet by German companies. *International Journal of Accounting Information Systems*, 5: 285-311.

Matuszak, G. (2007). *Enterprise 2.0: Fad or Future?* KPMG International. Disponible en: http://www.kpmg.com/global/en/issuesandinsights/articlesPublications/Enterprise-fad-future/Pages/default.aspx (consultado el 28 de febrero de 2010).

Mayr, P. & Tosques, F. (2005). Webometrische Analysen mit Hilfe der Google Web APIs. *Information Wissenschaft und Praxis*, 56(1): 41–48.

McAfee, A. (2006) Enterprise 2.0: The Dawn of Emergent Collaboration. *MIT Sloan Management Review*, 47(3): 21-28.

McGonagle, J.J. & Vella, C.M. (1999). *The Internet age of competitive intelligence*. Wesport, CT: Quorum.

McLuhan, M. (1995). *El medio es el masaje. Un inventario de efectos*. Paidos.

Mergent (2009a). *Asia-Pacific - Banking Sectors (June 2009)*. Disponible en: http://webreports.mergent.com (consultado el 24 de diciembre de 2009).

Mergent (2009b). *Europe - Banking Sectors (September 2009)*. Disponible en: http://webreports.mergent.com (consultado el 24 de diciembre de 2009).

Mettrop, W. & Nieuwenhuysen, P. (2001). Internet search engines—Fluctuations in document accessibility. *Journal of Documentation*, 57(5): 623–651.

Milgram, S. (1967). The small world problem. *Psychology Today*, 22: 60-67.

Milman, B.L. (1994). Individual co-citation clusters as nuclei of complete and dynamic informetric models of scientific and technological areas. *Scientometrics*, 31(1): 45-57.

Montoya Juárez, J. (2008). *Realismos del simulacro: imagen, medios y tecnología en la narrativa del Río de la Plata*. Tesis doctoral. Universidad de Granada.

Moon, M.J. (2002). The evolution of e-government among municipalities: rhetoric or reality? *Public Administration Review*, 62(4): 424-433.

Moore, A. & Murray, B.H. (2000). Sizing the Internet [White paper]. *Cyveillance*.

Morrison, A.H. (2009). An impossible future: John Perry Barlow's 'Declaration of the Independence of Cyberspace'. *New Media & Society*, 11(1-2): 53–72.

Moya-Anegón, F., Jiménez-Contreras, E. & Moneda-Corrochano, M. (1998). Research fronts in library and information science in Spain (1985-1994). *Scientometrics*, 42(2): 229-246.

Mulkay, M., Gilbert, G.N. & Woolgar, S. (1975). Problem Areas and Research Networks in Science. *Sociology*, 9(2): 187-203.

Musgrove, P., Binns, R., Page-Kennedy, T. & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking Web spaces. *Scientometrics*, 58(3): 657-672.

Naughton, J. (1999). *A Brief History of the Future: The Origins of the Internet*. London: Weidenfeld & Nicolson.

Nel, D., Van Niekerk, R., Berthon, J. & Davies, T. (1999). Going with the flow: Web sites and customer involvement. *Internet Research: Electronic Applications and Policy*, 9, 109-116.

Nelson, T. (1967). Getting it out of our system. En G. Schecter (Ed.) *Information retrieval: a critical view* (pp. 191-210). Washington, D.C.: Thompson Book Company.

Nelson, T.H. (2000). *Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning and Deep Re-Use*. Disponible en: http://xanadu.com.au/ted/XUsurvey/xuDation.html (consultado el 4 de Febrero de 2010).

Nentwich, M. (2003). *Cyberscience: Research on the age on the Internet*. Vienna: Austrian Academy of Sciences Press.

Neophytou, E. & Mar Molinero, C. (2004). Predicting Corporate Failure in the UK: A Multidimensional Scaling Approach. *Journal of Business Finance and Accounting*, 31 (5-6): 677-710.

NeSC (n.d.). Defining e-science. National e-Science Centre. Disponible en: http://www.nesc.ac.uk/nesc/define.html (consultado el 8 de febrero de 2010).

Newman, M.E.J. (2000). Models of the small world. *Journal of Statistical Physics*, 101(3-4): 819-841

Nissembaum, H. & Price, M.E. (Eds.) (2004). *Academy & the Internet*. New York: Peter Lang.

Nordstrom, R.D. & Pinkerton, R.L. (1999). Taking advantage of Internet sources to build competitive intelligence system. *Competitive Intelligence Review,* 10: 54-61.

Norris (2003). Preaching to the converted? Pluralism, participation and party websites. *Party politics*, 9(1): 21-46.

Noyons, E. (2001). Bibliometric mapping of science in a science policy context. *Scientometrics*, 50(1): 83-98.

O'Reilly, T. (2005). What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. Disponible en: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (consultado 17 de enero de 2009).

ONTSI (2009). *La sociedad en red. Informe anual de la Sociedad de la Información en España (Edición 2009).* Disponible en: http://www.ontsi.red.es/informes-anuales/articles/id/3779/informe-anual-2008-edicion-2009.html (consultado el 20 de febrero de 2010).

Oppenheim, C., Morris, A., McKnight, C. & Lowley, S. (2000). The evaluation of WWW search engines. *Journal of Documentation*, 56(2): 190-211.

Ortega, J.L. & Aguillo, I.F. (2009). Mapping world-class universities on the web. *Information Processing and Management*, 45: 272-279.

Ortega, J.L., Aguillo, I. & Prieto, J.A. (2006). Longitudinal study of content and elements in the web environment. *Journal of Information Science*, 32(4): 344-351.

Ortega, J.L., Aguillo, I., Cothey, V. & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area – an exploration of visual web indicators. *Scientometrics*, 74(2): 295-308.

Otlet, P. (1996). *Tratado de Documentacion: el Libro sobre el Libro.* Universidad de Murcia.

Park, H.W., Barnett, G.A. & Nam, I. (2002). Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science*, 53 (7): 592–601.

Pérez, C.C., Bolívar, M.P.R. & Hernández, A.M.L. (2008). e-Government process and incentives for online public financial information. *Online Information Review*, 32(3): 379-400.

Petricek, V., Escher, T., Cox, I.J. & Margetts, H. (2006). The Web Structure of E-Government – Developing a Methodology for Quantitative Evaluation. The 15th International World Wide Web Conference, 2006, May 22-26, Edinburgh, Scotland. Disponible en: http://www.governmentontheweb.org/downloads/papers/WWW2006-Web_Structure_of_E_Government.pdf (consultado el 10 de marzo de 2010).

Pickerill, J. (2001). Weaving a green web: environmental protest and computer mediated communication in Britain. En F. Webster (Ed.) *Culture and Politics in the Information Age: a new politics?* London: Routledge.

Pitkow, J. (1998). Summary of WWW characterizations. *Computer Networks and ISDN Systems*, 30(1-7): 551-558.

Pitkow, J.E. (1997). *Characterizing World Wide Web ecologies*. Doctoral dissertation. Georgia Institute of Technology. Disponible en: http://www.webir.org/resources/phd/Pitkow_1997.pdf (consultado el 8 de marzo de 2010).

Pool, I. & Kochen, M. (1978/1979). 'Contacts and influence'. En M. Kochen (Ed.) *The small world* (pp. 3-51). Norwood, N.J.: Ablex Publishing Corporation, 1989. Nota: publicado originalmente en *Social Networks* (1978/1979), 1:5-51.

Porat, M.U. (1977). *The information economy: Definition and measurement*. Washington, DC: U.S. Department of Commerce, Office of Telecommunications.

Porter, M.E. (1980) *Competitive Strategy: Techniques for Analyzing Industries and Competitors.* New York, NY: Free Press.

Porter, M.E. (1985). *The Competitive Advantage: Creating and Sustaining Superior Performance*. New York, NY: Free Press.

Priego, J.L.O. (2003). A Vector Space Model as a methodological approach to the Triple Helix dimensionality: A comparative study of biology and biomedicine centres of two European national research councils from a webometric view. *Scientometrics*, 58(2): 429-443.

Prime, C., Bassecoulard, E. & Zitt, M. (2002). Co-citations and co-sitations—A cautionary view on an analogy. *Scientometrics*, 54 (2): 291–308.

Probst, G., Raub, S. & Romhard, K. (1999). *Managing knowledge*. London: Wiley.

Prytherch, R. (Ed.) (2000). Search engine. En *Harrod's librarians' glossary* (9th ed.). Aldershot, U.K.: Gower.

Rasmussen, E. (2003). Indexing and retrieval from the Web. *Annual Review of Information Science and Technology*, 37: 91-124.

Reid, E. (2003). Using Web link analysis to detect and analyze hidden Web communities. En D. Vriens (Ed.) *Information and communications technology for competitive* intelligence (pp. 57-84). Hilliard, OH: Ideal Group.

Ribes, X. (2007). Web 2.0: El valor de los metadatos y de la inteligencia colectiva. *Telos. Cuadernos de comunicación e innovación*, 73 (Octubre-Diciembre 2007). Disponible en: http://www.campusred.net/TELOS/articuloperspectiva.asp?idarticulo=2&rev=73 (consultado el 18 de febrero de 2010).

Rimbach, F., Dannenberg, M. & Bleimann, U. (2007). Page ranking and topic-sensitive page ranking: micro-changes and macro-changes. *Internet Research*, 17(1): 38-48.

Risvik, K.M. & Michelsen, R. (2002). Search engines and Web dynamics. *Computer Networks*, 39: 289-302.

Rodríguez Bolivar, M.P., Caba Pérez, C.M. & López Hernández, A.M. (2006). Cultural contexts and governmental digital reporting. *International Review of Administrative Sciences*, 72(2): 269-290.

Rodriguez Gairín, J.M. (1997). Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red. *Revista Española de Documentación Científica*, 20: 175-181.

Rousseau, R. (1997). Sitations: An exploratory study. *Cybermetrics*, I(1). Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html (consultado el 7 de marzo de 2010).

Roy, J. (2003). E-government. Introduction to a special issue. *Social Science Computer Review*, 21(1): 3-5.

Schwartz, C. (1998). Web Search Engines. *Journal of American Society for Information Science*, 49(11): 973-982.

Scott, J. & Marshall, G. (2009). *Oxford Dictionary of Sociology. Third edition revised*. Oxford: Oxford University Press.

Serrano Cinca, C., Mar Molinero, C. & Gallizo, J.L. (2002). A multivariate study of the EU economy via financial statements analysis. *The Statistician, Journal of the Royal Statistical Society*, 51: 1-20.

Shackleton, P., Fisher, J. & Dawson, L. (2006). E-government services in the local government context: an Australian case study. *Business Process Management Journal*, 12(1): 88-100.

Shaw, D. (2001). Playing the links: interactivity and stickiness in .com and "not.com" websites. *First Monday*, 6(3). Disponible en: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746 (consultado el 10 de mayo de 2009).

Shaw, J. & Parussini, G. (2009). China bank regulator: Global regulatory cooperation very fragile (2009). Disponible en: http://www.nasdaq.com/aspx/stock-market-news-story.aspx?storyid=200910051100dowjonesdjonline000261&title=china-bank-regulatorglobal-regulatory-cooperation-very-fragile (consultado el 15 de octubre de 2009).

Sherman, C. & Price, G. (2001). *The invisible Web*. Medford, NJ: Information Today, Inc.

Shestakov, D. (2008). *Search interfaces on the Web: Querying and characterizing*. Doctoral Dissertation. Turku Centre for Computer Science. Disponible en: https://oa.doria.fi/bitstream/handle/10024/38506/diss2008shestakov.pdf?sequence=3 (consultado el 7 de marzo de 2010).

Shi, Y. (2006). E-government web site accessibility in Australia and China. *Social Science Computer Review*, 24(3): 378-385.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1). Disponible en: http://www.sigir.org/forum/F99/Silverstein.pdf (consultado el 9 de marzo de 2010).

Simpson, R., Renear, A., Mylonas, E. & van Dam, A. (1996). 50 years after 'As we may think': The Brown/MIT Vannevar Bush Symposium. *Interactions of the ACM*, 3(2): 47-67. Disponible en: http://http://www.cs.brown.edu/memex/Bush_Symposium_Interact_2.html (consultado el 18 de febrero de 2010).

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4): 265-269.

Small, H. (1999). Visualising science through citation mapping. *Journal of the Society for Information Science*, 50(9): 799-812.

Small, H. & Garfield, E. (1985). The geography of science: disciplinary and national mappings. *Journal of Information Science*, 11: 147-159.

Small, H.G. & Sweeney, E. (1985). Clustering the Science Citation Index using co-citations: 2 - mapping science. *Scientometrics*, 8(5-6): 321-340.

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics*, 54: 363-380.

Smith, A.G. (2003). Think local, search global? Comparing search engines for searching geographically specific information. *Online Information Review*, 27(2): 102-109.

SNAC (2005). Social networks and cyberinfrastructure. Workshop 'The role of social network research in enabling cyberinfrastructure and the role of cyberinfrastructure in enabling social network research'. 3-5 November. Disponible en: http://www.ncsa.illinois.edu/Conferences/SNAC/ (consultado el 18 de noviembre de 2009).

Snyder, H. & Rosenbaum, H. (1999). Can search engines be used as tools for Web-link analysis? A critical view. *Journal of Documentation*, 55(4): 375-384.

Spink, A., Jansen, B.J., Kathuria, V. & Koshman, S. (2006). Overlap among major web search engines. *Internet Research*, 16(4): 419-426.

Spink, A., Wolfram, D., Jansen, M.B.J. & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52: 226-234.

Strandberg, K. (2006). *Parties, candidates and citizens online – Studies of politics on the Internet*. Doctoral Dissertation. Åbo Akademi University. Disponible en: https://oa.doria.fi/bitstream/handle/10024/4181/TMP.objres.78.pdf?sequence=1 (consultado el 31 de marzo de 2010).

Strogatz, S. (2003). *Sync: the emerging science of spontaneous order*. London: Penguin Books limited.

Stuart, D. (2008). *Web Manifestations of Knowledge-based Innovation Systems in the U.K.* Doctoral dissertation. University of Wolverhampton.

Stuart, D. & Thelwall, M. (2005). What can university-to-government web links reveal about university-government collaborations? En P. Ingwersen & B. Larsen (Eds.) *Proceedings of the 10th International Conference of the International Society of Scientometrics and Informetrics* (vol. 1, pp. 188-192). Stockholm, Sweden: Karolinska University press.

Stuart, D. & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry. *Research Evaluation*, 5(2): 97–106.

Stuart, D., Thelwall, M. & Harries, G. (2007). UK academic web links and collaboration – an exploratory study. *Journal of Information Science*, 33(2): 231-246.

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing and Management*, 28(1): 1–3.

Tan, B., Foo, S. & Hui, S.C. (2001). Web information monitoring: an analysis of Web page updates. *Online Information Review*, 25(1): 6-18.

Tan, B., Foo, S. & Hui, S.C. (2002). Web information monitoring for competitive intelligence. *Cybernetics and Systems*, 33(3): 225-251.

Tang, R. & Thelwall, M. (2002). Exploring the pattern of links between Chinese university Web sites. *Proceedings of the 65th ASIST Annual Meeting* (vol. 39, pp. 417-424).

Tang, R. & Thelwall, M. (2003). U.S. academic departmental web-site interlinking in the United States: Disciplinary differences. *Library & Information Science Research*, 25(4): 437-458.

Tang, R. & Thelwall, M. (2004). Patterns of national and international Web inlinks to US academic departments: An analysis of disciplinary variations. *Scientometrics*, 60(3): 475-485.

Tapscott, D. & Williams, A.D. (2007). *Wikinomics. La nueva economía de las multitudes inteligentes*. Barcelona: Paidós.

Terveen, L. & Hill, W. (1998). Evaluating emergent collaboration on the web. En S. Poltrock & J. Grudin (Eds.) *Proceedings of the 1998 ACM conference on computer supported cooperative work* (pp. 355-362). ACM Press.

Thelwall, M. (2000a). Commercial Web sites: Lost in cyberspace? *Internet Research*, 10(2): 150-159.

Thelwall, M. (2000b). Web impact factors and search engine coverage. *Journal of Documentation*, 56(2): 185–189.

Thelwall, M. (2000c) Who is using the .co.uk domain? Professional and media adoption of the Web. *International Journal of Information Management*, 20(6): 441-453.

Thelwall, M. (2001a). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13): 1157–1168.

Thelwall, M. (2001b). The responsiveness of search engine indexes. *Cybermetrics*, 5(1), paper 1. Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html (consultado el 26 de febrero de 2009).

Thelwall, M. (2001c). A Web Crawler Design for Data Mining. *Journal of Information Science*, 27(5): 319-325.

Thelwall, M. (2001d). Exploring the link structure of the Web with network diagrams. *Journal of Information Science*, 27(6): 393-401.

Thelwall, M. (2001e). Commercial web site links. *Internet Research*, 11(2): 114-124.

Thelwall, M. (2002a). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. *Journal of the American Society for Information Science and Technology*, 53(12): 995-1005.

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation*, 58(5): 563-574.

Thelwall, M. (2002c). The top 100 linked-to pages on UK university web sites—High inlink counts are not usually associated with quality scholarly content. *Journal of Information Science*, 28(6): 483–491.

Thelwall, M. (2003a). Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science*, 29(1): 1-10.

Thelwall, M. (2003b). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3), paper 151. Disponible en: http://informationr.net/ir/8-3/paper151.html (consultado el 10 de mayo de 2009).

Thelwall, M. (2004). *Link Analysis: An Information Science Approach.* San Diego: Academic Press.

Thelwall, M. (2008a). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology*, 59(1): 38-50.

Thelwall, M. (2008b). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11): 1702-1710.

Thelwall, M. (2008c). Bibliometrics to webometrics. *Journal of Information Science*, 34(4): 605-621.

Thelwall, M. (2009). *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*. Morgan & Claypool.

Thelwall, M. & Harries, G. (2004). Can personal web pages that link to universities yield information about the wider dissemination of research? *Journal of Information Science*, 30(3): 243-256.

Thelwall, M. & Price, L. (2003). Disciplinary difference in academic web presence: A statistical study of the UK. *Libri, 53(*4): 242-253.

Thelwall, M. & Smith, A. (2002). Interlinking between Asia-Pacific university web sites. *Scientometrics*, 55(3): 363-376.

Thelwall, M. & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy and denial of service. *Journal of the American Society for Information Science and Technology*, 57(13): 1771-1779.

Thelwall, M. & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, 58(1): 153-179.

Thelwall, M. & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26: 162-176.

Thelwall, M. & Wilkinson, D. (2003). Graph structure in three national academic webs—Power laws with anomalies. *Journal of the American Society for Information Science and Technology*, 54(8): 706–712.

Thelwall, M. & Wilkinson, D. (2004). Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 40(3): 515-526.

Thelwall, M. & Wilkinson, D. (2008). A generic lexical URL segmentation framework for counting links, colinks or URLs. *Library & Information Science Research*, 30: 515–526.

Thelwall, M. & Zuccala, A. (2008). A university-centred European Union link analysis. *Scientometrics*, 75(3): 407-420.

Thelwall, M., Harries, G. & Wilkinson, D. (2003). Why do web sites from different academic subjects interlink? *Journal of Information Science*, 29(6): 453-471.

Thelwall, M., Tang, R. & Price, L. (2003). Linguistic patterns of academic Web use in Western Europe. *Scientometrics*, 56(3): 417-432.

Thelwall, M., Vaughan, L. & Björneborn, L. (2005). Webometrics. En B. Cronin (Ed.) *Annual review of information science and technology* (pp. 81–135). Medford, NJ: Information Today.

Thelwall, M., Vaughan, L., Cothey, V., Li, X. & Smith, A.G. (2003). Which academic subjects have most online impact? A pilot study and a new classification process. *Online Information Review*, 27(5): 333-343.

Thomas, O. & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science*, 26(6): 421-428.

Thompson, F. (1998). Public economics and public administration. En J. Rabin, W.B. Hildreth & G.J. Miller (Eds.) *Handbook of Public Administration* (2nd ed.; pp. 995-1063). New York, NY: Marcel Dekker.

Torres-Vargas, G.A. (2005). World Brain and Mundaneum: the ideas of Wells and Otlet concerning universal access. *VINE: The journal of information and knowledge management systems*, 35(3): 156-165.

Torres, L., Pina, V. & Acerete, B. (2006). E-governance developments in European Union cities: reshaping government's relationship with citizens. *Governance: An International Journal of Policy, Administration, and Institutions*, 19(2): 277-302.

Torres, L., Pina, V. & Royo, S. (2005). E-government and the transformation of public administrations in the EU countries: beyond NPM or just a second wave of reforms? *Online Information Review*, 29(5): 531-553.

Tredinnick, L. (2007). Post-structuralism, hypertext, and the World Wide Web. *Aslib Proceedings: New Information Perspectives*, 59(2): 169-186.

Tumarkin, R. & Whitelaw, R.F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3): 41-51.

Turow, J. & Tsui, L. (Eds.) (2008). *The Hyperlinked Society: Questioning Connections in the Digital Age*. Ann Arbor: University of Michigan Press and University of Michigan Library. Disponible en: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=nmw;;idno=5680986.0001.001 (consultado el 18 de febrero de 2010).

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35(4): 469-480.

Van Raan, A.F.J. (2001). Bibliometrics and the internet: Some observations and expectations. *Scientometrics*, 50(1): 59-63.

Vattimo, G. (1990). *La sociedad transparente*. Barcelona: Paidós-I.C.E. de la Universidad Autónoma de Barcelona.

Vaughan, L. (2004a) Exploring website features for business information. *Scientometrics*, 61(3): 467–477.

Vaughan, L. (2004b). Web hyperlinks reflect business performance—A study of US and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28(1): 17-31.

Vaughan, L. (2006). Visualizing Linguistic and Cultural Differences Using Web Co-Link Data. *Journal of the American Society for Information Science and Technology*, 57(9): 1178-1193.

Vaughan, L. & Hysen, K. (2002). Relationship between links to journal web sites and Impact Factors. *ASLIB Proceedings: New Information Perspectives*, 54: 356-361.

Vaughan, L. & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14): 1313-1322.

Vaughan, L. & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the American Society for Information Science and Technology*, 56(10): 1075-1087.

Vaughan, L. & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54: 29-38.

Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*, 40(4): 693-707.

Vaughan, L. & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: the case of Canadian universities. *Information Processing and Management*, 41: 347-359.

Vaughan, L. & Wu, G.Z. (2004). Links to commercial websites as a source of business information. *Scientometrics*, 60(3): 487-496.

Vaughan, L. & You, J. (2005). Mining Web hyperlink data for business information: The case of telecommunications equipment companies. *Proceedings of the First IEEE International Conference on Signal-Image Technology and Internet-Based Systems* (pp. 190–195). Yaoundé, Cameroon, Nov. 27–Dec. 1, 2005.

Vaughan, L. & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3): 611–628.

Vaughan, L. & You, J. (2008). Content assisted web co-link analysis for competitive intelligence. *Scientometrics*, 77(3): 433-444.

Vaughan, L. & You, J. (2009). Keyword enhanced Web structure mining for business intelligence. *Lecture Notes in Computer Science*, 4879: 161–168.

Vaughan, L. & Zhang, Y. (2007). Equal representation by search engines? A comparison of Websites across countries and domains. *Journal of Computer-Mediated Communication*, 12(3). Disponible en: http://jcmc.indiana.edu:80/vol12/issue3/vaughan.html

Vaughan, L., Gao, Y. & Kipp, M. (2006). Why are hyperlinks to business Websites created? A content analysis. *Scientometrics*, 67(2): 291–300.

Vaughan, L., Kipp, M. & Gao, Y. (2007a). Are co-linked business web sites really related? A link classification study. *Online Information Review*, 31(4): 440–450.

Vaughan, L., Kipp, M. & Gao, Y. (2007b). Why are Websites co-linked? The case of Canadian universities. Scientometrics, 72(1): 81-92.

Vaughan, L., Tang, J. & Du, J. (2009). Examining the robustness of Web co-link analysis. *Online Information Review*, 33(5): 956-972.

Vaughan, L., Tang, J. & Du, J. (2010) Constructing business profiles based on keyword patterns on Web sites. *Journal of the American Society for Information Science and Technology.*

Vickery, B. (1997). Knowledge discovery from databases: an introductory review. *Journal of Documentation*, 53: 107-122.

von Krogh, G., Ichijo, K. & Nonaka, I. (2000). *Enabling knowledge creation. How to unlock the mystery of tacit knowledge and release the power of innovation*. New York: Oxford University Press.

Vreeland, R.C. (2000). Law libraries in hyperspace: A citation analysis of World Wide Web sites. *Law Library Journal*, 92(1): 9-25.

Waite, K. & Harrison, T. (2007). Internet archaeology: uncovering pension sector web site evolution. *Internet Research*, 17(2): 180-195.

Wang, H. & Rubin, B.L. (2004). Embedding e-finance in e-government: a new e-government framework. *Electronic Government, an International Journal*, 1(4): 362-373.

Wang, Y. & Kitsuregawa, M. (2001). Link based clustering of web search results. *The Lecture Notes in Computer Science*, 2118: 225–236. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=D646C5F6B3AFC85037F9F64FBD7788E4?doi=10.1.1.19.998&rep=rep1&type=pdf (consultado el 10 de mayo de 2009).

Ward, S. & Gibson, R. (2003). On-line and on message? Candidate websites in the 2001 general election. *British Journal of Politics and International Relations*, 5(2): 188-205.

Watts, D.J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105 (2): 493-527.

Watts, D.J. (2003). *Six degrees: The science of a connected age*. London: Vintage, Random House.

Watts, D.J. & Strogatz, S. (1998). Collective dynamics of "small-world" networks. *Nature*, 393: 440-442.

Weare, C. & Lin, W.-Y. (2000). Content analysis of the World Wide Web: opportunities and challenges. *Social Science Computer Review*, 18(3): 272-292.

Webster, F. (1995). *Theories of the information society*. London: Routledge.

Weinberg, A.M. (1961). Impact of large-scale science on the United States: Big science is here to stay, but we have yet to make the hard financial and educationa choices it imposes. *Science*, 134(3473): 161-164.

Weinberger, D. (2007). *Everything Is Miscellaneous: The Power of the New Digital Disorder*. New York: Times book.

Wells, H.G. (1937). "World Brain: The Idea of a Permanent World Encyclopaedia". Encyclopédie Française. Disponible en https://sherlock.ischool.berkeley.edu/wells/world_brain.html (consultado el 17 de enero de 2010).

Wessels, B. & Craglia, M. (2009). Situated Innovations in e-Social Science. En N.W. Jankowski (Ed.) *e-Research. Transformation in Scholarly Practice* (pp. 291-309). New York, NY: Routledge.

White, H.D. (1990). Author co-citation analysis: Overview and defence. En C.L. Borgman (Ed.) *Scholarly communication and bibliometrics* (pp. 84–106). Newbury Park, CA: Sage Publications.

White, H.D., Wellman, B. & Nazer, N. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, 55(2): 111–126.

White, G.K. & Manning, B.J. (1998). Commercial WWW site appeal: How does it affect online food and drink consumers' purchasing behavior? *Internet Research: Electronic Networking Applications and Policy*, 8: 32-38.

Wildman, P. (1998). From the monophonic university to the polyphonic multiversities. *Futures*, 30(7): 625-633.

Wilkinson, D., Harries, G., Thelwall, M. & Price, E. (2003). Motivations for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication. *Journal of Information Science*, 29(1): 59-66.

Wolfram, D., Spink, A., Jansen, B.J. & Saracevic, T. (2001). Vox populi: The public searching of the Web. *Journal of the American Society for Information Science and Technology*, 52: 1073-1074.

Wouters, P. (1996). Cyberscience. *Kennis en Methode*, 20(2): 155-186.

Yan, E. & Zhu, Q. (2008). Hyperlink analysis for governmental websites of Chinese provincial capitals. *Scientometrics*, 76(2): 315-326.

Yao, Y.Y., Zhong, N., Liu, J. & Ohsuga, S. (2001). Web intelligence (WI): Research challenges and trends in the new information age. *Lecture Notes in Artificial Intelligence*, 2198: 1–17.

Zanasi, A. (1998). Competitive intelligence through data mining public sources. Competitive *Intelligence Review*, 9(1): 44-54.

Zhang, Y., Jansen, B.J. & Spink, A. (2009). Time series analysis of a Web search engine transaction log. *Information Processing and Management*, 45: 230-245.

Zizek, S. (1996). From Virtual Reality to the Virtualization of Reality". En T. Druckrey (Comp.) *Electronic Culture* (pp. 290-295). Ontario: Aperture Foundation Books.

Zook, M. (2005). The Geographies of the Internet. En B. Cronin (Ed.) *Annual Review of Information Science and Technology* (pp. 53-78). Medford, NJ: Information Today.

Zuccala, A. (2006a). Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...? *Journal of the American Society for Information Science and Technology*, 57(11): 1487-1502.

Zuccala, A. (2006b). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, 57 (2): 152–168.

ANEXOS

Anexo I. Artículo "World Brain: The Idea of a Permanent World Encyclopaedia" por H.G. Wells (1937)

It is probable that the idea of an encyclopaedia may undergo very considerable extension and elaboration in the near future. Its full possibilities have still to be realized. The encyclopaedias of the past have sufficed for the needs of a cultivated minority. They were written "for gentlemen by gentlemen" in a world wherein universal education was unthought of, and where the institutions of modern democracy with universal suffrage, so necessary in many respects, so difficult and dangerous in their working, had still to appear. Throughout the nineteenth century encyclopaedias followed the eighteenth-century scale and pattern, in spite both of a gigantic increase in recorded knowledge and of a still more gigantic growth in the numbers of human beings requiring accurate and easily accessible information. At first this disproportion was scarcely noted, and its consequences not at all. But many people now are coming to recognize that our contemporary encyclopaedias are still in the coach-and-horses phase of development, rather than in the phase of the automobile and the aeroplane. Encyclopaedic enterprise has not kept pace with material progress. These observers realize that modern facilities of transport, radio, photographic reproduction and so forth are rendering practicable a much more fully succinct and accessible assembly of fact and ideas than was ever possible before.

Concurrently with these realizations there is a growing discontent with the part played by the universities, schools and libraries in the intellectual life of mankind. Universities multiply, schools of every grade and type increase, but they do not enlarge their scope to anything like the urgent demands of this troubled and dangerous age. They do not perform the task nor exercise the authority that might reasonably be attributed to the thought and knowledge organization of the world. It is not, as it should be, a case of larger and more powerful universities co-operating more and more intimately, but of many more universities of the old type, mostly ill-endowed and uncertainly endowed, keeping at the old educational level.

Both the assembling and the distribution of knowledge in the world at present are extremely ineffective, and thinkers of the forward-looking type whose ideas we are

now considering, are beginning to realize that the most hopeful line for the development of our racial intelligence lies rather in the direction of creating a new world organ for the collection, indexing, summarizing and release of knowledge, than in any further tinkering with the highly conservative and resistant university system, local, national and traditional in texture, which already exists. These innovators, who may be dreamers today, but who hope to become very active organizers tomorrow, project a unified, if not a centralized, world organ to "pull the mind of the world together", which will be not so much a rival to the universities, as a supplementary and co-ordinating addition to their educational activities - on a planetary scale.

The phrase "Permanent World Encyclopaedia" conveys the gist of these ideas. As the core of such an institution would be a world synthesis of bibliography and documentation with the indexed archives of the world. A great number of workers would be engaged perpetually in perfecting this index of human knowledge and keeping it up to date. Concurrently, the resources of micro-photography, as yet only in their infancy, will be creating a concentrated visual record.

Few people as yet, outside the world of expert librarians and museum curators and so forth, know how manageable well-ordered facts can be made, however multitudinous, and how swiftly and completely even the rarest visions and the most recondite matters can be recalled, once they have been put in place in a well-ordered scheme of reference and reproduction. The American microfilm experts, even now, are making facsimiles of the rarest books, manuscripts, pictures and specimens, which can then be made easily accessible upon the library srceen. By means of the microfilm, the rarest and most intricate documents and articles can be studied now at first hand, simultaneously in a score of projection rooms. There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind. And not simply an index; the direct reproduction of the thing itself can be summoned to any properly prepared spot. A microfilm, coloured where necessary, occupying an inch or so of space and weighing little more than a letter, can be duplicated from the records and sent anywhere, and thrown enlarged upon

the screen so that the student may study it in every detail.

This in itself is a fact of tremendous significance. It foreshadows a real intellectual unification of our race. The whole human memory can be, and probably in a short time will be, made accessible to every individual. And what is also of very great importance in this uncertain world where destruction becomes continually more frequent and unpredictable, is this, that photography affords now every facility for multiplying duplicates of this - which we may call? - this new all-human cerebrum. It need not be concentrated in any one single place. It need not be vulnerable as a human head or a human heart is vulnerable. It can be reproduced exactly and fully, in Peru, China, Iceland, Central Africa, or wherever else seems to afford an insurance against danger and interruption. It can have at once, the concentration of a craniate animal and the diffused vitality of an amoeba.

This is no remote dream, no fantasy. It is a plain statement of a contemporary state of affairs. It is on the level of practicable fact. It is a matter of such manifest importance and desirability for science, for the practical needs of mankind, for general education and the like, that it is difficult not to believe that in quite the near future, this Permanent World Encyclopaedia, so compact in its material form and so gigantic in its scope and possible influence, will not come into existence.

Its uses will be multiple and many of them will be fairly obvious. Special sections of it, historical, technical, scientific, artistic, e.g. will easily be reproduced for specific professional use. Based upon it, a series of summaries of greater or less fullness and simplicity, for the homes and studies of ordinary people, for the college and the school, can be continually issued and revised. In the hands of competent editors, educational directors and teachers, these condensations and abstracts incorporated in the world educational system, will supply the humanity of the days before us, with a common understanding and the conception of a common purpose and of a commonweal such as now we hardly dare dream of. And its creation is a way to world peace that can be followed without any very grave risk of collision with the warring political forces and the vested institutional interests of today. Quietly and sanely this new encyclopaedia will, not so much overcome these archaic discords,

as deprive them, steadily but imperceptibly, of their present reality. A common ideology based on this Permanent World Encyclopaedia is a possible means, to some it seems the only means, of dissolving human conflict into unity.

This concisely is the sober, practical but essentially colossal objective of those who are seeking to synthesize human mentality today, through this natural and reasonable development of encyclopaedism into a Permanent World Encyclopaedia.

Anexo II. Evaluación de los supuestos del modelo de regresión múltiple (Apartado 4.1.3)

Sector: *Construcción*

Histograma de residuos estandarizados

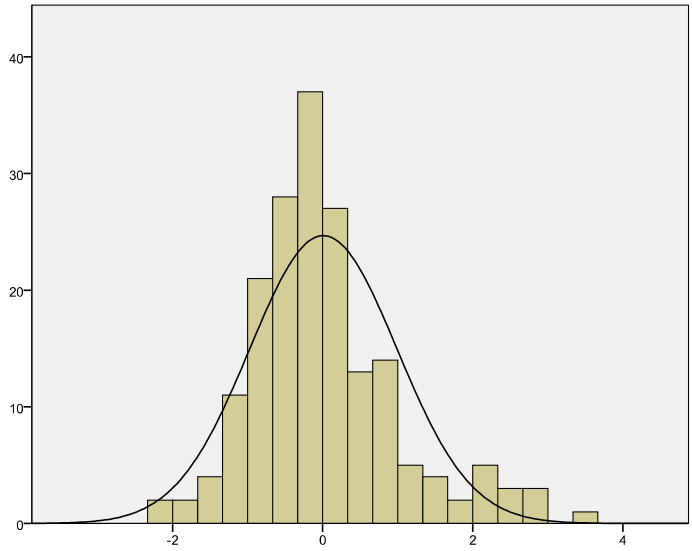


Gráfico de normalidad de los residuos estandarizados de la regresión

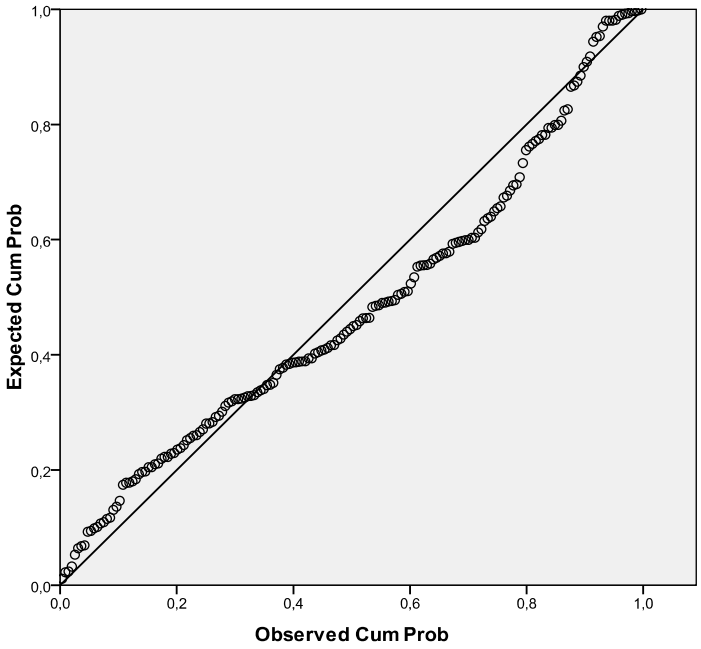
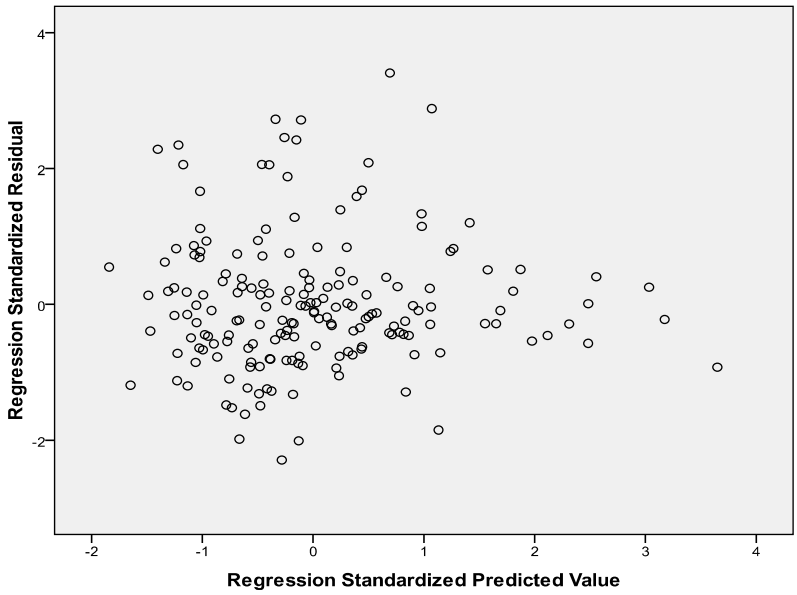
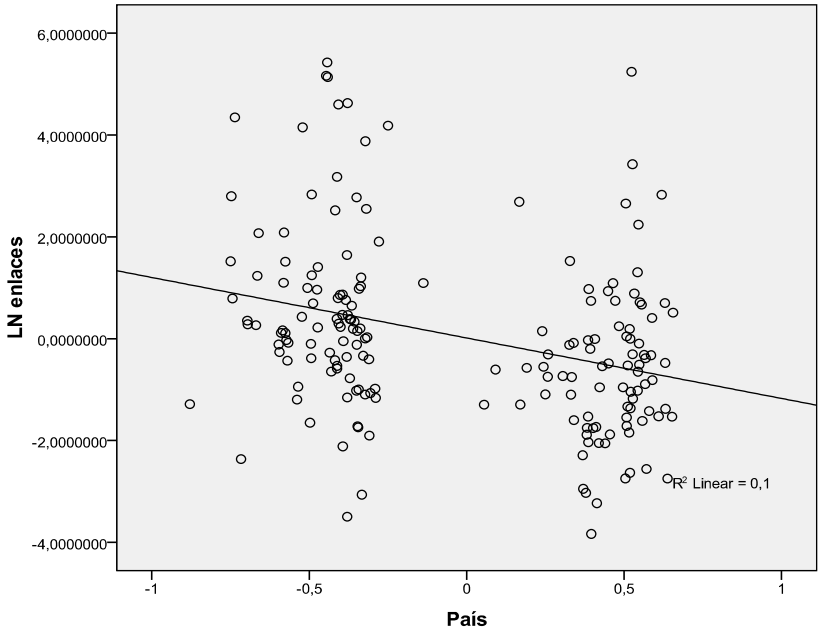
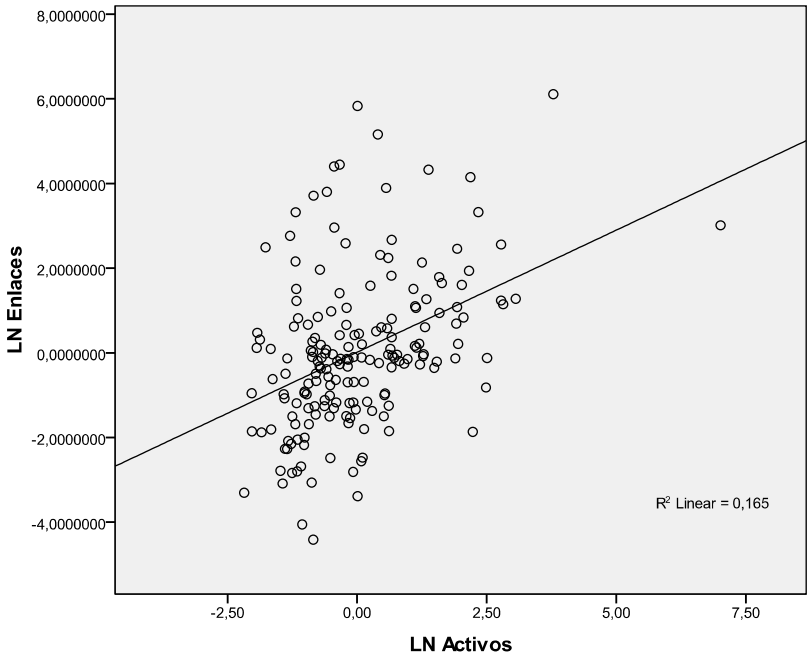


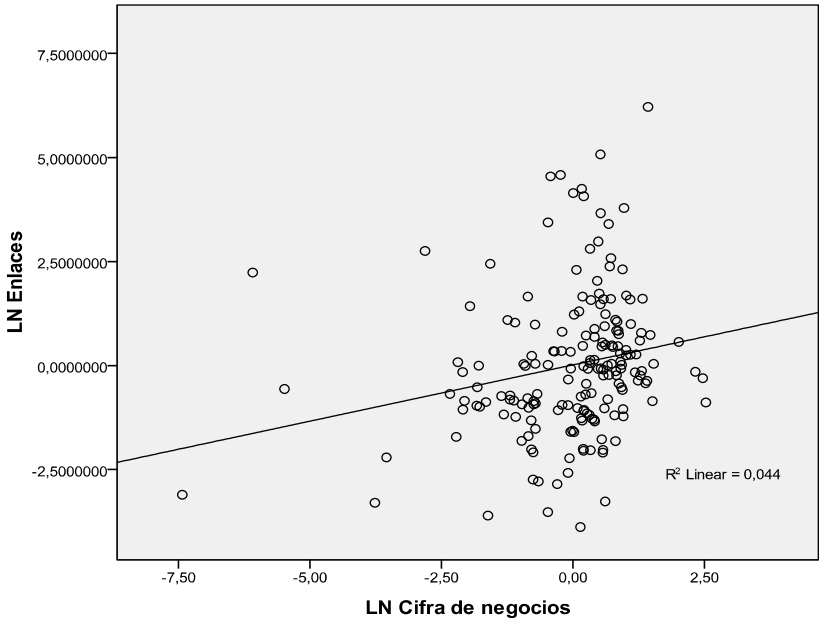
Diagrama de residuos estandarizados observados y esperados



Diagramas de regresión parcial de las variables del modelo




8,0000000
6,0000000
4,0000000
2,0000000
0,0000000
-2,0000000
-4,0000000
LN Enlaces
R² Linear = 0,165
-2,50
0,00
2,50
5,00
7,50
LN Activos


7,5000000
5,0000000
2,5000000
0,0000000
-2,5000000
LN Enlaces
R² Linear = 0,044
-7,50
-5,00
-2,50
0,00
2,50
LN Cifra de negocios

Sector: *Hostelería y restauración*

Histograma de residuos estandarizados

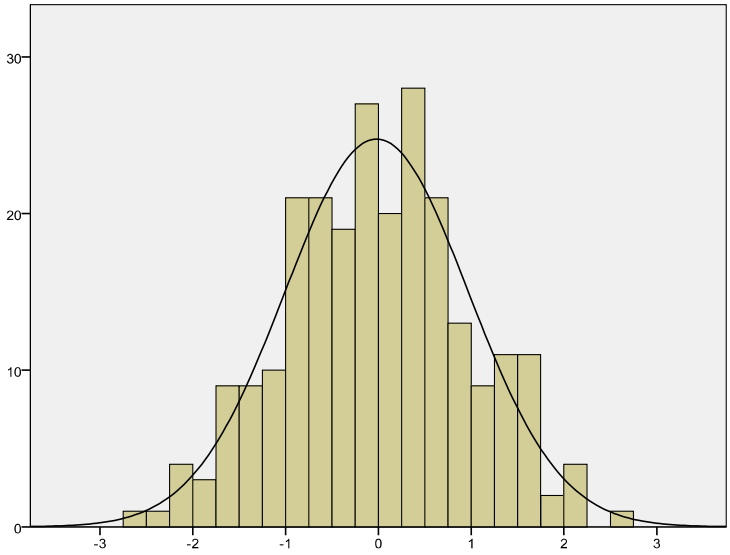


Gráfico de normalidad de los residuos estandarizados de la regresión

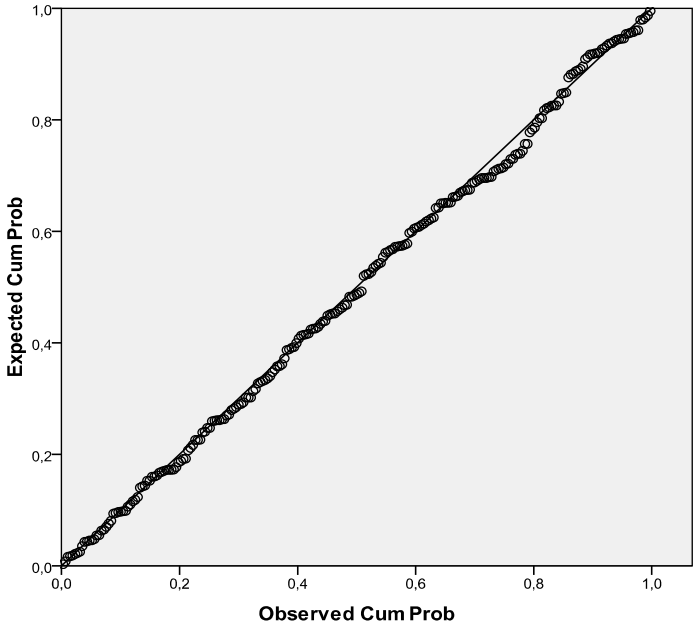
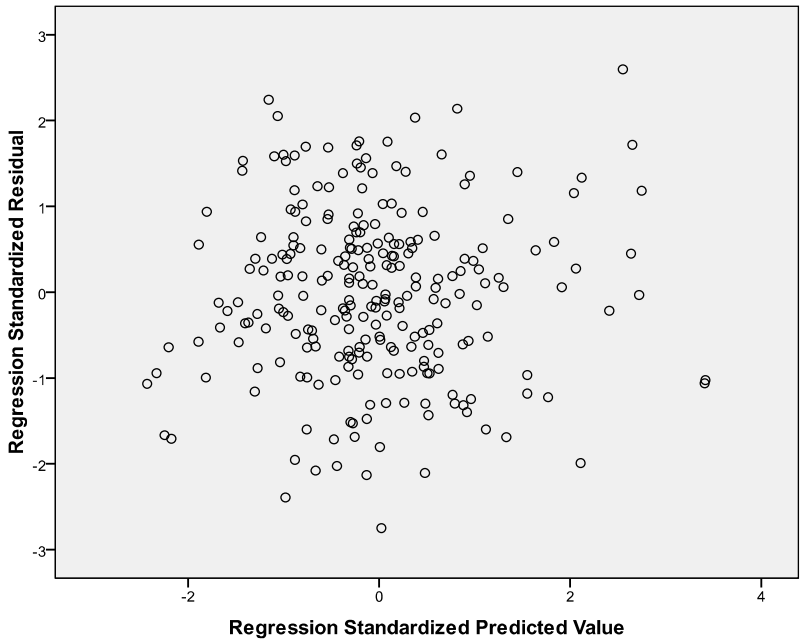


Diagrama de residuos estandarizados observados y esperados



Sector: *Servicios*

Histograma de residuos estandarizados

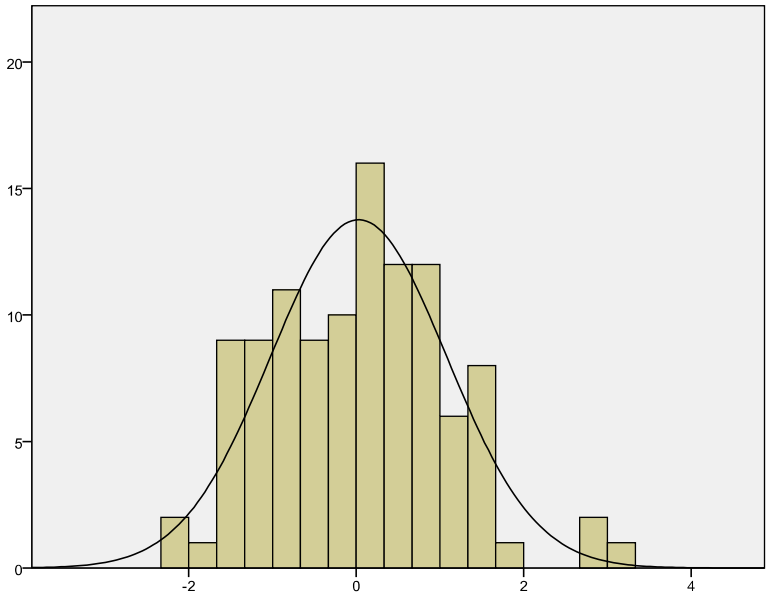


Gráfico de normalidad de los residuos estandarizados de la regresión

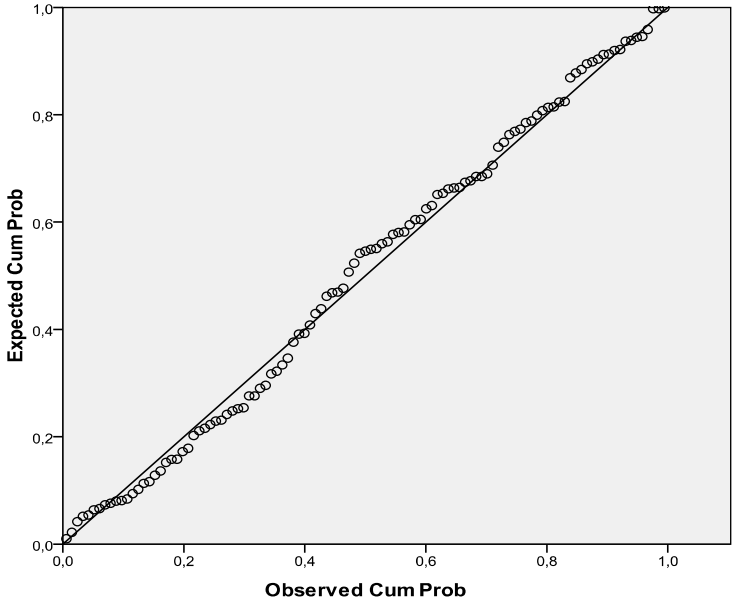
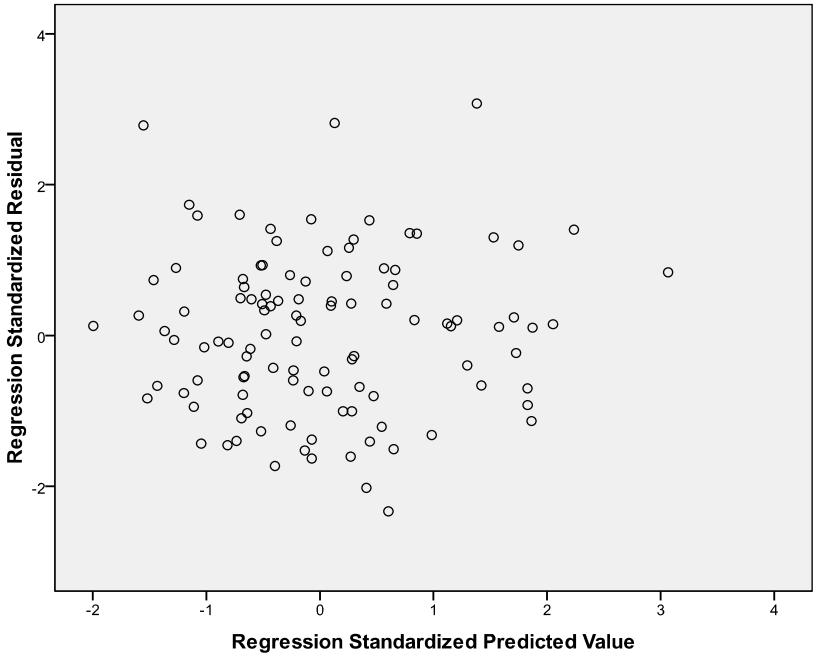


Diagrama de residuos estandarizados observados y esperados



Sector: *Industria editorial (excepto Internet)*

Histograma de residuos estandarizados

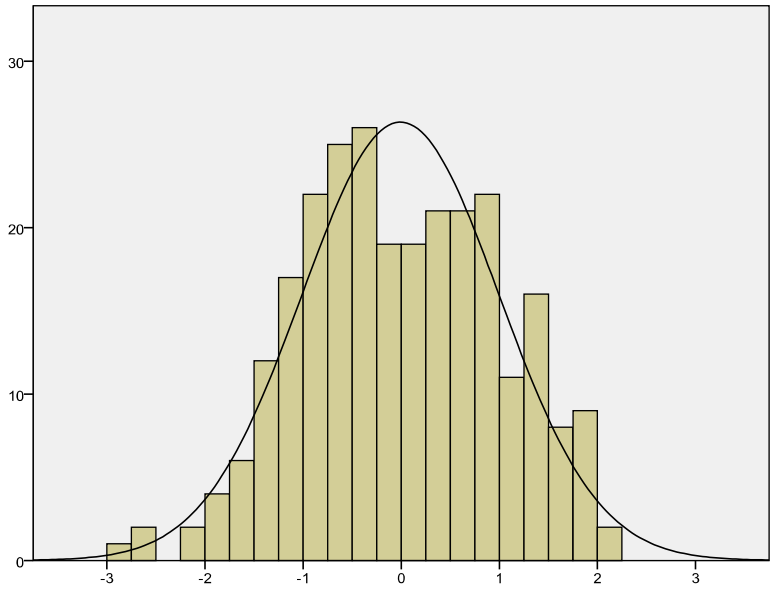


Gráfico de normalidad de los residuos estandarizados de la regresión

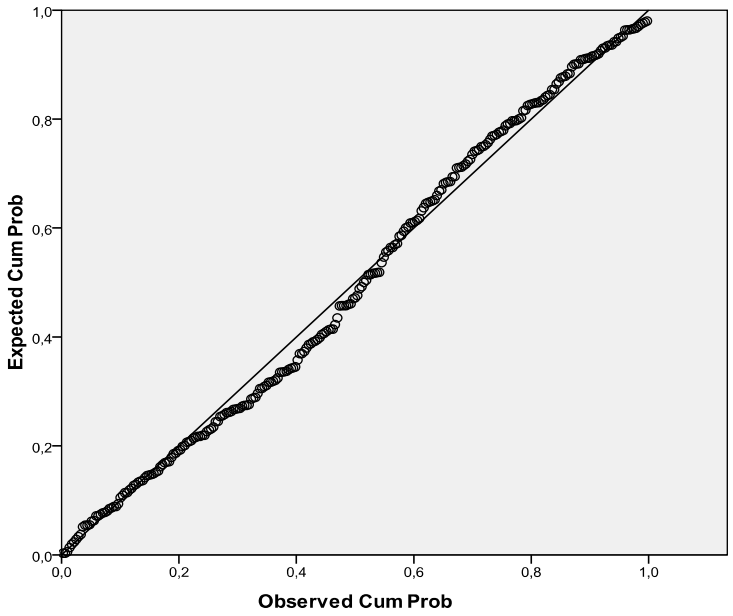
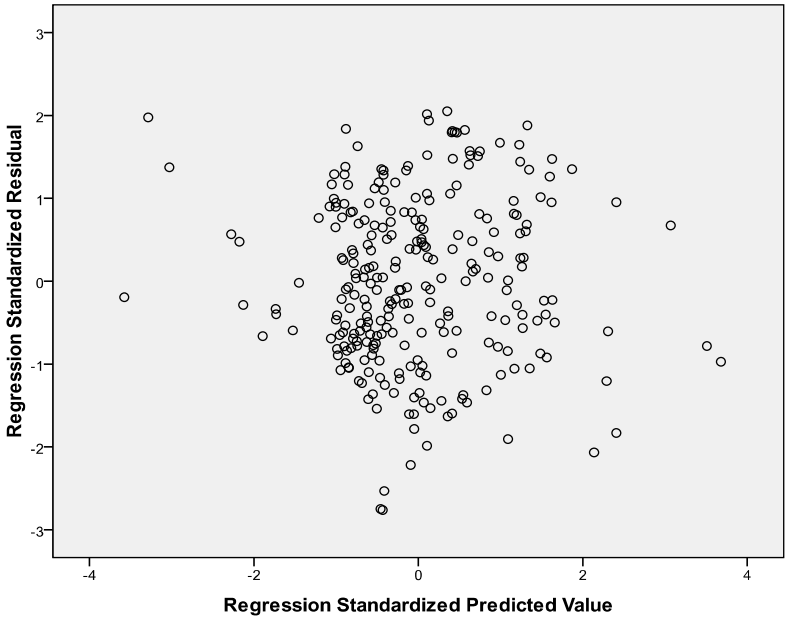
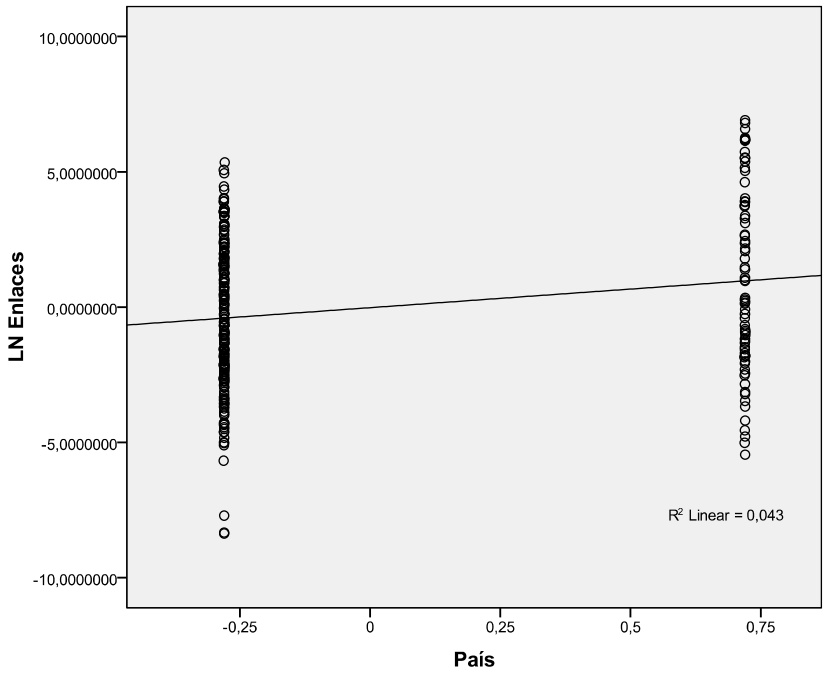
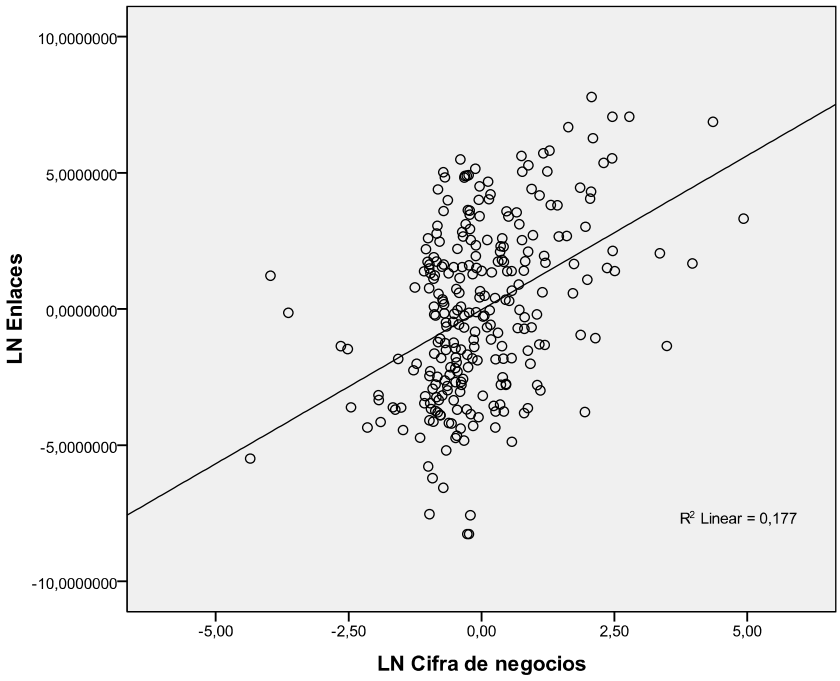


Diagrama de residuos estandarizados observados y esperados



Diagramas de regresión parcial de las variables del modelo




10,0000000
5,0000000
0,0000000
-5,0000000
-10,0000000
LN Enlaces
-5,00
-2,50
0,00
2,50
5,00
LN Cifra de negocios
R² Linear = 0,177

Sector: *Telecomunicaciones*

Histograma de residuos estandarizados

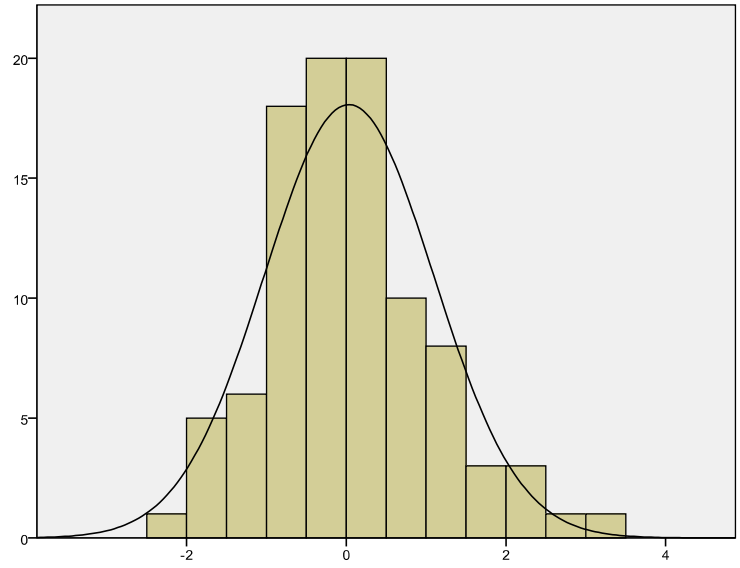


Gráfico de normalidad de los residuos estandarizados de la regresión

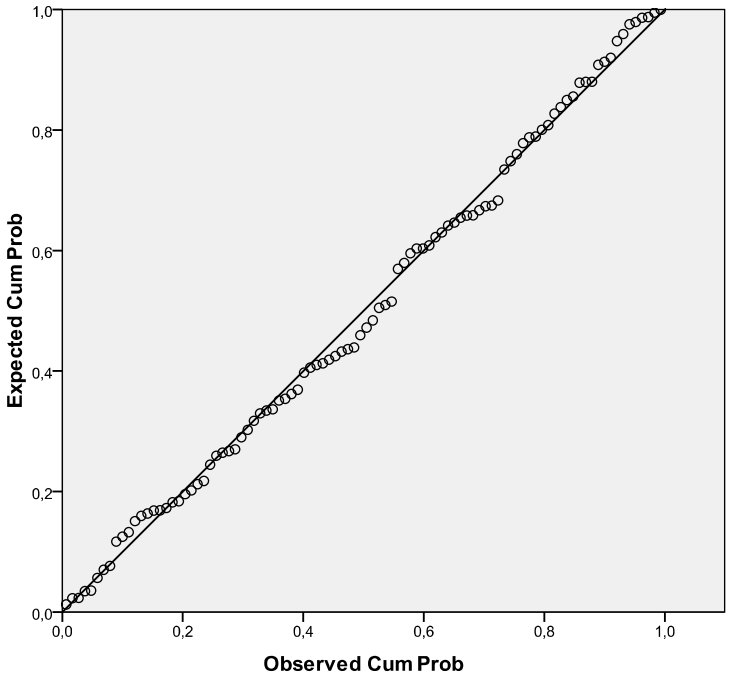
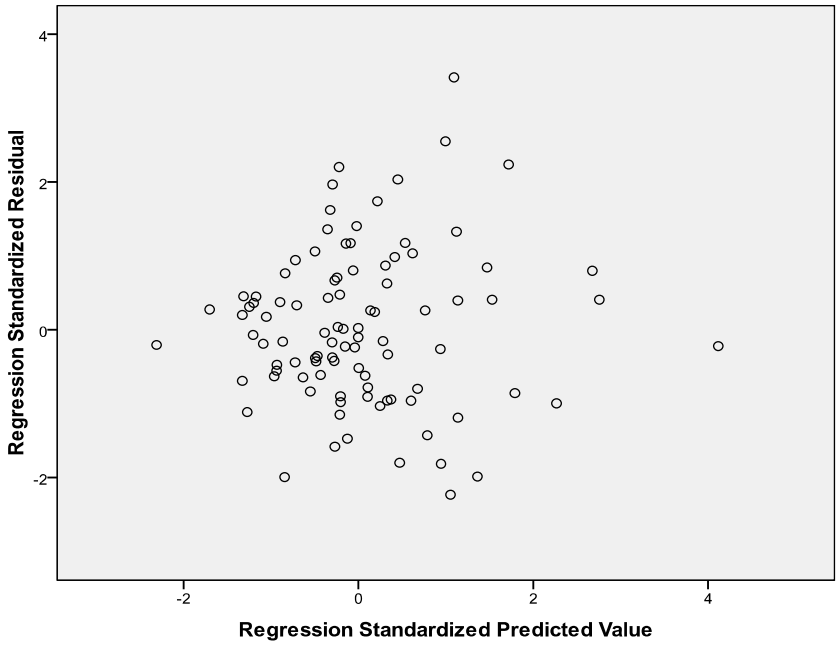
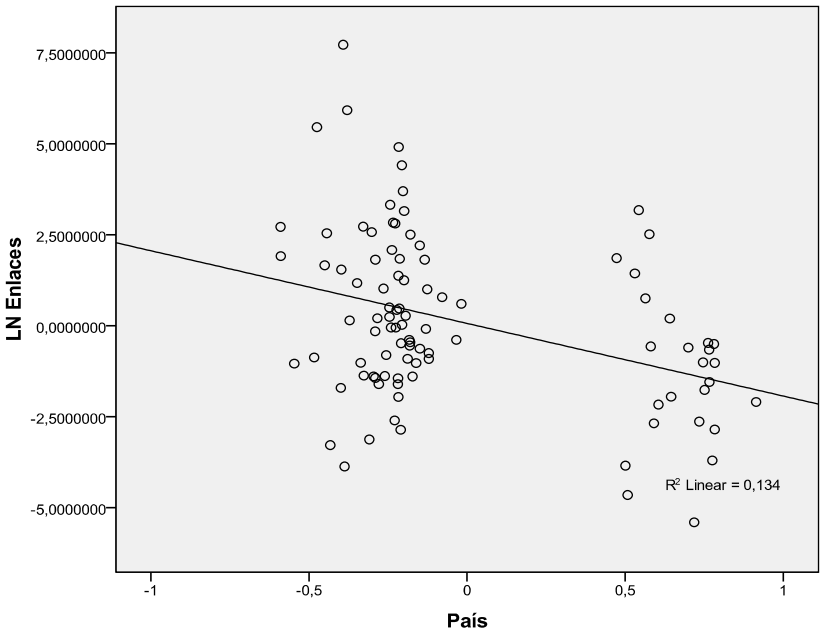
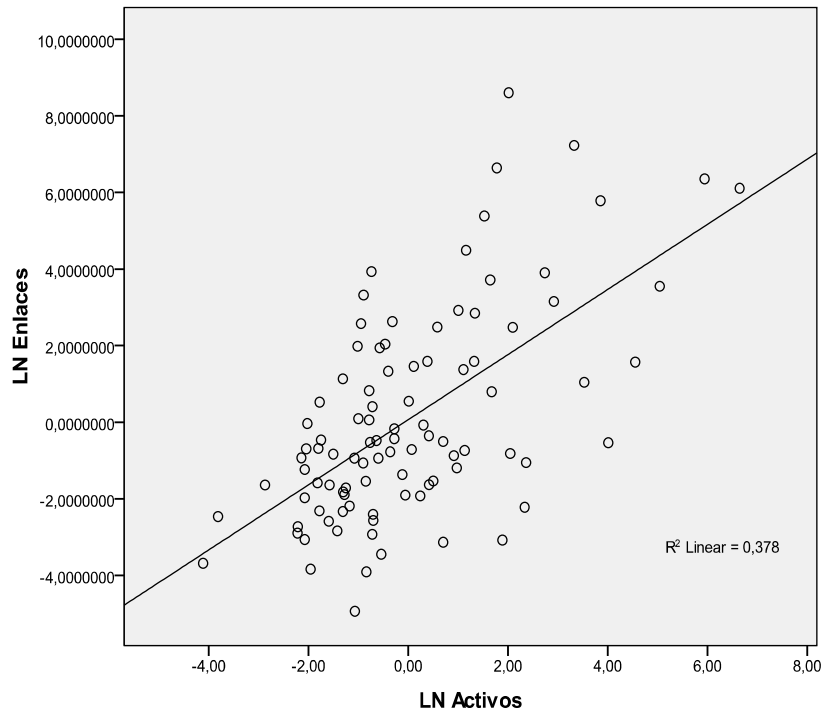


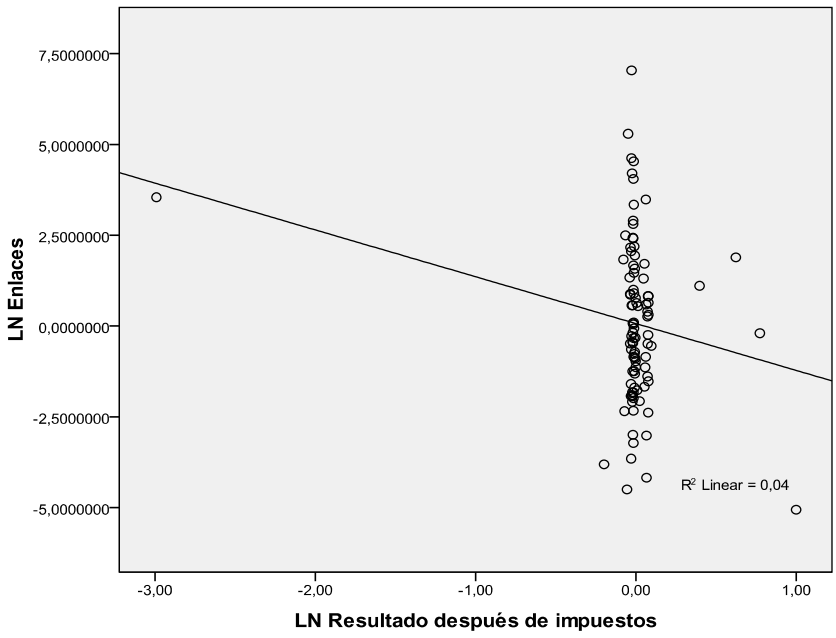
Diagrama de residuos estandarizados observados y esperados



Diagramas de regresión parcial de las variables del modelo




LN Enlaces
10,0000000
8,0000000
6,0000000
4,0000000
2,0000000
0,0000000
-2,0000000
-4,0000000
R² Linear = 0,378
-4,00
-2,00
0,00
2,00
4,00
6,00
8,00
LN Activos


LN Enlaces
7,5000000
5,0000000
2,5000000
0,0000000
-2,5000000
-5,0000000
R² Linear = 0,04
-3,00
-2,00
-1,00
0,00
1,00
LN Resultado después de impuestos

Anexo III. Empresas incluidas en el estudio de índices bursátiles (Apartado 4.2.1)

Empresas del Dow Jones Industrial				
Siglas	Empresa	URL principal	Sector	Enlaces recibidos
MMM	3M Co.	http://www.3m.com	Industria diversificada	251.000
AA	Alcoa Inc.	http://www.alcoa.com	Aluminio	42.900
AXP	American Express Co.	http://www.americanexpress.com	Finanzas de consumo	1.120.000
T	AT&T Inc.	http://www.att.com	Telecom. de línea fija	2.700.000
BAC	Bank of America Corp.	http://www.bankofamerica.com	Bancos	364.000
BA	Boeing Co.	http://www.boeing.com	Aerospacial	247.000
CAT	Caterpillar Inc.	http://www.cat.com	Camiones y vehículos comerciales	227.000
CVX	Chevron Corp.	http://www.chevron.com	Petróleo y gas integrados	233.000
C	Citigroup Inc.	http://www.citigroup.com	Bancos	128.000
KO	Coca-Cola Co.	http://www.coca-cola.com	Bebidas no alcohólicas	214.000
DD	E.I. DuPont de Nemours & Co.	http://www.dupont.com	Productos químicos	141.000
XOM	Exxon Mobil Corp.	http://www.exxonmobil.com	Petróleo y gas integrados	264.000
GE	General Electric Co.	http://www.ge.com	Industria diversificada	262.000
GM	General Motors Corp.	http://www.gm.com	Automóviles	242.000

HPQ	Hewlett-Packard Co.	http://www.hp.com	Hardware	2.900.000
HD	Home Depot Inc.	http://www.homedepot.com	Minoristas de productos de hogar	236.000
INTC	Intel Corp.	http://www.intel.com	Semiconductores	2.070.000
IBM	International Business Machines Corp.	http://www.ibm.com	Servicios informáticos	3.830.000
JNJ	Johnson & Johnson	http://www.jnj.com	Farmacéutico	91.300
JPM	JPMorgan Chase & Co.	http://www.chase.com	Bancos	232.000
KFT	Kraft Foods Inc. Cl A	http://www.kraft.com	Alimentación	142.000
MCD	McDonald's Corp.	http://www.mcdonalds.com	Restauración	636.000
MRK	Merck & Co. Inc.	http://www.merck.com	Farmacéutico	253.000
MSFT	Microsoft Corp.	http://www.microsoft.com	Software	45.500.000
PFE	Pfizer Inc.	http://www.pfizer.com	Farmacéutico	130.000
PG	Procter & Gamble Co.	http://www.pg.com	Productos para mantenimiento del hogar	272.000
UTX	United Technologies Corp.	http://www.utc.com	Aerospacial	42.800
VZ	Verizon Communications	http://www22.verizon.com	Telecom. de línea fija	332.000
WMT	Wal-Mart Stores Inc.	http://www.walmart.com	Minorista generalista	1.310.000

DIS	Walt Disney Co.	http://www.disney.go.com	Medios y entretenimiento	5.500.000

Empresas del FTSE 100 (35 empresas con mayor número de enlaces recibidos)				
Siglas	Empresa	URL principal	Sector	Enlaces recibidos
AZN	Astrazeneca	http://www.astrazeneca.com	Farmacéutico	59.600
BSY	B Sky B Group	http://www.sky.com	Medios; Televisión	2.040.000
BA	BAE Systems	http://www.baesystems.com	Aeroespacial y equipos	35.700
BARC	Barclays	http://www.barclays.co.uk	Financiero; Bancos	62.500
BLT	BHP Billiton	http://www.bhpbilliton.com	Metalúrgico	19.100
BP	BP	http://www.bp.com	Petróleo y gas natural	140.000
BAY	British Airways	http://www.britishairways.com	Transportes; Aerolíneas	241.000
BT-A	BT Group	http://www.bt.com	Telecomunicaciones	451.000
CW	Cable & Wireless	http://www.cw.com	Telecomunicaciones	20.300
CCL	Carnival	http://www.carnival.com	Ocio; Agente de viajes	69.800
DGE	Diageo	http://www.diageo.com	Bebidas	18.200
EXPN	Experian	http://www.experian.com	Servicios de inteligencia para empresas	423.000
FGP	Firstgroup	http://www.firstgroup.com	Transportes	30.600
GSK	Glaxosmithkline	http://www.gsk.com	Farmacéutico	124.000

HSBA	HSBC Hldg	http://www.hsbc.com	Financiero; Bancos	188.000
IHG	Intercont Hotels	http://www.intercontinental.com	Ocio; Alojamiento	142.000
LLOY	Lloyds Banking Grp	http://www.lloydstsb.com	Financiero; Bancos	31.300
LSE	LSE Group	http://www.londonstockexchange.com	Financiero; Bolsa de valores	149.000
MKS	Marks & Spencer	http://www.marksandspencer.com	Distribución, supermercados	36.100
NXT	Next	http://www.next.co.uk	Textil y piel; Ropa	19.700
PSON	Pearson	http://www.pearson.com	Medios; Libros	22.100
REL	Reed Elsevier	http://www.reed-elsevier.com	Medios; Prensa	828.000
RIO	Rio Tinto	http://www.riotinto.com	Metalúrgico	21.700
RR	Rolls-Royce Group	http://www.rolls-royce.com	Vehículos y equipos	23.700
RBS	Royal Bank Scotland Group	http://www.rbs.com	Financiero; Bancos	26.600
RDSB	Royal Dutch Shell-B	http://www.shell.com	Petróleo y gas natural	202.000
SGE	Sage Group	http://www.sage.com	Tecnologías de la Información; Software	73.000
STAN	Standard Chartered	http://www.standardchartered.com	Financiero; Bancos	37.400
TSCO	Tesco plc	http://www.tesco.com	Distribución, supermercados	146.000
TCG	Thomas Cook Group	http://www.thomascook.com	Ocio; Agente de viajes	52.000
TRIL	Thomson Reuters	http://www.reuters.com	Medios; Minorista especializado	10.400.000

TT	Tui Travel	http://www.tuitravelplc.com	Ocio; Agente de viajes	92.600
ULVR	Unilever	http://www.unilever.com	Alimentación	54.400
VOD	Vodafone Group	http://www.vodafone.com	Telecom.; Móviles	148.000
WPP	WPP	http://www.wpp.com	Servicios; Anuncios	34.700

Empresas del Euro Stoxx 50

Siglas	Empresa	URL principal	País	Sector	Enlaces recibidos
AEGN	Aegon	http://www.aegon.com	NL	Financiero; Seguros de vida	11.600
AIRP	Air Liquide	http://www.airliquide.com	FR	Químico	18.100
ALVG	Allianz	http://www.allianz.com	DE	Financiero; Seguros	38.600
ALSO	Alstom	http://www.alstom.com	FR	Maquinaria industrial	24.800
ISPA	ArcelorMittal	http://www.arcelormittal.com	LU	Metales y minería industrial	17.900
GASI	Assicurazioni Generali	http://www.generali-gw.com	IT	Financiero; Seguros	122
AXAF	Axa	http://www.axa.com	FR	Financiero; Seguros	32.300
BASF	Basf	http://www.basf.com	DE	Químico	102.000
BAYG	Bayer	http://www.bayer.com	DE	Químico	110.000
BBVA	Banco Bilbao Vizcaya Argentaria	http://www.bbva.com	ES	Financiero; Bancos	33.800
SAN	Banco	http://www.santander.com	ES	Financiero;	366.000

	Santander			Bancos	
BNPP	BNP Paribas	http://www.bnpparibas.com	FR	Financiero; Bancos	112.000
CARR	Carrefour Supermarché	http://www.carrefour.com	FR	Minorista de alimentación y generalista	32.300
CAGR	Crédit Agricole	http://www.credit-agricole.com	FR	Financiero; Bancos	26.200
DAIGn	Daimler	http://www.daimler.com	DE	Automóviles	47.200
DBKGn	Deutsche Bank	http://www.db.com	DE	Financiero; Bancos	76.900
DB1Gn	Deutsche Boerse	http://www.deutsche-boerse.com	DE	Financiero, Servicios de inversión	88.800
DTEGn	Deutsche Telekom	http://www.telekom.de	DE	Telecom.; Móviles	222.000
EONGn	E.ON	http://www.eon.com	DE	Servicios	68.200
ENEI	Enel	http://www.enel.com	IT	Servicios	12.900
ENI	Eni	http://www.eni.it	IT	Petróleo y gas	93.600
FOR	Fortis	http://www.fortis.com	NL	Financiero; Bancos	21.800
FTE	France Telecom	http://www.francetelecom.com	FR	Telecom.; Línea fija	885.000
GSZ	GDF Suez	http://www.gdfsuez.com	FR	Servicios	16.000
DANO	Danone	http://www.danone.com	FR	Alimentación y bebidas	18.300
SOGN	Société Générale	http://www.societegenerale.fr	FR	Financiero; Bancos	52.800
IBE	Iberdrola	http://www.iberdrola.es	ES	Servicios	50.700

ING	ING grp	http://www.ing.com	NL	Financiero; Seguros	57.400
ISP	Intesa Sanpaolo	http://www.intesasanpaolo.com	IT	Financiero; Bancos	40.900
OREP	L'Oreal	http://www.loreal.com	FR	Bienes de consumo; Productos de uso personal	29.900
LVMH	LVMH Moet Hennessy	http://www.lvmh.com	FR	Bienes de consumo; Productos de uso personal	18.200
MUVGn	Muenchener Rueck	http://www.munichre.com	DE	Financiero; Seguros	8.870
NOK1V	Nokia	http://www.nokia.com	FI	Tecnología, *hardware* y equipos	1.340.000
PHG	Philips Electronics	http://www.philips.com	NL	Bienes de consumo; Productos electrónicos	485.000
RENA	Renault	http://www.renault.com	FR	Bienes de consumo; Automóviles	59.100
REP	Repsol YPF	http://www.repsol.com	ES	Petróleo y gas	13.000
RWEG	RWE	http://www.rwe.com	DE	Servicios	39.500
SGOB	Saint Gobain	http://www.saint-gobain.com	FR	Industria, construcción y materiales	61.300
SASY	Sanofi-Aventis	http://www.sanofi-aventis.com	FR	Farmacéutico	18.200

SAPG	SAP	http://www.sap.com	DE	Tecnología, *software*	297.000
SCHN	Schneider Electric	http://www.schneider-electric.com	FR	Industria, componentes eléctricos	41.800
SIEGn	Siemens	http://w1.siemens.com	DE	Industria diversificada	48.500
TLIT	Telecom Italia	http://www.telecomitalia.com	IT	Telecom.; Línea fija	5.140
TEF	Telefónica	http://www.telefonica.es	ES	Telecom.; Línea fija	279.000
TOTF	Total	http://www.total.com	FR	Petróleo y gas	60.900
CRDI	Unicredit	http://www.unicreditgroup.eu	IT	Financiero; Bancos	58.000
UNc	Unilever NV	http://www.unilever.com	NL	Bienes de consumo; Alimentación	54.000
SGEF	Vinci	http://www.vinci.com	FR	Industria, construcción y materiales	6.370
VIV	Vivendi	http://www.vivendi.com	FR	Medios	23.400
VOWG	Volkswagen	http://www.volkswagen.com	DE	Automóviles	183.000
Las claves correspondientes a los países son: DE (Alemania), ES (España), FR (Francia), FI (Finlandia), IT (Italia), LU (Luxemburgo), NL (Holanda).					

Empresas del CAC 40 (selección de empresas relevantes para el análisis)				
Siglas	Empresa	URL principal	Sector	Enlaces recibidos
AC	Accor	http://www.accor.com	Ocio; Viajes; Hoteles	75.900
ALU	Alcatel-Lucent	http://www.alcatel-lucent.com	Tecnología, *hardware* y equipo	69.100
BNP	BNP Paribas	http://www.bnpparibas.com	Financiero; Bancos	111.000
CAP	Cap Gemini	http://www.capgemini.com	Software y servicios informáticos	77.900
ACA	Crédit Agricole	http://www.credit-agricole.com	Financiero; Bancos	26.300
FTE	France Telecom	http://www.francetelecom.com	Telecom. de línea fija	881.000
MMB	Lagardere	http://www.lagardere.com	Medios	5.890
UG	Peugeot	http://www.peugeot.com	Automóviles	47.200
GLE	Société Générale	http://www.societegenerale.fr	Financiero; Bancos	51.500
STM	STMicroelectronics	http://www.st.com	Tecnología, *hardware* y equipo	75.000
SEV	Suez Environnement	http://www.suez-environnement.com	Bienes y servicios industriales	21.800
VK	Vallourec	http://www.vallourec.com	Bienes y servicios industriales	2.600
VIV	Vivendi	http://www.vivendi.com	Medios	23.500

Empresas del IBEX 35				
Siglas	Empresa	URL principal	Sector	Enlaces recibidos
ABG	Abengoa	http://www.abengoa.es	Ingeniería	3.030
ABE	Abertis Infraestructuras Serie A	http://www.abertis.com	Autopistas y aparcamientos	142.000
ANA	Acciona	http://www.acciona.es	Construcción	23.700
ACX	Acerinox	http://www.acerinox.es	Metales y minería industriales	772
ACS	ACS Actividades Const. y Servicios	http://www.grupoacs.com	Construcción	3.710
BBVA	Banco Bilbao Vizcaya Argentaria	http://www.bbva.es	Bancos	214.000
SAB	Banco de Sabadell	https://www.bancsabadell.com	Bancos	5.860
BTO	Banco Español de Credito	http://www.banesto.es	Bancos	33.200
POP	Banco Popular Español	http://www.bancopopular.es	Bancos	18.200
SAN	Banco Santander Central Hispano	http://www.santander.com	Bancos	367.000
BKT	Bankinter	https://www.bankinter.com	Bancos	8.490

BME	Bolsas y Mercados Españoles	http://www.bolsasymercados.es	Servicios de inversión	15.800
CIN	Cintra Conc. Infra. de Transporte	http://www.cintra.es	Autopistas y aparcamientos	712
MAP	Corporacion Mapfre	http://www.mapfre.com	Seguros	24.400
CRI	Criteria Caixacorp	http://www.criteriacaixacorp.es	Holding	1.570
ENG	Enagas	http://www.enagas.es	Servicios	1.670
ELE	Endesa	http://www.endesa.es	Servicios	61.500
FER	Ferrovial	http://www.ferrovial.es	Construcción	13.100
FCC	Fomento Constr. y Contratas	http://www.fcc.es	Construcción	20.600
GAM	Gamesa Corporación Tecnológica	http://www.gamesa.es	Equipos industriales	4.570
GAS	Gas Natural	http://www.gasnatural.com	Servicios	28.800
TL5	Gestevision Telecinco	http://www.telecinco.es	Medios y anuncios	378.000
GRF	Grifols	http://www.grifols.com	Farmacéutico	1.350
IBE	Iberdrola	http://www.iberdrola.es	Servicios	50.600
IBR	Iberdrola Renovables	http://www.iberdrolarenovables.es	Energías renovables	1.270
IBLA	Iberia	http://www.iberia.es	Transporte	87.700
ITX	Ind. de Diseño	http://www.inditex.com	Textil	3.020

	Textil (inditex)			
IDR	Indra, serie A	http://www.indra.es	Electrónica y *software*	37.700
OHL	Obrascón Huarte Lain	http://www.ohl.es	Construcción	1.520
REE	Red Eléctrica de España	http://www.ree.es	Servicios	19.900
REP	Repsol YPF	http://www.repsol.com	Petróleo	13.100
SYV	Sacyr Vallehermoso	http://www.gruposyv.com	Construcción	6.330
TRE	Técnicas Reunidas	http://www.tecnicasreunidas.es	Ingeniería	344
TEF	Telefónica	http://www.telefonica.es	Telecomunicaciones	279.000
UNF	Unión Fenosa	http://www.unionfenosa.es	Servicios	18.400

Anexo IV. Listado de bancos cotizados en la NYSE y el importe de fondos federales recibidos (Apartado 4.2.2)

Empresa	Siglas	URL principal	Fondos federales a 20/8/2009 (en millones)
Bank of America Corporation	BAC	http://www.bankofamerica.com	25.000
Citigroup Inc.	C	http://www.citigroup.com	25.000
JPMorgan Chase & Co.	JPM	http://www.chase.com	25.000
Wells Fargo & Co.	WFC	http://www.wellsfargo.com	25.000
US Bancorp	USB	http://www.usbank.com	6.599
Suntrust Banks Inc.	STI	http://www.suntrust.com	4.850
Regions Financial Corporation	RF	http://www.regions.com	3.500
BB&T Corporation	BBT	http://www.bbt.com	3.133,64
KeyCorp	KEY	http://www.key.com	2.500
Comerica Inc.	CMA	http://www.comerica.com	2.250
Marshall & Ilsley Corporation	MI	http://www.mibank.com	1.715
Synovus Financial Corp.	SNV	http://www.synovus.com	967,87
First Horizon National Corporation	FHN	http://www.firsthorizon.com	866,54
City National Corporation	CYN	http://www.cnb.com	400
Webster Financial Corporation	WBS	http://www.websteronline.com	400
TCF Financial Corporation	TCB	http://www.tcfbank.com	361,172
Wilmington Trust Corporation	WL	http://www.wilmingtontrust.com	330
Valley National Bancorp	VLY	http://www.valleynationalbank.com	300
* Flagstar Bancorp, Inc.	FBC	http://www.flagstar.com	266,657
Western Alliance Bancorporation	WAL	http://www.westernalliancebancorp.com	140
Central Pacific Financial Corp.	CPF	http://www.centralpacificbank.com	135

F.N.B. Corporation	FNB	http://www.fnb-online.com	100
Old National Bancorp	ONB	http://www.oldnational.com	100
Astoria Financial Corporation	AF	http://www.astoriafederal.com	0
BancorpSouth Inc	BXS	http://www.bancorpsouthonline.com	0
Bank of Hawaii Corporation	BOH	http://www.boh.com	0
BankAtlantic Bancorp, Inc.	BBX	http://www.bankatlantic.com	0
Capitol Bancorp Limited	CBC	http://www.capitolbancorp.com	0
Colonial Bancgroup Inc.	CNB	http://www.colonialbank.com	0
Community Bank System, Inc.	CBU	http://www.communitybankna.com	0
Cullen/Frost Bankers, Inc.	CFR	http://www.frostbank.com	0
Downey Financial Corporation	DSL	http://www.downeysavings.com	0
First Commonwealth Financial Corporation	FCF	http://www.fcbanking.com	0
Firstfed Financial Corporation	FED	http://www.firstfedca.com	0
Guaranty Financial Group Inc.	GFG	http://www.guarantygroup.com	0
Imperial Capital Bancorp, Inc.	IMP	http://www.itlacapital.com	0
Irwin Financial Corporation	IFC	http://www.irwinfinancial.com	0
M&T Bank Corporation	MTB	http://www.mandtbank.com	0
New York Community Bancorp, Inc.	NYB	http://www.mynycb.com	0
NewAlliance Bancshares, Inc.	NAL	http://www.newalliancebank.com	0
Oriental Financial Group, Inc.	OFG	http://www.orientalonline.com	0
PNC Financial Services Group	PNC	http://www.pnc.com	0
Provident Financial Services, Inc.	PFS	http://www.providentnj.com	0

Sovereign Bancorp, Inc.	SOV	http://www.sovereignbank.com	0
Sterling Bancorp	STL	http://www.sterlingbancorp.com	0
Wachovia	WB	http://www.wachovia.com	0
Washington Mutual Inc.	WM	http://www.wamu.com	0
* En el periodo entre enero y agosto únicamente Flagstar Bancorp, Inc. recibió fondos. En ese mismo periodo otros bancos devolvieron parte o la totalidad de los fondos federales recibidos.			

Anexo V. Listado de los 50 mayores bancos mundiales (Apartado 4.3.1)

Siglas	Empresa	URL principal	País	Enlaces dic. 08	Enlaces junio 09
RBS	The Royal Bank of Scotland Group plc	http://www.rbs.com	Reino Unido	26.700	35.200
DB	Deutsche Bank	http://www.db.com	Alemania	66.100	88.100
BNP	BNP Paribas SA	http://www.bnpparibas.com	Francia	143.000	112.000
Barclays	Barclays PLC	http://www.barclays.com	Reino Unido	42.900	28.200
Credit-Agri	Crédit Agricole SA	http://www.credit-agricole.com	Francia	28.700	30.900
UBS	UBS AG	http://www.ubs.com	Suiza	147.000	148.000
SocGen	Société Générale	http://www.socgen.com	Francia	54.300	42.000
ABN	ABN AMRO Holding NV	http://www.abnamro.com	Holanda	52.100	42.700
Unicredit	UniCredit SpA	http://www.unicreditgroup.eu	Italia	79.900	50.000
ING	ING Bank NV	http://www.ing.com	Holanda	58.800	64.900
MUFG	The Bank of Tokyo-Mitsubishi UFJ Ltd	http://www.mufg.jp	Japón	364.000	454.000
Santander	Banco Santander SA	http://www.santander.com	España	395.000	364.000
Chase	JPMorgan Chase Bank National Association	http://www.chase.com	Estados Unidos	223.000	144.000
BoA	Bank of America NA	http://www.bankofamerica.com	Estados Unidos	339.000	304.000
Citi	Citibank NA	http://www.citibank.com	Estados Unidos	140.000	98.700

Suisse	Credit Suisse Group	http://www.credit-suisse.com	Suiza	61.500	70.100
Fortis	Fortis Bank SA/NV	http://www.fortis.com	Bélgica	34.000	20.700
ICBC	Industrial & Commercial Bank of China Limited	http://www.icbc.com.cn	China	985.000	937.000
CCB	China Construction Bank Corporation	http://www.ccb.com	China	304.000	539.000
BoS	Bank of Scotland plc	http://www.bankofscotland.co.uk	Reino Unido	13.700	9.390
HSBC	HSBC Bank plc	http://www.hsbc.co.uk	Reino Unido	85.100	65.700
Intesa	Intesa Sanpaolo SpA	http://www.intesasanpaolo.com	Italia	47.800	46.700
SMBC	Sumitomo Mitsui Banking Corp.	http://www.smbc.co.jp	Japón	105.000	103.000
Commerz	Commerzbank AG	http://www.commerzbank.com	Alemania	10.600	9.100
Calyon	Calyon	http://www.calyon.com	Francia	10.900	6.590
Rabo	Rabobank Nederland	http://www.rabobank.com	Holanda	11.100	11.300
Dresdner	Dresdner Bank Group	http://www.dresdner-bank.com	Alemania	5.280	3.220
Caisse	Caisse Nationale des Caisses d'Epargne et de Prévoyance	http://www.caisse-epargne.com	Francia	28.800	31.700
Lloyds	Lloyds TSB	http://www.lloydstsb.com	Reino	40.400	28.900

	Group plc		Unido		
ABC	Agricultural Bank of China	http://www.abchina.com	China	193.000	273.000
BOC	Bank of China Limited	http://www.boc.cn	China	399.000	485.000
Danske	Danske Bank A/S	http://www.danskebank.com	Dinamar-ca	7.990	7.820
Wachovia	Wachovia Bank NA	http://www.wachovia.com	Estados Unidos	143.000	83.900
RBC	Royal Bank of Canada	http://www.royalbank.com	Canadá	40.600	40.500
Hypo	Bayerische Hypo- und Vereinsbank AG	http://www.hypovereinsbank.de	Alemania	14.000	16.300
Natixis	Natixis	http://www.natixis.fr	Francia	7.690	4.720
Nochu	The Norinchukin Bank	http://www.nochubank.or.jp	Japón	5.790	2.660
DZ	DZ Bank AG	http://www.dzbank.de	Alemania	9.650	14.300
Natwest	National Westminster Bank Plc	http://www.natwest.com	Reino Unido	21.800	18.300
Nordea	Nordea Group	http://www.nordea.com	Suecia	41.200	35.900
Mizuho	Mizuho Bank Ltd	http://www.mizuhobank.co.jp	Japón	124.000	113.000
LBBW	Landesbank Baden-Württemberg	http://www.lbbw.de	Alemania	6.190	7.070
MizuhoCBK	Mizuho Corporate Bank Ltd	http://www.mizuhocbk.co.jp	Japón	5.710	3.570
KfW	Kreditanstalt für Wiederaufbau	http://www.kfw.de	Alemania	44.300	62.700

BBVA	Banco Bilbao Vizcaya Argentaria SA	http://www.bbva.com	España	43.100	53.200
NAB	National Australia Bank Ltd	http://www.nab.com.au	Australia	19.800	21.200
Wellsfargo	Wells Fargo Bank NA	http://www.wellsfargo.com	Estados Unidos	346.000	291.000
Bayernlb	Bayerische Landesbank	http://www.bayernlb.de	Alemania	5.810	4.700
KBC	KBC Bank NV	http://www.kbc.com	Bélgica	17.000	20.300
TD	The Toronto-Dominion Bank	http://www.td.com	Canadá	37.300	39.800

Anexo VI. Versión en español del Capítulo 5: "Resumen y conclusiones finales"

> *«Si cada uno de nosotros extiende su atención de manera igual entre grupos de diferentes tamaños, desde lo personal a lo global, ayudamos a evitar esos extremos. Vínculo a vínculo construimos sendas de entendimiento a través de la red de la humanidad. Somos los hilos que cohesionan el mundo. Cuando hacemos esto, acabamos teniendo de manera natural unos cuantos sitios web muy demandados, y una continua disminución de la enorme cantidad de sitios con muy pocos visitantes. En otras palabras, por muy atractiva que pueda ser la igualdad entre pares, semejante estructura no es óptima por su uniformidad. No presta la suficiente atención a la coordinación global, y puede requerir muchos clics para ir desde el problema a la solución.»*
>
> *Tim Berners-Lee*
>
> *Tejiendo la red (1999:189)*

Antecedentes

Durante sus 20 años de vida, la Web se ha convertido en un laboratorio para las ciencias sociales (Apartado 2.3). Desde su creación, los científicos se han esforzado por estudiar y analizar los diversos fenómenos que tienen lugar en la Web, lo cuales se caracterizan por la gran cantidad de datos disponibles y por su continua variación y crecimiento. La posibilidad de obtener datos a gran escala hace posible describir la Web como una enorme base de datos de carácter heterogéneo y no estructurado, la cual, a pesar de su apariencia, no está construida al azar. Bajo esta perspectiva, los datos obtenidos de la Web se pueden explotar desde diferentes perspectivas (basándose en el contenido, la estructura y el comportamiento del usuario) con el objeto de estudiar fenómenos que suceden únicamente en red y otros que ocurren fuera, pero que de algún modo se ven reflejados.

La combinación de ciencia e Internet ha dado lugar a varios conceptos, tales como

e-Ciencia o e-Investigación (Apartado 2.4.1). La e-Ciencia se refiere a la ciencia a gran escala que se lleva a cabo a través de redes globales de colaboración, basadas fundamentalmente en Internet. Muchas iniciativas de e-Ciencia han subrayado la importancia de la capacidad de procesamiento informático de forma distribuida y la computación grid, aunque en muchas ocasiones, especialmente cuando nos referimos a las ciencias sociales, la colaboración entre investigadores a través de Internet no exige el empleo de este tipo de recursos. Por ejemplo, la investigación basada en información acerca del contenido y la estructura de la Web no requiere necesariamente este tipo de infraestructura por parte del investigador. Existen muchas potenciales bases de datos disponibles; por ejemplo, los motores de búsqueda podrían usarse para recoger datos con el objetivo de investigar en diferentes aspectos de la Web (Apartado 3.3.3). La principal ventaja que proporcionan los buscadores es el libre acceso a sus datos así como su potencial para indexar gran parte de la Web, proporcionando datos acerca del contenido de las páginas y de los enlaces que las vinculan. Estos datos podrían emplearse para descubrir patrones y tendencias que, de otro modo, quedarían ocultos. Una potencial fuente de información en la Web proviene de la visualización de redes de hiperenlaces. La explotación de este tipo de información se ha realizado aplicando técnicas de minería de datos.

En esta línea, la Webmetría ha surgido como una nueva disciplina que aplica, a la Web conceptos y métodos que proceden de la Bibliometría y de las Ciencias de la Información (Capítulo 3). Ha sido definida por Björneborn (2004; en Björneborn y Ingwersen, 2004: 1217) como "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches". Sin embargo, el estudio de la Web difiere significativamente del estudio de la producción científica y del ámbito académico. Las diferencias se pueden entender más fácilmente a través del análisis de algunas de las principales características de este medio (Apartado 2.4.2). Su conocimiento también nos puede ayudar a entender las implicaciones y limitaciones de los hallazgos realizados mediante este tipo de investigación. Por ejemplo, la naturaleza dinámica de la Web (Ke et al., 2005) implica que cualquier investigación de este tipo sea siempre el análisis de una fotografía fija en un

momento específico, sujeta en todo caso a las limitaciones de las herramientas empleadas para la obtención de la información (por ejemplo, los motores de búsqueda). Se trata también de investigaciones que son imposibles de reproducir en el futuro.

Como se ha mencionado, los motores de búsqueda son una de las fuentes de datos de la Web más importantes, especialmente cuando pretendemos realizar un análisis en su conjunto. Sin embargo, existe una serie de limitaciones significativas que se derivan de la finalidad comercial que tienen los motores de búsqueda (Apartado 3.3.3.5). Aquí podemos resumir algunas de ellas, las cuales permiten subrayar algunas de las especificidades de la Web, en tanto que objeto de investigación:

- Los motores de búsqueda no indexan toda la Web (Lawrence y Giles, 1998; 1999b; Sherman y Price, 2001; Bar-Ilan, 2004; Thelwall, Vaughan y Björneborn, 2005). En ocasiones esto no se debe a cuestiones puramente técnicas, sino al diseño de los sitios web, que pueden prohibir a las arañas de los buscadores el acceso a algunas de sus partes (Koster, 2009a; 2009b). Vaughan y Zhang (2007) señalan que el porcentaje de cobertura de los motores de búsqueda se ha ido incrementado.

- Los sistemas de clasificación de los resultados por parte de los buscadores eliminan páginas que son idénticas o similares, con el fin de evitar proporcionar información poco relevante a los usuarios (Gomes y Smith, 2003; Thelwall, 2008a).

- Los algoritmos empleados para indexar la Web y para proporcionar las consultas constituyen secretos comerciales y, por lo tanto, los criterios exactos empleados en la elaboración de *rankings* de resultados son desconocidos (Thelwall, Vaughan y Björneborn, 2005).

- El número total de resultados ofrecidos por los motores de búsqueda son estimaciones, dado que emplean algoritmos que priorizan el tiempo de respuesta frente a la exhaustividad (Björneborn y Ingwersen, 2001).

- Los resultados pueden estar afectados por sesgos debidos al país o a la lengua (Vaughan y Thelwall, 2004).

- Los resultados están sujetos a fluctuaciones y cambios a lo largo del tiempo (Bar-Ilan, 2000; Mettrop y Nieuwenhuysen, 2001). Además, sólo un pequeño número de páginas son accesibles (por lo general, un máximo de mil), generando problemas en los casos en los que, por ejemplo, se necesita una muestra aleatoria. Sin embargo, algunos trabajos también indican que los motores de búsqueda son cada vez más estables a la hora de efectuar consultas (por ejemplo: Thelwall, 2001a; 2001b; Vaughan y Thelwall, 2003; Vaughan, 2004a). Uyar (2009) ofrece datos recientes sobre la exactitud del número de resultados que ofrecen los motores de búsqueda al realizar una consulta.

A pesar de estas limitaciones, estudios previos en diferentes campos han puesto de relieve la existencia de patrones en redes de enlaces y correlaciones significativas entre fenómenos *online* y *offline*, validando los resultados mediante el empleo de análisis de contenidos de las páginas que establecen los enlaces. Sin embargo, es esencial tener en mente estas limitaciones cuando se analizan e interpretan los datos obtenidos de la Web.

Esta tesis se centra en el análisis de hiperenlaces (Apartado 3.4), esto es, en el análisis de la estructura de la Web. Un hiperenlace puede definirse como una referencia, incluida en un documento, a otra sección del mismo documento o a otro documento distinto. De algún modo, los enlaces constituyen la estructura oculta de la Web, conectando diferentes sitios y páginas web que, de otro modo, permanecerían aislados a menos que se conociera su URL en particular (Berners-Lee, 1999). Los enlaces pueden considerarse como un respaldo o apoyo a la página web que enlazan, especialmente si el autor de los mismos los ha creado porque apuntan a un recurso que considera de interés. Esta idea recuerda

claramente a la propuesta de Garfield (1979) de emplear las citas bibliográficas como base para establecer *rankings* de las producciones científicas, una técnica que se emplea para evaluar la calidad de la investigación académica (Apartado 3.1). La creación de hiperenlaces no es un fenómeno accesorio, sino que tiene importantes implicaciones sociales (Turow y Lokman, 2008). Su gran potencial se ilustra muy bien a través del funcionamiento de los motores de búsqueda (Batelle, 2006); por ejemplo, el dominio del motor de búsqueda de Google (Apartado 3.3.3.4) deriva en buena medida del sistema *PageRank* (Brin y Page, 1998).

La aplicación de técnicas webmétricas a sitios web comerciales no está tan desarrollada como en el caso de los sitios web académicos (Thelwall, Vaughan y Björneborn, 2005). La Inteligencia Competitiva (*Competitive Intelligence*; Kahaner, 1996) es una de las áreas donde esta investigación ha tenido más éxito y es más prometedora (Tan et al., 2002; Reid, 2003). Consideramos que la aplicación de técnicas webmétricas y otras metodologías de investigación en Internet a las empresas puede proporcionar nuevos recursos a las organizaciones para generar y mantener ventajas competitivas (Porter, 1980; 1985). Por ejemplo, un estudio reciente (Choi y Varian, 2009), llevado a cabo por investigadores de Google, utilizó datos del servicio Google Trends para anticipar el comportamiento del consumidor en varios sectores empresariales.

La investigación empírica desarrollada en esta tesis se basa en dos perspectivas: en el análisis de impacto de enlaces y en el análisis de co-enlaces. Por un lado, el análisis de impacto de enlaces está basado en el número de páginas y sitios web que enlazan a un conjunto de páginas o sitios web incluidos en el estudio. El propósito de este tipo de investigación es, de acuerdo con Thelwall (2009: 28), "to evaluate whether a given website has a high link-based web impact compared to its peers". El número de enlaces recibidos puede constituir un indicador indirecto de otros atributos de la organización representados por el sitio web. Por ejemplo, se han empleado de manera habitual, dentro del campo académico, como un estimador potencial del desempeño investigador (Smith y Thelwall, 2002). En relación con los sitios web de empresas, Vaughan (2004a; 2004b) y Vaughan y Wu (2004) encontraron evidencia significativa de relaciones positivas entre el número

de enlaces recibidos por el sitio web de un grupo de empresas y sus variables financieras. Los estudios incluidos en los Apartados 4.1.1 y 4.1.2 amplían esta investigación, proporcionando nuevas evidencias en diferentes países y sectores de actividad. El estudio en el Apartado 4.1.3 representa uno de los primeros intentos de determinar las variables explicativas del número de enlaces recibidos por el sitio web de una empresa.

Por otra parte, el análisis de co-enlaces se podría considerar como un tipo de técnica de representación gráfica de enlaces (*link relationship mapping technique*) (Thelwall, 2009). Está basado en los enlaces que interconectan un conjunto de sitios web con el objeto de dibujar un gráfico que ilustra las relaciones entre ellos. En concreto, el análisis de co-enlaces está basado en el número de páginas web que enlazan al mismo tiempo a dos páginas o sitios web pertenecientes al conjunto de elementos en el estudio. Los co-enlaces son análogos al concepto bibliométrico de co-citación (Small, 1973). El análisis de co-enlaces ha demostrado ser una herramienta útil para revelar la estructura cognitiva o intelectual de un campo de estudio específico (Zuccala, 2006a). Este método es especialmente apropiado cuando los sitios web se enlazan poco entre sí de manera directa. Es el caso de los sitios web comerciales que, muy pocas veces, enlazan sitios web de empresas competidoras, sobre todo cuando pertenecen a un mismo sector de actividad (Vaughan, Gao y Kipp, 2006). La explicación se debe principalmente a que las empresas procuran evitar el desvío del tráfico de visitas y, en definitiva, de la atención del cliente a entidades competidoras (Shaw, 2001).

Además, como apunta Vaughan (2006), los datos de co-enlaces son más robustos que los datos de enlaces recibidos, dado que los primeros son más difíciles de manipular. El análisis de co-enlaces en empresas ha empezado centrándose en sectores de actividad particulares (Vaughan y You, 2006) o en subsectores dentro de una industria específica (Vaughan y You, 2008; 2009). Los estudios incluidos en los Apartados 4.2.1 y 4.2.2 exploran el empleo del análisis de co-enlaces para investigar empresas que pertenecen a diferentes sectores y para analizar la evolución de la crisis financiera en los bancos estadounidenses. Finalmente, la investigación incluida en el Apartado 4.3.1 combina ambos métodos con el fin de

proporcionar una visión global del sector bancario internacional.

Preguntas de investigación: hallazgos y contribuciones

En esta tesis se han analizado diferentes cuestiones relativas a sitios web de empresas. Algunos de los resultados observados eran esperados; sin embargo, otros no. Los métodos webmétricos se han empleado para analizar los patrones de enlaces a sitios web de empresas, ampliando la evidencia observada en la investigación previa. Hasta donde conocemos, esta es la primera aproximación a la investigación webmétrica, vinculada a variables contables y financieras, realizada desde el área de los estudios de empresa. Por lo tanto, una de las principales contribuciones de esta tesis es la de presentar y desarrollar nuevos métodos de investigación en este ámbito, partiendo de una perspectiva interdisciplinar.

De forma global, el objetivo de esta tesis es establecer un "enlace transversal", empleando el lenguaje técnico expuesto en páginas anteriores, entre los estudios de empresa y las Ciencias de la Información en relación con la investigación en la Web. Adicionalmente, hemos intentado mejorar la comprensión de los patrones de enlaces a sitios web comerciales, ampliando los estudios que se habían llevado a cabo previamente. Para conseguir este objetivo, hemos trabajado en tres líneas principales:

- verificando si las correlaciones encontradas entre el número de enlaces recibidos y las variables financieras existen en diferentes sectores, regiones y tipos de empresas;
- investigando si el análisis de co-enlaces permite generar nuevos conocimientos sobre la economía en su conjunto, no únicamente centrado en sectores particulares; y,
- combinando los métodos anteriormente señalados para proporcionar una perspectiva global de un sector empresarial específico.

Los objetivos de esta tesis se resumen en cinco preguntas de investigación que hemos intentado responder a través de los distintos apartados de la tesis.

- **Pregunta de investigación 1: ¿En qué medida se relacionan las variables financieras clave de una empresa con su presencia en la Web, medida a través del número de enlaces que reciben sus sitios web?**

Hemos respondido a la primera pregunta de investigación estudiando las relaciones entre el número de enlaces recibidos por los sitios web de empresas y sus variables financieras en diferentes regiones del mundo y en múltiples sectores de actividad. En términos generales, los datos de enlaces recibidos obtenidos se correlacionan con variables financieras, como ya apuntaron trabajos previos. Las correlaciones entre variables web y variables *offline* permiten reforzar las conclusiones que ya se habían sugerido en investigaciones anteriores. Trabajos previos (Vaughan, 2004a; 2004b; Vaughan y Wu, 2004) se limitaban al sector de tecnologías de la información y a tres países (China, Estados Unidos y Canadá). Ello hacía necesario extender la investigación a:

- una mayor variedad de sectores, y
- a otras regiones, especialmente Europa.

Además, la mayoría de empresas europeas incluidas en la base de datos Amadeus no están cotizadas en mercados de valores, a diferencia de lo que ocurre con las empresas de cuyos datos disponemos en Estados Unidos.

El estudio incluido en el Apartado 4.1.1 amplía la investigación previa y proporciona evidencia de correlaciones significativas en diversos sectores en los Estados

Unidos. El estudio analiza cinco sectores diferentes (Bancos, Construcción, Comercio general, Minería y Servicios), así como las empresas que componen el Dow Jones Industrial. Los resultados ponen de manifiesto que existen correlaciones significativas en sectores distintos al de Tecnologías de la información. Así, se concluye que los datos de hiperenlaces podrían emplearse como una variable significativa en múltiples sectores, no exclusivamente en aquellos que están centrados en información.

En el Apartado 4.3.1, analizamos los 50 principales bancos en el contexto internacional. El test de correlación de Spearman indica que la mayoría de los coeficientes de correlación entre enlaces recibidos y un conjunto de variables financieras (por ejemplo, activos totales, ingresos totales, resultado neto) son significativos al 1%. Solamente la variable ROA no es significativa. Los coeficientes de correlación para el sector bancario estadounidense son mayores que los del sector bancario internacional. Esto se explica por el contexto competitivo homogéneo que existe en los Estados Unidos frente a los mercados heterogéneos a los que pertenecen las empresas en este estudio. Los diferentes países tienen diferentes condiciones económicas y financieras, lo cual explica que los coeficientes de correlación sean menores cuando se analizan grupos de bancos pertenecientes a diferentes países. Otro factor significativo podría ser el grado en el que Internet se emplea con propósitos comerciales en los distintos países.

Finalmente, el estudio incluido en el Apartado 4.1.2 extiende la investigación en el contexto europeo. Se seleccionaron como países de referencia España y Reino Unido dado que representan dos grandes economías de la Unión Europea y sus dos lenguas se encuentran entre las más comúnmente empleadas. El estudio confirmó la evidencia hallada en los trabajos previos. También se ha descubierto que los datos de hiperenlaces reflejan algunas características propias de la economía de la Unión Europea. Por ejemplo, las correlaciones entre las variables de enlaces recibidos y las variables financieras no cambian significativamente cuando los datos de los dos países se consideran en conjunto, reflejando posiblemente el hecho de que comparten un mercado común. Cuando se comparan los resultados del estudio actual con los de estudios previos de otros países, como

los Estados Unidos, se han observado cuestiones que precisan una exploración adicional para alcanzar una comprensión más profunda de la relación entre el número de enlaces y las variables financieras. La mayoría de los estudios realizados hasta el momento indican que, solamente cuando las empresas pertenecen a un mismo sector, las correlaciones son significativas. Sin embargo, en este caso, el estudio en España y Reino Unido muestra correlaciones significativas cuando empresas de distintos sectores se consideran conjuntamente.

A partir de estos estudios, podríamos afirmar que el número de enlaces recibidos por el sitio web de una empresa está significativamente correlacionado con variables de posición financiera (activos totales) y, en menor medida, con variables de desempeño financiero (ingresos totales). El resultado y algunos ratios, tales como el ROA o el ROE, en la mayoría de los casos no presentan correlaciones significativas. Los resultados muestran que el número de enlaces recibidos se podría emplear como un indicador de variables económicas en términos absolutos, tanto variables de posición financiera como de desempeño financiero de la empresa. Sin embargo, no hay evidencia de que esto sea así cuando prestamos atención a variables financieras de carácter relativo, tales como el ROA. La explicación puede estar vinculada a la naturaleza de la variable "enlaces recibidos" (*inlink*), que representa el número de enlaces que apuntan a una página o sitio web desde su creación. Las funciones que ofrecen los motores de búsqueda, en este momento, no permiten recuperar información sobre enlaces que fueron creados durante un periodo determinado de tiempo. De algún modo, esto es similar a la naturaleza de las variables de posición financiera, que tienden a incrementarse a lo largo del tiempo por un fenómeno de acumulación. Las variables de desempeño financiero como los ingresos o el resultado tienden generalmente a incrementarse con el tiempo basadas en el desempeño de años anteriores. De acuerdo con la evidencia encontrada, los enlaces recibidos tienden a reflejar principalmente el tamaño de la empresa y, en menor medida, algunas medidas del rendimiento.

Finalmente, estos estudios nos permiten decir que las correlaciones significativas observadas en trabajos previos se verifican también en distintos sectores, fuera de lo que es el sector de las tecnologías de la información, así como en otras regiones

del mundo, en este caso Europa.

- **Pregunta de investigación 2: ¿Cuáles son las variables financieras que explican el número de enlaces que recibe el sitio web de una empresa?**

La segunda pregunta de investigación se ha abordado mediante el diseño y comprobación de un modelo explicativo del número de enlaces recibidos a partir de las variables financieras de empresas en España y en Reino Unido (Apartado 4.1.3). Se trata de un intento de explicar la variable "enlaces recibidos" por sitios web comerciales, utilizando para ello un análisis de regresión múltiple. Anteriormente, Vaughan y Thelwall (2005) aplicaron un análisis de regresión múltiple para explicar los enlaces que apuntaban a las universidades canadienses. Nuestro modelo ha sido aplicado a cinco sectores diferentes, alcanzando algunas conclusiones generales que dan respuesta a la pregunta de investigación:

- La variable "activos totales" (transformada empleando el logaritmo natural) es la variable más relevante para explicar el número de enlaces recibidos por los sitios web de las empresas. Solamente en el sector Editorial esta variable no es significativa. Consideramos que la variable debería aparecer de manera consistente en un modelo explicativo del número de enlaces recibidos por los sitios web de las empresas en la mayor parte de los sectores.

- La variable "activos intangibles" no resulta significativa en ningún sector. Se pueden sugerir diversas explicaciones: inconsistencias en el modo en que las empresas proporcionan esta información, dificultades de las normas contables para reconocer y medir los activos intangibles, o falta de relación entre la naturaleza de los activos intangibles reconocidos y medidos en la contabilidad y los activos intangibles de la empresa que están relacionados con la Web.

- La variable "cifra de negocios" se incluye como una variable significativa

para los sectores de la Construcción y Editorial; mientras que el "resultado después de impuestos" solamente en el caso del sector de Telecomunicaciones. También el país de origen de la empresa es una variable significativa en tres de los cinco sectores.

Para concluir, los "activos totales" (variable de posición financiera) y la "cifra de negocios" (variable de desempeño financiero) son las variables más relevantes para explicar el número de enlaces recibidos por los sitios web comerciales. También el país de origen parece ser una variable a tener en cuenta, aunque no está claro en el estudio si esta variable es representativa de un componente geográfico, lingüístico o bien una mezcla de ambos.

- **Pregunta de investigación 3: ¿Cómo se interrelacionan empresas pertenecientes a sectores de actividad distintos analizadas a través de su estructura de enlaces en la Web?**

La tercera pregunta de investigación ha sido abordada estudiando diferentes índices bursátiles e incluyendo empresas que pertenecen a diferentes sectores (Apartado 4.2.1). Esta investigación se basa en el análisis de co-enlaces. El análisis de co-enlaces en empresas se ha centrado en el estudio de sectores específicos (Vaughan y You, 2006) o en un análisis detallado de un subsector dentro de una industria determinada (Vaughan y You, 2008, 2009). La metodología empleada en estos artículos ha sido comprobada y verificada en diferentes países y sectores; por ejemplo, la industria química y la industria de la electrónica en China (Vaughan, Tang y Du, 2009). Al margen de estos trabajos en sitios web comerciales, cabe mencionar la investigación realizada sobre la teoría de la triple hélice (Stuart y Thelwall, 2006; García-Santiago y Moya-Anegón, 2009), la cual combina sitios web de diferentes ámbitos sociales y económicos. Esta teoría analiza la transferencia de conocimiento entre la empresa, la universidad y la administración pública, incluyendo entidades de naturaleza muy heterogénea.

El estudio amplía la aplicación del análisis de co-enlaces a sitios web de empresas de distintos sectores pertenecientes a cinco índices bursátiles. Se emplea el escalamiento multidimensional para elaborar mapas que representan las posiciones relativas de las empresas en los índices. Finalmente, se comparan los resultados de los cinco índices bursátiles, los cuales corresponden a diferentes contextos económicos, geográficos y culturales. Hay una variable principal que podría determinar explicar un patrón general que se observa en todos ellos; esta es, el grado en el que el modelo de negocio se centra en la información. Los sectores más intensivos en información forman grupos localizados en posiciones distantes del resto de empresas, que aparecen en posiciones centrales. Los sectores centrados en la información, de acuerdo con nuestra interpretación, comprenden principalmente cuatro grandes grupos: empresas basadas en la producción de contenidos (los medios de comunicación, *Media*, en el estudio), empresas que proporcionan servicios e infraestructuras relacionadas con la información (sector de tecnologías de la información, *IT*), empresas que aplican intensamente técnicas de comercio electrónico (sector del turismo, *Leisure*) y empresas del sector financiero. Estos sectores se hayan inmersos en un continuo proceso de transformación en sus modelos de negocio. Un claro ejemplo es el caso de la profunda crisis del sector de los medios de comunicación a raíz del proceso de digitalización de sus contenidos. Además de su posición en los mapas, las empresas de medios de comunicación y de tecnologías de la información reciben más enlaces que ningún otro tipo de empresas pertenecientes a sectores más tradicionales. Este patrón de enlaces refleja modelos de negocio fuertemente expuestos a Internet. La observación de los distintos mapas de los índices bursátiles tomados individualmente no nos permite confirmar la expectativa de que empresas de un mismo sector aparezcan agrupadas. Se trata de un aspecto que requiere realizar estudios adicionales.

En el trabajo, el Euro Stoxx 50 es el único índice que está compuesto por empresas de países diferentes. Este hecho nos permite observar si la integración económica en la Eurozona tiene algún tipo de reflejo en los datos de enlaces. Los resultados muestran que la variable "país de origen" de las empresas permite explicar de manera más precisa los *clusters* observados, frente a la variable "sector

empresarial". Como contraste, el mapa del Dow Jones está definido más claramente en base a los sectores de actividad, sugiriendo que existe una integración económica más profunda en el mercado de los Estados Unidos, como por otra parte es lógico pensar.

Los datos de co-enlaces constituyen una fuente de información para obtener nuevas perspectivas en el contexto de la empresa. El estudio de empresas pertenecientes a diferentes sectores podría permitir a la dirección de la empresa, analistas financieros y otros grupos de interés situar las actividades de sus empresas en relación a otras e identificar diferentes modelos de negocio, así como seguir su evolución a lo largo del tiempo. En el caso de que una empresa intentara incrementar su presencia en la red, el número de enlaces recibidos constituiría una medida más directa del éxito de los esfuerzos empresariales en alcanzar este objetivo. El análisis de co-enlaces podría ser empleado también por la dirección de la empresa para observar los progresos de la entidad en relación con otras empresas en el mismo sector o en diferentes sectores. Se trata de una herramienta que puede ser especialmente útil para las empresas que están en proceso de cambio de sus estrategias de tecnologías de la información vinculadas a la Web. Las conclusiones son de carácter principalmente exploratorio, si bien son relevantes para responder a la pregunta de investigación propuesta y para abrir nuevas direcciones en la investigación webmétrica en asuntos de empresa y económicos en general.

- **Pregunta de investigación 4: ¿Podemos emplear la información sobre enlaces para estudiar eventos económicos determinados?**

La cuarta pregunta de investigación ha sido abordada mediante el desarrollo de un análisis de co-enlaces para estudiar la evolución de la crisis financiera en los bancos de Estados Unidos cotizados en la New York Stock Exchange. Esta investigación se ha inspirado en un estudio reciente (Vaughan y You, 2008) que propuso un método que combina información sobre el contenido de las páginas web con datos de co-enlaces con el fin de alcanzar una visión más detallada del

panorama competitivo de un subsector dentro de una industria particular. En este caso, se ha aplicado el método al sector bancario con el objeto de estudiar el impacto de la actual crisis financiera. Las palabras clave seleccionadas para filtrar las páginas de co-enlaces fueron "crisis", "bailout" y "subprime", las cuales han sido empleadas frecuentemente para describir los problemas financieros de los bancos. El empleo de las palabras clave pretende, tanto para la captura de enlaces recibidos como de co-enlaces, excluir de la muestra aquellas páginas que no tienen relación con el tema indicado. Con este trabajo (Apartado 4.2.2), hemos intentado explorar si el método puede proporcionar información sobre la crisis financiera. Nuestro objetivo era que los datos pudieran ser utilizados para visualizar *clusters* de bancos con mayores dificultades financieras. Si bien los resultados preliminares, basados en datos de enero de 2009, ofrecían resultados positivos; una segunda ronda de datos en agosto de 2009 no ha permitido confirmar esta evidencia. Algunas de las razones que pueden explicar esta situación son:

- El análisis de co-enlaces puede no constituir un método apropiado para alcanzar los objetivos pretendidos. No todas las páginas web que enlazan a un banco con dificultades financieras enlazan simultáneamente a otro banco en una situación similar. Parece que un análisis basado únicamente en enlaces directos puede ser una mejor forma de afrontar la cuestión.

- La variable empleada para medir la crisis financiera (el dinero de rescate de activos recibido del gobierno federal) se encuentra sesgada por un efecto tamaño debido a que los bancos más grandes son bancos que lógicamente reciben mayores fondos y, por lo tanto, es posible que se esté midiendo el tamaño de la entidad en lugar de la profundidad de la crisis financiera.

En relación con esta pregunta de investigación, es preciso llevar a cabo trabajos adicionales en otros asuntos con el fin de evaluar la utilidad de esta perspectiva. Este estudio puede proporcionar indicaciones para evitar problemas futuros.

- **Pregunta de investigación 5: ¿Cómo se pueden combinar distintos tipos de análisis webmétrico para ofrecer una visión global de un sector empresarial?**

La quinta y última pregunta de investigación ha sido respondida a través de una combinación de métodos webmétricos: el análisis de impacto de enlaces y el análisis de co-enlaces dentro de un mismo estudio, con el objeto de analizar un sector determinado. Este trabajo (Apartado 4.3.1) se ha centrado en el sector bancario internacional, empleando diversas técnicas webmétricas. Primero, pretendemos averiguar si existe alguna correlación entre el número de hiperenlaces que el sitio web de un banco atrae y las variables financieras del citado banco. Segundo, empleamos el análisis de co-enlaces para generar unos mapas que muestren las posiciones competitivas de estos bancos en el mercado global en dos momentos de tiempo. El estudio ha alcanzado el objetivo de emplear una combinación de métodos para proporcionar una visión global de este sector. Se han observado correlaciones significativas entre enlaces recibidos y varias variables financieras. Adicionalmente, una comparación entre el número de enlaces recibidos por los bancos asiáticos frente a los otros bancos en la muestra indica que los primeros reciben un mayor número de enlaces. Esto es consistente con el hecho de que los bancos asiáticos han sido capaces de sortear la crisis financiera global con más éxito.

Los mapas de co-enlaces, realizados empleando el escalamiento multidimensional, indican que el sector bancario global está parcialmente organizado en mercados regionales, a pesar de la existencia de algunos actores globales principales debido al proceso de internacionalización de las actividades financieras. Algunos de ellos parecen encontrarse más aislados, como por ejemplo, los bancos chinos. Esta es posiblemente una de las razones por la que éstos no han sufrido tan seriamente como otras entidades la crisis financiera mundial. Los datos recogidos en un segundo momento revelan que estos bancos han cambiados de posición, desplazándose cerca de las principales entidades de Estados Unidos y del Reino Unido. Esto refleja la percepción cambiante del sector financiero global hacia los bancos chinos como resultado de su mayor resistencia a la crisis financiera

internacional.

Aunque deben de desarrollarse metodologías más sofisticadas e integradas para analizar empresas empleando distintos métodos webmétricos, este estudio muestra como diferentes procedimientos pueden proporcionar visiones complementarias acerca de un asunto económico determinado.

Limitaciones

A pesar de que la investigación ha proporcionado evidencia complementaria sobre aspectos anteriormente estudiados, así como ha planteado nuevas líneas de investigación futura en el campo de la empresa, podemos mencionar algunas limitaciones.

La naturaleza dinámica de la Web hace que, como fuente de datos, constituya un verdadero reto que afronta dificultades importantes, lo cual debe de ser tenido en cuenta a la hora de interpretar cualquier investigación basada en hiperenlaces. La investigación en la Web representa siempre una fotografía fija de un determinado momento y situación, cambiando a cada momento que transcurre. Esto hace, por ejemplo, que sea imposible llevar a cabo una investigación basada en enlaces referida a momentos pasados o, por otra parte, reproducir en un momento futuro los resultados obtenidos. Algunos de los problemas han sido puestos de relieve cuando abordamos las limitaciones de los motores de búsqueda como fuente de datos para la investigación webmétrica. Por lo tanto, es siempre discutible hasta qué punto los resultados obtenidos son generalizables.

El empleo de motores de búsqueda para obtener los datos de enlaces se ve afectado por las características de este sector, que consiste en un oligopolio con tres operadores principales: Google, Yahoo! y Bing (Microsoft). Desde una perspectiva global, estos tres copan la gran mayoría del mercado de las búsquedas, si bien, en determinadas áreas de la Web existen otros buscadores, como por ejemplo, Technorati, especializado en blogs. Esta cuestión ha sido tratada por algunos estudios en los últimos años (Lewandowski, Wahlig y Meyer-Bautor, 2006;

Evans, 2007). En 2009 Yahoo! y Microsoft firmaron un acuerdo que, a medio plazo, podría afectar a las funciones de consulta sobre enlaces que actualmente ofrece Yahoo!. Ello podría tener destacadas implicaciones en el desarrollo de la investigación webmétrica basada en datos de motores de búsqueda.

En relación con el análisis de co-enlaces, los estudios podrían presentar algunas limitaciones derivadas de la interpretación de los resultados, la cual podría diferir en función de los criterios empleados en el análisis o del conocimiento del investigador acerca del asunto en particular abordado en el estudio.

Investigación futura

A pesar de que la investigación ha permitido responder algunas preguntas acerca de sitios web comerciales y de los enlaces que reciben, también ha planteado nuevas cuestiones. El descubrimiento de nueva evidencia en las relaciones entre número de enlaces recibidos por los sitios web de las empresas y sus variables financieras refuerza las conclusiones previamente alcanzadas (Vaughan, 2004a; 2004b; Vaughan y Wu, 2004). Con todo, es preciso desarrollar modelos para ofrecer esta información de manera adecuada para su explotación empresarial. Por ejemplo, la elaboración de un modelo en línea con el análisis de regresión múltiple, llevado a cabo en el Apartado 4.1.3, se podría emplear para predecir el número de enlaces que un sitio web espera recibir basándose en sus variables financieras y en el sector al que pertenece. Esto podría ayudar a determinar qué empresas lo están haciendo mejor o peor en la Web. El número de enlaces recibidos, como medida de la presencia de las empresas en la Web, también podría usarse para cuantificar activos intangibles de la empresa relacionados con Internet.

El empleo del análisis de co-enlaces para estudiar empresas de distintos sectores presenta resultados positivos. Es preciso realizar trabajos adicionales para comprobar la utilidad del método con el objeto de identificar sectores que evolucionan hacia modelos de negocio más intensivos en información. Además, se debe explorar el empleo de nuevas técnicas de visualización de datos que permitan

extraer más información de los mismos.

El empleo conjunto del análisis de impacto de enlaces y del análisis de co-enlaces ha proporcionado también resultados positivos para el análisis de sitios web de empresas. Sin embargo, es preciso desarrollar un método sistemático para interpretar los resultados y combinar las metodologías, de modo que se proporcione una visión global más completa de un sector específico. Se deben explorar también variables webmétricas adicionales, como, por ejemplo, los enlaces salientes o los co-enlaces salientes o espacios web relacionados, para el estudio de sitios web de empresas.

Otra perspectiva que es prometedora es la de llevar a cabo estudios de tipo longitudinal para analizar la evolución de los sectores, índices bursátiles o empresas individuales. Esta perspectiva se ha introducido en los estudios realizados en los Apartados 4.2.2 y 4.3.1.

Es preciso llevar a cabo una investigación cualitativa para complementar la investigación cuantitativa, ya que proporciona evidencia confirmatoria acerca de la relevancia de los enlaces que están siendo analizados (Vaughan, Gao y Kipp, 2006). Esto también permite conocer la naturaleza de las páginas web, especialmente si corresponden a herramientas o servicios de la Web 2.0 o no. El fuerte desarrollo de la Web 2.0 (blogs, redes sociales, wikis, etc.) en los últimos cinco años ha cambiado significativamente el escenario de la Web, siendo posible medir su impacto a través de técnicas webmétricas. Una de las ventajas de estudiar la Web 2.0 es la posibilidad de usar fuentes alternativas para obtener datos, por ejemplo, *delicious* (un servicio para compartir enlaces favoritos), *Flickr* (una comunidad basada en la fotografía), *Youtube* (una comunidad basada en los vídeos), *Technoraty* (un motor de búsqueda especializado en blogs), etc. El desarrollo de la Web semántica y el empleo de APIs puede ampliar las opciones de investigación.

Finalmente, otra sugerencia para la investigación futura estriba en aplicar el análisis de impacto de enlaces y el análisis de co-enlaces para investigar las relaciones competitivas y de cooperación en otras áreas, tales como el sector público o los

partidos políticos, siendo especialmente interesante en el ámbito del gobierno electrónico o la democracia electrónica.