Características de la Web de España

Por **Ricardo Baeza-Yates**, **Carlos Castillo** y **Vicente López**

Ricardo Baeza-Yates, doctor en ciencias de la computación (Universidad de Waterloo, Canadá, 1989), es profesor Icrea en la Universitat Pompeu Fabra en Barcelona y director del Centro de Investigación de la Web en la Universidad de Chile. Es autor de numerosas publicaciones en las áreas de recuperación de información, tecnologías de búsqueda para la Web, bibliotecas digitales, algoritmos y estructuras de datos. Junto a B. Ribeiro-Neto publicó en 1999 el libro Modern information retrieval, un referente en la recuperación de información.

Carlos Castillo es doctor en ciencias de la computación (Universidad de Chile, 2004). Entre sus áreas de investigación está la recuperación de información en la Web, en particular el análisis de enlaces y caracterización de la Web. Actualmente lleva a cabo una investigación como asociado post doctoral en la Università degli Studi di Roma "La sapienza". Es fundador y socio de Newtenberg Ltd.

Vicente López, doctor en química cuántica (Universidad Autónoma de Madrid, 1981), es catedrático de ciencias de la computación e inteligencia artificial en la Universitat Pompeu Fabra, Barcelona, donde es también director de la Cátedra Telefónica de Producción Multimedia, y director general del Barcelona Media-Centre d'Innovació. Hasta 1989 fue investigador en química física y desde 1989 en el campo de la inteligencia artificial, el procesamiento de información en sistemas naturales, y tecnología de la producción multimedia. Es autor de más de 50 artículos de investigación.

Resumen: *En septiembre y octubre de 2004 se llevó a cabo una recolección masiva de páginas de la Web de España. Esto permitió encontrar más de 16 millones de páginas en alrededor de 300.000 sitios, que fueron analizadas en términos de contenido, conectividad y tecnologías. El análisis de estos datos entrega numerosas similitudes estadísticas con las observaciones sobre las características de la Web de otros estudios realizados sobre subconjuntos de la Web global, fundamentalmente relacionados con la presencia de leyes de potencia. En el plano cualitativo se aprecia que el rol de centros académicos y de la administración pública es fundamental para articular la conectividad de la navegación entre sitios de la Web de España.*

Palabras clave: *Caracterización de la Web, Análisis de enlaces, Web de España.*

Title: Characteristics of the web of Spain

Abstract: *In September and October 2004, a massive harvesting of pages from the web of Spain was carried out. This effort permitted us to find over 16 million pages, on approximately 300,000 web sites, which have been analysed in terms of contents, connectivity and technologies. The analysis of the data reveals several statistical similarities with observations made about the characteristics of the web by other studies conducted on subsets of the global web, mostly related to the presence of power laws. On the qualitative side, the role of academic and governmental centres is fundamental for enhanced connectivity in navigating among the sites of the web of Spain.*

Keywords: *Web characterisation, Link analysis, Web of Spain.*

***Baeza-Yates**, **Ricardo**; **Castillo**, **Carlos**; **López**, **Vicente**. "Características de la Web de España". En: El profesional de la información, 2006, enero-febrero, v. 15, n. 1, pp. 6-17.*

1. ¿Cómo es la Web?

Una de sus grandes ventajas es la capacidad de relacionar información mediante vínculos o enlaces, que permiten a los usuarios una gran flexibilidad en el momento de buscar información. Por esto, el modelo de web que nos planteamos es de grafo dirigido, en el que cada página es un nodo y cada arco representa un enlace –con una dirección marcada con una flecha– entre dos páginas.

Estos enlaces no están puestos al azar. Las páginas normalmente tienen vínculos hacia otras del mismo tema (**Davison**, 2000) y las mejores tienden a ser más referenciadas que el promedio. La web como grafo tiene una estructura que se puede clasificar como una red li-

Artículo recibido el 06-10-05
Aceptación definitiva: 27-10-05

bre de escala que, al contrario que las redes aleatorias, se identifica por una distribución sesgada de sus enlaces. Estas redes han sido el tema de una serie de estudios, entre los que cabe resaltar por su claridad el de **Barabási** (2002). Se caracterizan por ser redes en las que la probabilidad de que una página tenga *k* enlaces es proporcional a k^{-t}, donde *«t»* es un número real positivo (típicamente entre 1 y 3) que es el parámetro de la distribución. Esto se llama una ley potencial o de potencia (*power-law*).

«La Web de España está compuesta por más de 300.000 sitios que contienen más de 16 millones de páginas. Muchas de sus características son muy similares a las de la Web global»

En las redes libres de escala, unas pocas páginas tienen muchos enlaces, mientras que la mayoría tienen muy pocos. Esta distribución la encontraremos en la Web en casi todos los aspectos y es la misma que encontró el economista **Vilfredo Pareto** en 1896 para la distribución de la riqueza: el 80% de la misma estaba repartida entre el 20% de la población. También es idéntica a la que encontró **George Kingsley Zipf** en 1932 para la frecuencia de las palabras en los textos, y que más tarde resultó ser aplicable a muchos otros dominios (**Zipf**, 1949). Al representar gráficamente estas distribuciones en escala logarítmica aparece una línea recta, tal como se observa en muchos de los gráficos de este estudio.

1.1. Estudiando la Web de un país

Las redes libres de escala son, a su vez, auto-similares, en el sentido de que una pequeña muestra tiene propiedades de la totalidad. Es el caso de la Web de España, que presenta características muy parecidas a la red mundial, a pesar de contener menos de 2/1.000 de las páginas disponibles en el mundo, si consideramos que la Web indexable contiene al menos 11 mil millones de páginas (**Gulli**; **Signorini**, 2005).

Una Web nacional es el conjunto de páginas relacionadas con un país. En términos técnicos es difícil distinguir perfectamente cuando una página pertenece a un cierto país, por lo cual utilizamos algunas heurísticas: decimos que una página es de España si su dirección IP está asignada a alguna red físicamente en España, o si el nombre del dominio al que pertenece termina en *.es*. Esto nos permite conseguir muchas más que las que obtendríamos si sólo tuviéramos en cuenta el sufijo del dominio, ya que de acuerdo con nuestras observaciones, solamente el 16% de los dominios con páginas de España están bajo *.es*.

En los últimos años se ha estudiado la Web de más de diez países con distintas metodologías (**Baeza-Yates**; **Castillo**, 2005b). Específicamente respecto a la Web de España existe un estudio en profundidad sobre 27 sitios específicos de universidades e instituciones públicas (**Alonso**, et al., 2003), uno sobre 64 universidades (**Thelwall**, 2004, capítulo 13), otro sobre 500 sitios escogidos al azar (**Amat**, 2003) y un estudio preliminar sobre una muestra grande de sitios (**Baeza-Yates**, 2003), aproximadamente la mitad de los que analizamos aquí.

1.2. Recolección de páginas

La colección fue obtenida entre los meses de septiembre y octubre del 2004 utilizando un programa desarrollado por *Akwan* (**da Silva**, et al., 1999). Esta aplicación comienza descargando un conjunto de direcciones iniciales, que en nuestro caso fue obtenido a partir de las referencias incluidas en el antiguo buscador *Buscopio (http://www.buscopio.net/)*. Luego extrae de las páginas descargadas enlaces a otras nuevas, y continúa recursivamente realizando esta operación mientras se cumplan los criterios que enunciamos más arriba.

Páginas web	16.171.267
Texto total	43 GB
Texto promedio por página	2.855 B
Sitios web	308.822
Páginas promedio por sitio	52,08
Texto promedio por sitio	146 KB
Dominios	118.248
Sitios promedio por dominio	2,61
Páginas promedio por dominio	136,75
Texto promedio por dominio	373 KB

Cuadro 1. Resumen de la colección estudiada

El cuadro 1 resume las características principales de la colección obtenida. En todos los casos, cuando se menciona el tamaño de una página nos referimos a su contenido en términos de texto, sin las marcas html, puesto que ésta fue la información que almacenamos.

1.3. La Web como colección documental

La web es una colección descentralizada en la que distintos autores pueden aportar contenido independientemente y sin una instancia de control que decida qué se publica y qué no. Es su ventaja más importante, pero también la principal causa de dificultades tan-

to para buscar información como para caracterizar colecciones de páginas extraídas desde ella.

Hemos detectado varias anomalías que constituyen violaciones de estándares o situaciones especiales que dificultan la caracterización de las páginas. Esto incluye errores en la implementación de html (el lenguaje de marcado de las páginas), de http (el protocolo de comunicaciones) y de *DNS* (el servicio de resolución de nombres), entre otros.

Además, existen numerosas réplicas de colecciones completas de gran volumen. La información replicada se estima entre un 20% y un 40% del total en la Web (**Fetterly**, et al., 2005). Como consecuencia directa es que hay muchos sitios con el mismo contenido. Además, como estas colecciones normalmente son bastante grandes, los sitios tienen una cantidad de texto bastante mayor que lo esperable.

Dado que hay un incentivo económico para aparecer en los primeros lugares en las máquinas de búsqueda, existen numerosos sitios que practican el *spam*. Es un término que en este contexto se aplica a acciones maliciosas orientadas a engañar a los sistemas de búsqueda en la Web y dar a algunas páginas una posición más alta que la que merecen en el resultado de una búsqueda (**Gyöngyi**; **García-Molina**, 2005). Estas acciones incluyen cambios en el texto, los metadatos y/o los enlaces de las páginas.

«Los sitios con mayor cantidad de texto público son las Cortes de Castilla-La Mancha, la Universidad de Barcelona, la empresa EuroVia, la organización RedIris y el diario El mundo»

Reconocer fehacientemente qué es *spam* es un área activa de investigación; hoy en día se estima que hasta un 8% de lo que indexan las máquinas de búsqueda en la Web lo es, siendo uno de los signos destacados más obvios la presencia de regularidades en la distribución de ciertas variables estadísticas (**Fetterly**, et al., 2004). Por ejemplo, en la figura 6 se observa que hay grupos de páginas que comparten una conectividad sospechosamente similar, y efectivamente al inspeccionarlas manualmente se observa que muchas de ellas son generadas automáticamente para mejorar fraudulentamente la posición de otras en los buscadores.

Otra técnica usual para esto es el uso de servidores de nombres de dominio configurados para producir cientos o miles de nombres distintos para el mismo contenido, a esto se le llama *DNS wildcarding*, o *DNS comodín* (**Barr**, 1996). En la Web de España constatamos que esta técnica es usada con bastante frecuencia.

«La mayoría de los sitios con gran cantidad de información en la Web de España son réplicas de documentación o centros de documentación gubernamentales y académicos»

En lo que sigue, presentamos nuestras observaciones en este contexto a varios niveles: de páginas (sección 2); de sitios (sección 3), y de dominios (sección 4). Por último incluimos nuestras propias conclusiones.

2. Páginas

Presentamos el análisis de las páginas individualmente, sin considerar su agrupación en sitios o dominios web.

2.1. Títulos de las páginas

Examinamos esta información y encontramos que más de un 9% de las páginas no lo tiene y un 3% tiene alguno por omisión como *Untitled document*, *Documento sin título*, *Página nueva 1*, etc.

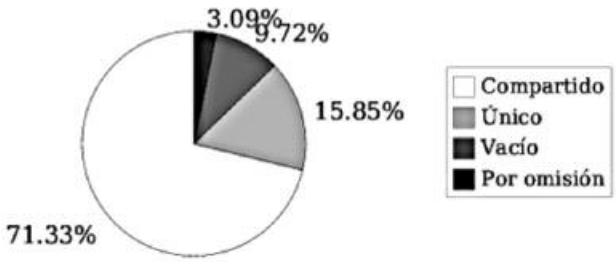

Figura 1. Distribución del título de las páginas

La figura 1 muestra la distribución de los títulos de las páginas. Lo más común es que una página tenga un título, pero que no sea único, por lo tanto, es muy probable que no sea suficientemente descriptivo de su contenido. Muchas máquinas de búsqueda dan más importancia a las palabras que aparecen en el título de las páginas como términos de recuperación, por lo que cuando este dato no aparece o es poco descriptivo tendrá menos posibilidades de ser encontrada en un buscador que otra idéntica pero con un título apropiado.

Observamos que, en promedio, un mismo título válido es compartido por aproximadamente 4 páginas; siendo similar a lo percibido en la Web de Portugal: un título distinto por cada 5 páginas (**Gomes**; **Silva**, 2003). Además, normalmente un sitio grande puede tener la misma proporción de títulos distintos que uno pequeño, por lo que la calidad de los sitios en este sentido es, en términos generales, independiente de su tamaño.

2.2. Texto en las páginas

Después de extraer el texto, almacenamos solamente los primeros 300 KB de cada página. Si representamos gráficamente la distribución de los tamaños, obtenemos el gráfico de la figura 2, aunque es difícil apreciar la distribución. Sin embargo, si utilizamos una escala logarítmica en ambos ejes (figura 3) podemos ver los datos con mayor claridad y observamos que hay muchas páginas con muy poco texto y unas pocas con un tamaño enorme.

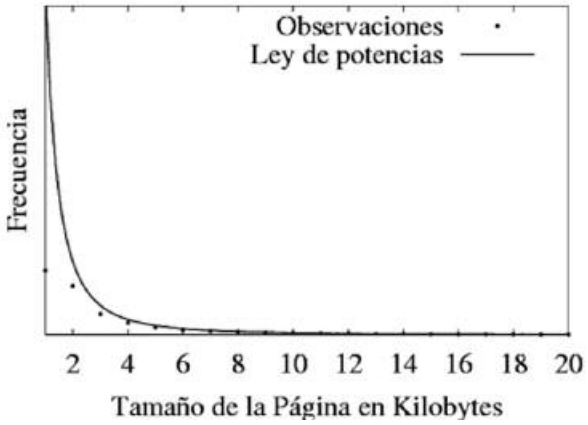

Figura 2. Distribución del tamaño de las páginas

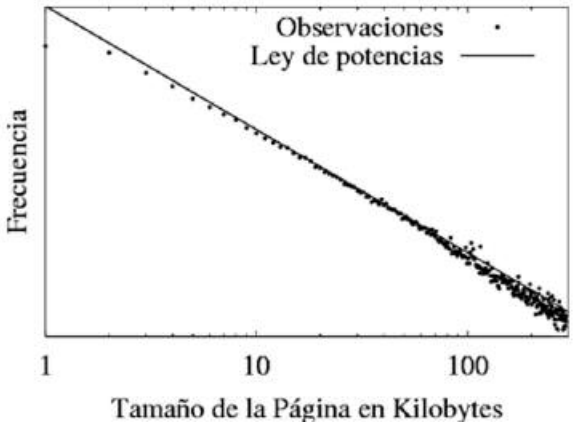

Figura 3. Distribución del tamaño de las páginas, usando escala logarítmica

Una ley de potencia con parámetro 2,25 se ajusta muy bien a la distribución de los tamaños de las páginas. Al inspeccionarlas manualmente, se puede ver que varias que contienen poco texto corresponden a sitios web construidos principalmente con elementos gráficos como imágenes o animaciones, mientras que las más grandes son o bien índices generados automáticamente o largos textos de temáticas diversas (legales, técnicos, etc.).

2.3. Idioma

Se utilizó un sistema estadístico de análisis de textos llamado *Bow*[1] que permite hacer, entre otras cosas, una clasificación basada en n-gramas utilizando *Naïve Bayes*. El sistema se entrenó con las lenguas más frecuentes en la Web más los idiomas propios de España: catalán, gallego y euskera, lográndose clasificar aproximadamente un 64% de las páginas (el resto mayoritariamente contiene muy poco texto, que no permite este tipo de detección de idioma). La distribución para estas páginas aparece en la figura 4.

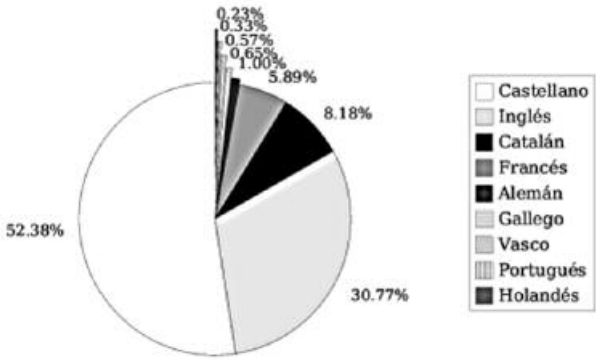

Figura 4. Distribución del idioma en el que están escritas las páginas

La proporción total de páginas escritas en los idiomas oficiales de España es de aproximadamente 62%. Esto lo podemos poner en contexto comparando con otros países de los cuales tenemos la siguiente información:

—Tailandia: 34% en tailandés.

—Portugal: 73% en portugués.

—Brasil: 75% en portugués.

—Chile: 90% en castellano.

Las páginas en inglés en la Web de España se corresponden mayoritariamente a copias de documentación sobre lenguajes de programación o sistemas computacionales, así como a sitios relacionados con turismo.

2.4. Vocabulario

Analizamos 1 GB de texto extraído de páginas en cada uno de los idiomas. La distribución de la frecuencia de las palabras sigue una ley de potencias con parámetro 0,7 para el inglés y cercano a 0,8 para el castellano y el catalán. Los términos más frecuentes naturalmente son en su mayoría *stopwords* o palabras funcionales, que no tienen significado por sí mismas. En el cuadro 2 se muestran solamente las palabras más frecuentes que son sustantivos. En el caso del castellano son prácticamente las mismas que las que aparecían en 2002 (**Baeza-Yates**, 2003). Es interesante notar que el nombre del país resulta ser un término habitual en este tipo de muestras, como se ha observado previamente en Brasil (**Veloso,** et al., 2000) y Chile (**Baeza-Yates**; **Castillo**, 2000). Este dato en relación a las páginas en inglés indica que se trata en su mayoría de documentación técnica; en cuanto al catalán, nos

muestra una fuerte presencia de páginas relacionadas con universidades o centros de enseñanza.

Castellano	Inglés	Catalán
Artículo	Java	Barcelona
Información	String	Catalunya
Trabajo	Meted	Informació
Ley	Directory	Universitat
Servicio	Name	Servei
Madrid	Object	Departament
Año	File	Treball
Universidad	Information	Centre
Forma	Server	Dia
España	Data	Curs

Cuadro 2. Sustantivos más frecuentes en la Web (incluyendo nombres propios)

2.5. Documentos que no están en html

Se localizaron aproximadamente 200.000 enlaces a ficheros no-html, lo que si bien es un número grande, representa sólo un 1% de las páginas totales en la Web. Los formatos texto plano (típicamente con extensión *.txt*) y *Adobe pdf* son los más usados, y juntos corresponden a más del 80% de los documentos que no están en html. La distribución se muestra en la figura 5.

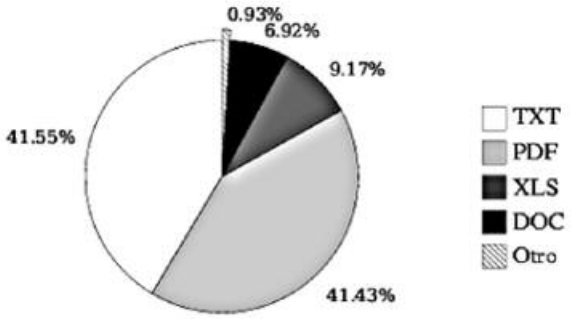

Figura 5. Distribución de enlaces a documentos, excluyendo enlaces a páginas html

Pdf también es el más usado para documentos que no son html en Austria, Brasil, Chile, Corea del Sur y Portugal, con porcentajes que van desde el 45% al 65% de los documentos no escritos en html (**Rauber**, et al., 2002; **Modesto**, et al., 2005; **Baeza-Yates**; **Castillo**, 2005a; **Baeza-Yates**; **Lalanne**, 2004; **Gomes**; **Silva**, 2003). A pesar de que *Microsoft Windows* es el sistema operativo más usado, las extensiones asociadas con aplicaciones de *Microsoft Office* como *Word* o *Excel* en total suman sólo alrededor del 16% de los ficheros. Es decir, los estándares abiertos son preferidos por los autores de las páginas para publicar documentos en la Web.

2.6. Enlaces entre páginas web

El número de enlaces que recibe una página web se llama su "grado interno", nombre que proviene del hecho de que estamos tratando la Web como un grafo. Por otro lado, a la cantidad de vínculos que salen de ella se le llama su "grado externo". La distribución de ambas cantidades se muestra en la figura 6.

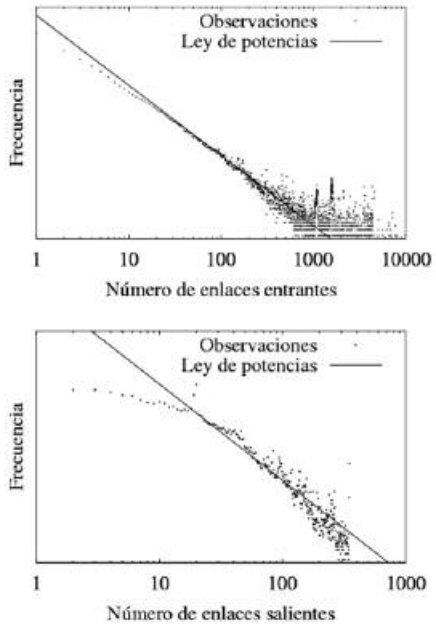

Figura 6. Número de enlaces hacia y desde cada página

El grado interno de una página es una medida de su popularidad en la Web, mientras que el externo refleja más bien una opción de diseño. En otras palabras, tener una página en la que aparezcan muchos enlaces es fácil, pero recibirlos desde otras en gran cantidad es más difícil. Todo ello se refleja en la distribución estadística.

«Alrededor del 50% de las páginas están en castellano, seguidas de 30% en inglés y 8% en catalán. El contenido en gallego y vasco constituye aproximadamente el 2%»

Al ajustar una distribución de **Zipf** a los datos, se obtienen los parámetros 2,11 para el grado interno y 2,84 para el externo. Eso es similar a los valores que se observan para estos conceptos en otros subconjuntos de la Web, siendo los valores más usuales 2,1 y 2,7 (**Pandurangan,** et al., 2002); los rangos usuales de estos parámetros para otros son 1,5-2,0 para el grado interno, y entre 2,6-4,1 para el externo (**Boldi**, et al., 2002; **Modesto**, et al., 2005; **Baeza-Yates**; **Castillo**, 2005a; **Baeza-Yates**; **Lalanne**, 2004).

También se puede observar que existen ciertas anomalías en términos de grupos de páginas que comparten el mismo grado interno o externo. Estos aparecen como "saltos" en la frecuencia de páginas para números de enlaces bastante altos. Usualmente constituyen *spam*, como ha sido comentado anteriormente (**Fetterly**, et al., 2004; **Thelwall**; **Wilkinson**, 2003).

3. Sitios web

Los definimos como un conjunto de páginas que comparten el fragmento del nombre del servidor de la url. Así, *http://ejemplo.es/paginaA.html* y *http://ejemplo.es/paginaB.html* pertenecen ambas al sitio *ejemplo.es*.

3.1. Número de páginas

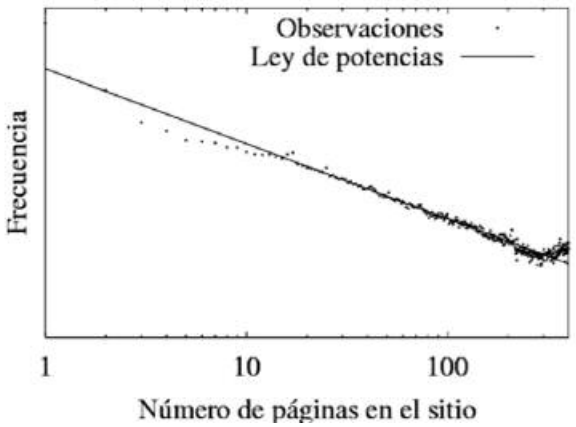

Figura 7. Número de páginas por sitio web

Observamos un promedio de 52 por sitio. Su distribución es muy sesgada, como se muestra en la figura 7. Alrededor de 400 páginas por sitio hay una disminución en la frecuencia de los sitios, puesto que el recolector fue configurado para extraer 400 páginas máximo de los sitios *.com*. Ajustando una ley de potencias en la parte central de la distribución se obtiene el parámetro 1,14, que se puede comparar con el de otros países en los cuales este dato está en el rango 1,6-2,5 (**Modesto**, et al., 2005; **Baeza-Yates**; **Castillo**, 2005a; **Baeza-Yates**; **Lalanne**, 2004), infiriéndose que en la Web de España hay comparativamente una mayor cantidad de sitios muy grandes.

3.2. Sitios que tienen sólo una página

Amat (2003) manifestó que al tomar una muestra de sitios al azar, aproximadamente sólo un 34% tenía algún tipo de contenido. En la misma línea, nosotros observamos que solamente un 40% tenía más de una página, y analizamos el motivo por el cual el recolector no pudo encontrar más en ellos. Nos dimos cuenta que principalmente se trata de sitios construidos utilizando *Java*, *Flash* o *Javascript*, o sea que es necesario interpretar estos lenguajes para poder navegar por ellos, tarea que pocos recolectores de páginas web realizan. En otros casos, la página es sólo una redirección o abre recuadros (*frames*) que se encuentran en otro sitio. La distribución de casos se muestra en la figura 8. La proporción de sitios que efectivamente sólo constan de una página, sin ningún enlace, es cercana al 30%.

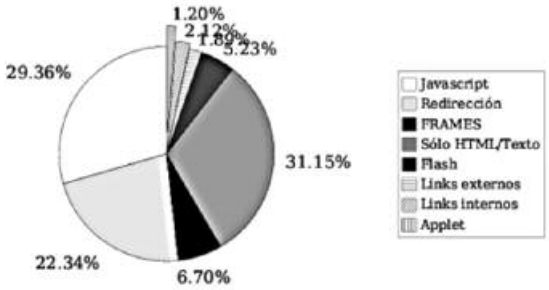

Figura 8. Distribución de los sitios que tienen sólo una página detectada por el recolector

Es importante el hecho de que en la Web de España existen por lo tanto alrededor de 90.000 sitios que utilizan únicamente *Javascript* o *Flash* en su portada y por tanto son difíciles o imposibles de indexar por los sistemas de búsqueda actuales. Es decir, un 30% son de complicado acceso para las máquinas de búsqueda. Dado que la consulta es un mecanismo fundamental para encontrar información de la Web, esos sitios tienen menos presencia y por tanto menos visitantes que otros.

3.3. Enlaces entre sitios web

Un enlace entre dos sitios web representa uno o varios enlaces entre sus páginas, manteniendo la dirección del enlace. Esto significa que si existe al menos uno entre, por ejemplo, *http://www.A.es/paginaA.html* y *http://www.B.es/paginaB.html*, entonces diremos que existe un enlace entre *www.A.es* y *www.B.es*. Los enlaces internos, a páginas dentro del mismo sitio, no son considerados. Esto se ha llamado también *hostgraph* o grafo de servidores (**Dill**, et al., 2002).

«Los sustantivos que más aparecen en páginas web incluyen los nombres de las ciudades 'Madrid' (11%) y 'Barcelona' (7%)»

Existen 122.190 sitios con más de una página. De ellos, 77.718 (63%) no reciben ninguna referencia desde otro sitio de España, y 109.787 (90%) no tienen ningún enlace hacia otro sitio de España. Esto es compatible con lo observado en **Amat** (2003) en cuanto a que 2/3 de los sitios están aislados respecto al resto de la Web y el porcentaje es un poco mayor que en Chile (**Baeza-Yates**; **Castillo**, 2005a). La distribución del

grado interno y externo en los sitios web también revela una red libre de escala, como se muestra en la figura 9.

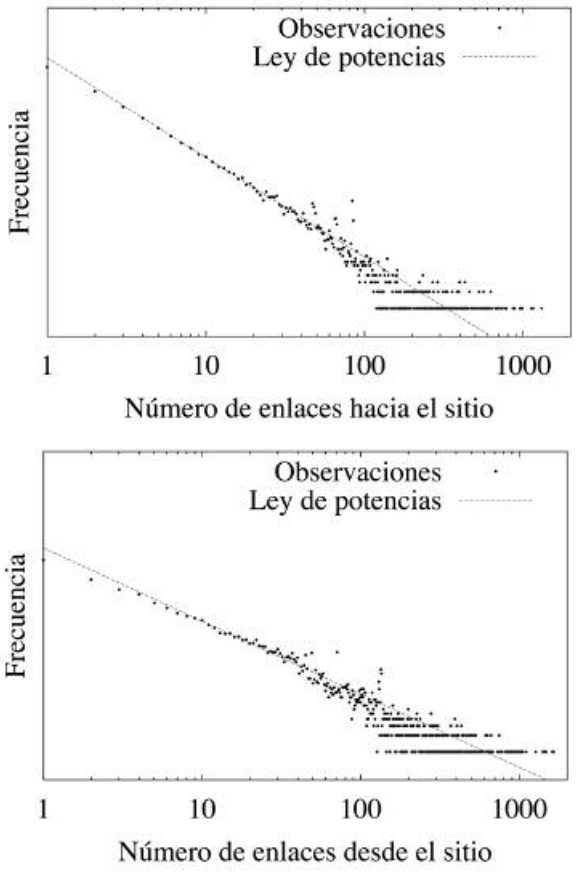

Figura 9. Distribución del número de enlaces entre sitios

Los parámetros del ajuste de la ley de potencias son 1,82 y 1,34 para grado interno y externo respectivamente, siendo los rangos observados en otras colecciones 1,2-2,1 para el grado interno y cercano a 1,8 para el grado externo. En el caso de la Web global, una estimación de este parámetro para el grado interno es 2,34 (**Dill**, et al., 2002).

3.4. Estructura macroscópica

En un grafo se dice que una parte es una componente conexa si es posible ir desde cualquier nodo de esa parte a otro dentro de la misma parte; se la denomina fuertemente conexa si esto es posible respetando la dirección de los enlaces Dentro de una parte fuertemente conexa es posible ir desde cualquier sitio a cualquier sitio siguiendo enlaces. No toda la Web de España es fuertemente conexa, es decir, si escogemos dos sitios al azar, no siempre habrá un camino desde el primero hasta el segundo siguiendo enlaces.

Las redes libres de escala como la Web presentan un componente fuertemente conexo gigante, que puede ser usado como el punto de partida para distinguir ciertos componentes estructurales y que fueron definidos por **Broder**, et al. (2000):

—*Main*: los sitios en la componente fuertemente conexa.

—*Out*: los sitios que son alcanzables desde *Main*, pero que no tienen enlaces hacia *Main*.

—*In*: los sitios que pueden alcanzar a *Main*, pero que no tienen enlaces desde *Main*.

—*Islas*: no son accesibles ni hacia ni desde *Main*.

—*Tentáculos*: sólo se conectan con *In* o *Out*, pero en el sentido inverso de los enlaces.

—*Túnel*: una componente que une *Out* e *In* sin pasar por *Main*.

«Los sitios .es tienen más contenido, están mucho mejor conectados y presentan bastante menos spam que los sitios de España en otros dominios como .com o .net»

En **Baeza-Yates** y **Castillo** (2001) extendimos esta notación distinguiendo en *Main* los siguientes subcomponentes:

—*Main-Main*: son los sitios que pueden ser alcanzados directamente desde *In* o que pueden alcanzar directamente *Out*.

—*Main-In*: pueden ser alcanzados directamente desde *In* pero no están en *Main-Main*.

—*Main-Out*: sitios que pueden alcanzar directamente a *Out* pero no pertenecen a *Main-Main*.

—*Main-Norm*: no pertenecen a las subcomponentes definidas anteriormente.

La figura 10 muestra todos los componentes mencionados anteriormente.

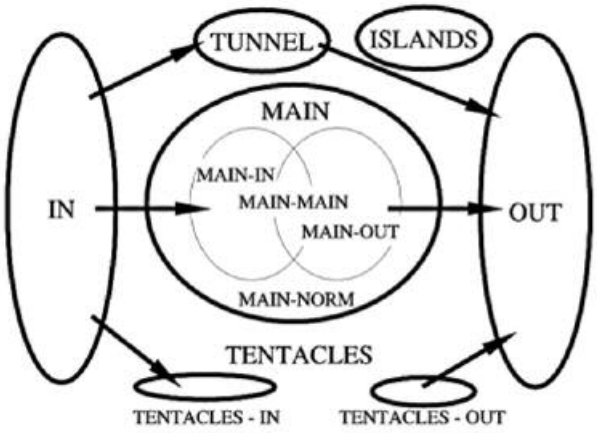

Figura 10. Estructura macroscópica de la Web

Componente	Sitios	Páginas
Main-In	0,8%	1,2%
Main-Main	4,2%	25,6%
Main-Norm	3,0%	3,4%
Main-Out	7,1%	27,0%
Main (total)	15,0%	57,1%
In	2,6%	3,2%
Out	74,1%	36,4%
Tentáculos in	5,9%	1,6%
Tentáculos out	1,0%	0,6%
Túnel	0,3%	0,6%
Islas	1,2%	0,6%

Cuadro 3. Distribución de sitios en las componentes de la Web. La distribución de páginas indica qué porcentaje de las páginas están en los sitios de cada componente

La distribución de sitios web en componentes se muestra en el cuadro 3. Nótese que los sitios web en *In* e *Islas* se encuentran sólo si se conoce a priori la dirección de su página principal, puesto que no son alcanzables siguiendo enlaces. Además, se incluye el porcentaje sobre el total de sitios. Finalmente, incluimos también la distribución del número de páginas en los sitios de cada componente.

Componente	Total de sitios	*.es*	*.com*	*.net*	*.org*	Otro
In	2%	100%	0%	0%	0%	0%
Main (total)	14%	100%	0%	0%	0%	0%
Out	74%	23%	55%	8%	12%	2%
Tentáculos in	6%	25%	61%	7%	4%	3%
Tentáculos out	1%	100%	0%	0%	0%	0%
Túnel	0%	100%	0%	0%	0%	0%
Islas	1%	47%	45%	5%	1%	2%

Cuadro 4. Distribución de los dominios en que están los sitios de cada componente. La primera columna muestra la proporción de sitios en cada componente. Las siguientes columnas indican cuántos de esos sitios están en cada dominio (los valores pueden no sumar 100% debido a que están aproximados al entero más cercano)

También estudiamos en qué dominio residen los sitios de cada componente. El resultado se muestra en el cuadro 4; lo que más se destaca es que el 100% de los sitios en el *Main* de nuestra colección de páginas* están alojados bajo *.es*, mientras que los otros dominios son mucho más frecuentes, por ejemplo, para la componente *Out*. Esto tiene una interpretación muy clara: si bien una gran cantidad de dominios de España se registran bajo dominios genéricos como *.com*, *.net*, etc., *.es* concentra la mayoría de los sitios bien conectados, y por lo tanto constituye el eje central de la Web de España.

4. Dominios

Definimos el dominio de una página como un sufijo de su nombre de sitio web, siguiendo la siguiente regla: si es de la forma *www.A.es* o *www.xxx.A.es*, entonces el dominio es *A.es*.

Para dominios en los cuales lo usual es utilizar nombres de tercer o cuarto nivel, se hizo una excepción. Esto ocurre con dos proveedores españoles de subdominios gratuitos bajo *.es.vg* y *.es.fm*. Además, *.uk* no admite registros directamente, por lo que la mayoría de los sitios se ubican bajo *.co.uk*; existen numerosos casos que utilizan esta extensión, no sólo por la relación comercial y diplomática entre España y el Reino Unido, sino también debido a la presencia de sitios asignados a Gibraltar. Por último, *.eu.int* es utilizado por numerosas instituciones de la Unión Europea que operan en España. Para los efectos de este estudio, estos dominios son considerados de primer nivel. Por ejemplo, en *www.A.co.uk* el dominio es *A.co.uk*.

En total se encontraron 118.248 dominios distintos.

4.1. Número de sitios por dominio

A pesar de que un 92% de los dominios tienen un sólo sitio, el promedio de sitios por dominio es cercano a 2,6. Esta aparente contradicción se explica porque hay dominios muy grandes, por ejemplo hay casi 30 con más de 1.000 sitios cada uno. La distribución del número de sitios para cada uno de los 10.000 dominios más grandes se muestra en la figura 11.

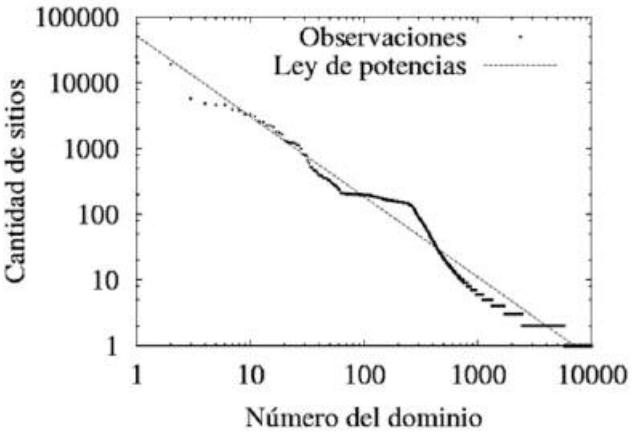

Figura 11. Distribución del número de sitios por dominio

Al inspeccionarlos manualmente, se observa que prácticamente todos los dominios con más de 100 sitios, excepto muy pocas excepciones, son creados para producir *spam*, y muchos de ellos usan *DNS Wildcard*, un sistema de nombres de dominio con comodín, o están configurados para entregar la misma dirección

* La colección está sesgada por las direcciones iniciales que en su mayoría son .es.

sea cual sea el nombre usado. Por ejemplo: *http://X.bcnlink.com/* para cualquier secuencia "X" retorna siempre una dirección IP válida y que apunta a la misma página o a otras con ligeras variaciones.

4.2. Tamaño total de los dominios

Aproximadamente un 25% de los dominios españoles tienen solamente una página, y el promedio es de 133 por dominio. Este número es mucho menor que el 60% de sitios que tienen sólo una página, y posiblemente se debe a que crear un sitio nuevo una vez que se tiene el nombre de dominio no tiene ningún costo, mientras que tener un dominio nuevo sí lo tiene y, una vez que se ha pagado, resulta natural instalar más sitios web para rentabilizar la inversión realizada.

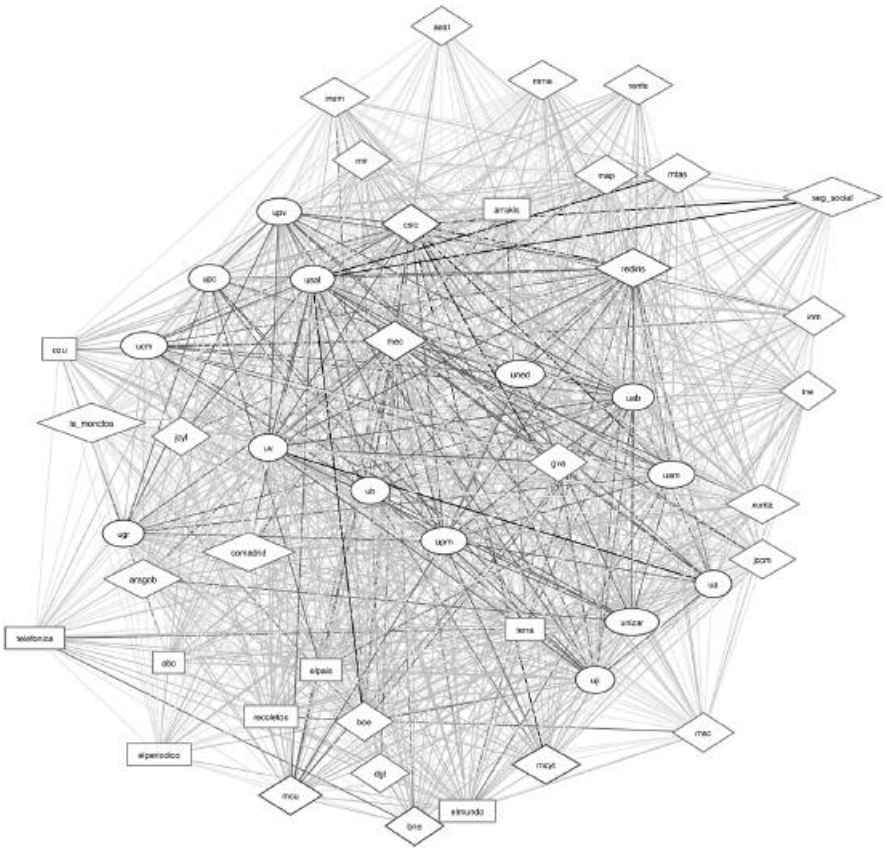

Figura 12. Representación gráfica de los enlaces entre dominios

El tamaño promedio de un dominio web completo considerando solamente el texto es de aproximadamente 370 KB. La distribución del tamaño total de páginas por dominios sigue una ley de potencias con parámetro 1,19 en su parte central.

Una gran mayoría (25 de los 30 dominios más grandes) son de universidades, centros de investigación o bases de datos para uso académico. Esto es similar al caso de Chile (**Baeza-Yates**; **Castillo**, 2005a) y de Tailandia (**Sanguanpong**, et al., 2000) donde también existe una fuerte presencia de sitios de tipo académico. El caso contrario es el de Corea del Sur, alli la mayoría de los sitios son de tipo comercial (**Baeza-Yates**; **Lalanne**, 2004).

Detectamos además que existe una gran abundancia de réplicas o *mirrors* de documentación. Por ejemplo, encontramos 6 réplicas completas de las notas técnicas *RFC*, 7 copias íntegras del proyecto *LuCAS* (*Linux en castellano*), 30 de toda la documentación de *Tomcat* de *Apache* y 36 del proyecto de documentación de *Linux* (*LDP*), entre otros.

Comercial		Universidades		Gobierno	
Adobe	843	U. Complutense de Madrid	381	Boletín Oficial del Estado	595
El Mundo	520	U. Politècnica de Catalunya	342	M. de Educación y Ciencia	518
El País	503	U. Politécnica de Madrid	341	Consejo Superior Inv. Científicas	452
Terra	500	U. d'Alacant	338	Generalitat Valenciana	448
ABC	400	U. de Zaragoza	323	M. Trabajo y Asuntos Sociales	394
Arrakis	380	U. de Barcelona	314	M. Ciencia y Tecnología	380
Recoletos	351	U. de Valencia	312	Instituto Nacional de Estadística	376
Ozu	317	U. Jaume I	306	M. Administraciones Públicas	374
El Periódico	296	U. Autónoma de Madrid	305	RedIris	364
Telefónica	295	U. Politècnica de Valencia	294	M. Cultura	354

Cuadro 5. Dominios con mayor número de referencias, separados por tipo de dominio. El número que aparece junto a cada nombre corresponde a la cantidad de referencias

4.3. Enlaces entre dominios

A continuación medimos este dato con el propósito de obtener una representación gráfica de las relaciones entre dominios. Para dibujarla utilizamos el programa *Neato* del paquete *Graphviz*[2].

En la figura 12 hemos incluido los 50 dominios que reciben más enlaces en la Web de España.

Para mayor claridad, los hemos dividido en tres grupos: comercial (rectángulos), educacional (elipses) y administración pública (rombos); una línea más oscura significa un mayor número de enlaces. En el grafo se aprecia que dominios del mismo tipo tienden a agruparse juntos, incluso si consideramos que en algunos ámbitos no es usual enlazar a sitios similares; por ejemplo, entre los de los medios de comunicación normalmente hay muy pocos enlaces.

«Los dominios que reciben más enlaces internamente son: Adobe, BOE, El mundo, Mº de Educación y Ciencia y El país. El resto más referenciados son mayoritariamente universidades y sitios gubernamentales»

El dominio *adobe.es* es el más enlazado, principalmente por la página para descargar el programa *Acrobat Reader*, que además es la más enlazada de la Red global[3]. Dado que este vínculo no indica una relación entre los dominios más allá que la descarga de este software, no ha sido incluido en la figura 12.

A diferencia del año 2002 (**Baeza-Yates**, 2003), se observa una mayor presencia de sitios de la administración pública entre los primeros lugares. Los dominios más referenciados de cada tipo se muestran en el cuadro 5. Los medios informativos y proveedores de acceso son los que más destacan entre los sitios de carácter comercial. En el caso de las universidades, la diferencia en el número de enlaces entre el primer y el décimo lugar es apreciablemente menor que en los otros grupos.

4.4. Enlaces a otros dominios

Encontramos enlaces a aproximadamente 50 millones de sitios distintos fuera de España. Para cada sitio web externo encontrado obtuvimos el nombre del dominio de primer nivel, que usualmente corresponde a un código de país, o alguno de los nombres genéricos *.com*, *.net*, etc.

La mitad de los sitios externos enlazados desde la Web de España están alojados en *.com*; en cuanto a los genéricos *.org*, *.info* y *.biz* aparecen con bastante más frecuencia que la que correspondería al número de servidores en cada uno de estos dominios. Respecto a los códigos de país, los 20 más referenciados se muestran en el cuadro 6. Los porcentajes están calculados sobre el número de sitios totales referenciados, incluyendo las referencias a dominios genéricos de primer nivel.

Dominio	Nombre	Fracción de sitios
tk	Tokelau	3,25%
uk	Reino Unido	3,18%
de	Alemania	3,13%
it	Italia	1,85%
fr	Francia	1,20%
ca	Canadá	0,91%
nl	Holanda	0,90%
ch	Suiza	0,82%
jp	Japón	0,79%
us	EUA	0,67%
se	Suecia	0,58%
cl	Chile	0,57%
be	Bélgica	0,48%
dk	Dinamarca	0,47%
pt	Portugal	0,44%
au	Australia	0,42%
ru	Rusia	0,41%
at	Austria	0,37%
pl	Polonia	0,33%
no	Noruega	0,32%

Cuadro 6. Los 20 países más referenciados desde España

Un estudio similar de conectividad que incluyó varios países (**Bharat**, et al., 2001) y realizado a nivel de sitios mostraba que los sitios externos referenciados desde *.es* en 2001 estaban principalmente en Alemania, Reino Unido, Francia y el dominio *.int* de organizaciones internacionales.

«El 63% de los sitios web estudiados no es apuntado por otro sitio web de España, lo que los hace más difíciles de encontrar»

Finalmente tomamos datos de la *United Nations Statistics Division*[4] acerca del volumen de exportaciones españolas a otros países y comparamos esto con el número de enlaces encontrados, excepto los dominios *.com*, *.net* y las importaciones y exportaciones a empresas en EUA, que posiblemente los utilizan con más frecuencia que *.us*. Existe una relación significativa entre el número de enlaces y el intercambio comercial cuando ambos se observan en una escala logarítmica. En tal caso, el coeficiente de correlación es 0,5 entre

enlaces e importaciones y 0,6 entre enlaces y exportaciones. Esta última relación es apreciable a simple vista, como se muestra en la figura 13.

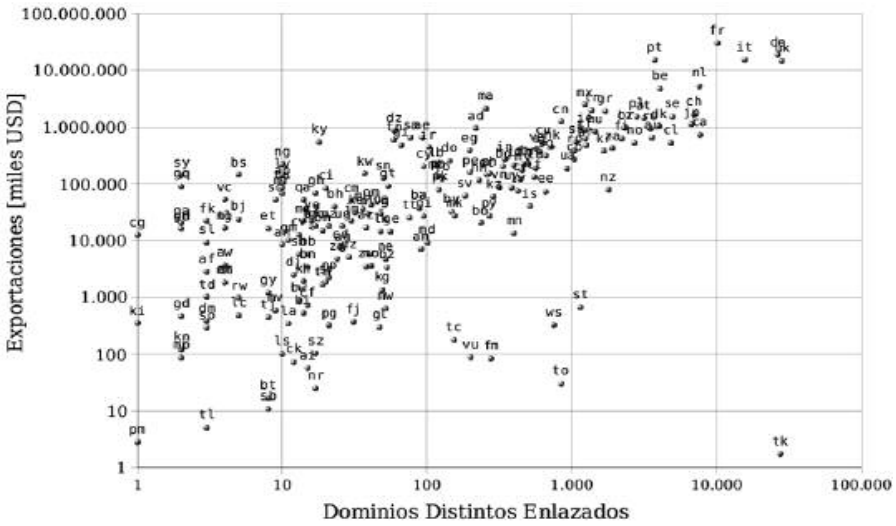

Figura 13. Relación entre el número de enlaces externos desde sitios web españoles y la suma total de exportaciones de empresas españolas

Se observan algunos rasgos destacados en esta figura. Los primeros tanto en exportaciones como en número de enlaces son algunos de los socios comerciales más importantes de España: Alemania, Reino Unido, Italia y Francia. También existe un grupo de nombres de dominio que tienen muchos enlaces, pero que no responden a la lógica de exportaciones a esos países. Estos son *.tc*, *.vu*, *.fm*, *.ws*, *.st*, *.to* y particularmente *.tk*, que son usados en general por su facilidad para ser recordadas en castellano o en otros idiomas.

5. Conclusiones

Un estudio como el que hemos presentado tiene varias aplicaciones, siendo la más directa la que tiene que ver con el desarrollo de mejores sistemas de búsqueda en la Web, en particular, con la implementación de mejores estructuras de datos para almacenar información sobre los metadatos o enlaces entre páginas web, y para jerarquizar los resultados de una consulta.

Nuestros datos también sirven para mostrar la heterogeneidad de la Web que, por un lado, es positiva debido a su diversidad, pero además es negativa, dada su calidad, por la presencia de numerosos sitios aislados, con poco contenido, con pocas referencias, etc.

«Los formatos más usados para documentos exceptuando html, son Adobe pdf y texto simple, cada uno con aproximadamente un 40% de participación»

Una particularidad que destaca en la Web de España es que una gran cantidad de sitios no utilizan el dominio de primer nivel correspondiente al país (*.es*), prefiriendo *.com* o *.org*. Mientras que por una parte el hecho de que *Es.Nic* –el organismo registrador de .es, propiedad ahora de la empresa estatal Red.es– solicite una serie de requisitos para la inscripción de un nombre ha contribuido a mantener el dominio *.es* relativamente libre de malas prácticas, por otra se usa mucho menos de lo esperado. Con la reciente liberalización del dominio *.es* esto sin duda cambiará, aunque quizá resulte irrelevante para los usuarios de la Web de España, puesto que nuestra impresión es que muy pocos usuarios están atentos a la barra de direcciones del navegador para diferenciar entre navegar por una página *.com* o *.es*. Además, no disponemos de datos respecto al número de visitas recibidas por cada sitio, pero sí podemos inferir que debido a que los sitios de España en *.es* tienen más contenido y están en las componentes mejor conectadas de la Web, seguramente reciben más visitas.

«Los países más referenciados desde España son Alemania, Reino Unido, Italia, Francia y Canadá. Existe una fuerte correlación entre el intercambio comercial con estos países y el número de enlaces»

Otra particularidad es que una gran cantidad de la información disponible es generada por universidades u organismos de la administración pública. Prácticamente todos los dominios más referenciados pertenecen a estas dos categorías, y esto es particularmente notorio en las páginas escritas en catalán. La prensa escrita también tiene una participación importante tanto en número de páginas como en cantidad de referencias.

Hemos encontrado que existe en la Web de España gran cantidad de información que puede tener diversos usos. Durante el presente estudio se encontraron, por ejemplo, más de 250.000 páginas en catalán que sirvieron para crear *CucWEB*[5], un corpus del catalán en la Web anotado con información lingüística.

Otro uso es la minería de datos en la Web, en particular su relación con la sociedad, por ejemplo en términos de la relación encontrada entre el intercambio comercial y las estructuras de enlaces a otros países.

Finalmente, es posible observar en esta muestra propiedades estadísticas muy similares a los de otras,

lo que indica que puede ser usada para estudios que sean al menos parcialmente extrapolables a la realidad de la red global.

Agradecimientos

María Eugenia Fuenmayor y **Paulo Golgher** realizaron la operación del recolector durante el proceso de descarga de las páginas. La clasificación de páginas por idiomas fue llevada a cabo por **Bárbara Poblete**, **Gemma Boleda**, **Stefan Bott** y **Toni Badia**. Financiado por la *Cátedra Telefónica de Producción Multimedia-Universitat Pompeu Fabra*.

Una versión extendida de los datos presentados en este artículo está disponible en línea en:
http://www.catedratelefonica.upf.es/webes/2005/

Notas

1. *Bow* es un conjunto de herramientas para modelamiento estadístico del lenguaje, recuperación de textos, clasificación y clustering. Consultado en octubre 2005.
http://www.cs.cmu.edu/~mccallum/bow/

2. Verificado en octubre 2005
http://www.graphviz.org/

3. *Internet Archive*, comunicación personal.

4. *Commodity trade statistics database*, división de estadísticas de las *Naciones Unidas*. Consultado en octubre 2005.
http://unstats.un.org/unsd/comtrade/

5. Consultado en octubre 2005
http://www.catedratelefonica.upf.es/

Bibliografía

Alonso, J.; **García, L.**; **Zazo, F.** *Cibermetría: nuevas técnicas de estudio aplicables al web*. España: Ediciones Trea, 2003. Isbn 84-9704-114-3.

Amat, C. B. «Caracterización de una muestra de sedes web españolas bajo dominio .es». En: *Boletín de la RedIris*, 2003, abril, n. 64, pp. 2014.

Baeza-Yates, R. «The web of Spain». En: *Upgrade*, 2003, v. 3, n. 3, pp. 82-84.

Baeza-Yates, R.; **Castillo, C.** «Caracterizando la Web chilena». En: *Encuentro chileno de ciencias de la computación*, 2000.

Baeza-Yates, R.; **Castillo, C.** «Relating web characteristics with link based web page ranking». En: *String processing and information retrieval (Spire*), 2001, pp. 21-32.

Baeza-Yates, R.; **Castillo, C.** «Características de la Web chilena 2004». Informe técnico, Center for web Research, Universidad de Chile, 2005(a).

Baeza-Yates, R.; **Castillo, C.** «Characterization of national web domains». Informe técnico, Universitat Pompeu Fabra, 2005(b).

Baeza-Yates, R.; **Lalanne, F.** «Characteristics of the korean web». Informe técnico, Korea-Chile IT Cooperation Center ITCC, 2004.

Barabási, A. *Linked: the new science of networks*. EUA: Perseus Books Group, 2002, Isbn 0-738-20667-9.

Barr, D. *RFC 1912: common DNS operational and configuration errors*, 1996. Consultado en: 01-10-05.
http://www.ietf.org/rfc/rfc1912.txt

Bharat, K.; **Chang, B. W.**; **Henzinger, M.**; **Ruhl, M.** «Who links to whom: mining linkage between web sites». En: *International conference on data mining (ICDM)*, 2001.

Boldi, P.; **Codenotti, B.**; **Santini, M.**; **Vigna, S.** «Structural properties of the african web». En*: Poster session, eleventh international conference on world wide web*, 2002.

Broder, A.; **Kumar, R.**; **Maghoul, F.**; **Raghavan, P.**; **Rajagopalan, S.**; **Stata, R.**; **Tomkins, A.**; **Wiener, J.** «Graph structure in the web: experiments and models». En: *Ninth conference on world wide web*, 2000, pp. 309-320.

da Silva, A. S.; **Veloso, E. A.**; **Golgher, P. B.**; **Laender, A. H. F.**; **Ziviani, N.** «Cobweb-a crawler for the brazilian web». En: *String processing and information retrieval (Spire)*, 1999, pp. 184-191.

Davison, B. D.»Topical locality in the web». En: *23rd annual international ACM Sigir conference on research and development in information retrieval*, 2000, pp. 272-279.

Dill, S.; **Kumar, R.**; **Mccurley, K. S.**; **Rajagopalan, S.**; **Sivakumar, D.**; **Tomkins, A.** «Self-similarity in the web». En: *ACM transactions on internet technology*, 2002, v. 2, n. 3, pp. 205-223.

Fetterly, D.; **Manasse, M.**; **Najork, M.** «Spam, damn spam, and statistics: using statistical analysis to locate spam web pages». En: *Seventh workshop on the web and databases (webDB)*, 2004, pp. 1-6.

Fetterly, D.; **Manasse, M.**; **Najork, M.** «Detecting phrase-level duplication on the world wide web». En: *28th annual international ACM Sigir conference on research and development in information retrieval*, 2005, pp. 170-177.

Gomes, D.; **Silva, M. J.** «A characterization of the Portuguese web». En: *3rd ECDL workshop on web archives*, 2003.

Gulli, A.; **Signorini, A.** «The indexable web is more than 11.5 billion pages». En: *Poster session, 14th international conference on world wide web*, 2005, pp. 902-903.

Gyöngyi, Z.; **García-Molina, H.** «Web spam taxonomy». En: *First international workshop on adversarial information retrieval on the web*, 2005.

Modesto, M.; **Pereira, Á.**; **Ziviani, N.**; **Castillo, C.**; **Baeza-Yates, R.** «Um novo retrato da web brasileira». En: *XXXII Semish*, 2005, pp. 2.005-2.017.

Pandurangan, G.; **Raghavan, P.**; **Upfal, E.** «Using Pagerank to characterize web structure». En: *8th Annual international computing and combinatorics conference (Cocoon)*, 2002, pp. 330-390.

Rauber, A.; **Aschenbrenner, A.**; **Witvoet, O.**; **Bruckner, R. M.**; **Kaiser, M.** «Uncovering information hidden in web archives". En: *D-Lib magazine*, 2002, v. 8, n. 12. *Doi: 10.1045/december2002-rauber*

Sanguanpong, S.; **Nga, P. P.**; **Keretho, S.**; **Poovarawan, Y.**; **Warangrit, S.** «Measuring and analysis of the thai world wide web». En: *Asia Pacific advance network conference*, 2000, pp. 225-230.

Thelwall, M. *Link analysis: an information science approach.* EUA: Elsevier Academic Press, 2004, Isbn 0-12-088553-0.

Thelwall, M.; **Wilkinson, D.** «Graph structure in three national academic webs: power laws with anomalies». En: *Journal of the American Society for Information Science and Technology*, 2003, v. 54, n. 8, pp. 706-712.

Veloso, E. A.; **de Moura, E.**; **Golgher, P.**; **da Silva, A.**; **Almeida, R.**; **Laender, A.**; **Ribeiro-Neto, B.**; **Ziviani, N.** «Um retrato da Web brasileira». En: *Simpósio brasileiro de computacão*, 2000.

Zipf, G. *Human behavior and the principle of least effort: an introduction to human ecology*. EUA: Addison-Wesley, 1949, Isbn 0-58-220471-3.

***Ricardo Baeza-Yates**, profesor ICREA, Departamento de Tecnología, Universitat Pompeu Fabra.*

ricardo.baeza@upf.edu

***Carlos Castillo**, **Vicente López**, Cátedra Telefónica de Producción Multimedia, Departamento de Tecnología, Universitat Pompeu Fabra.*

carlos.castillo@upf.edu

vicente.lopez@upf.edu