Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información.

José Luis Alonso Berrocal (berrocal@gugu.usal.es)
Carlos G. Figuerola (figue@gugu.usal.es)
Ángel Francisco Zazo Rodríguez (afzazo@gugu.usal.es)

Universidad de Salamanca. Facultad de Documentación.
C/ Francisco Vitoria, 6-16, 37008 - Salamanca

0.1 Resumen

Los sistemas más habituales de Recuperación de la Información se basan en conseguir una representación homogénea y procesable de documentos y consultas, y en el cálculo subsiguiente de alguna función que exprese el grado de similaridad entre una consulta dada y cada uno de los documentos de una colección. Por regla general, tales representaciones toman como base las palabras o términos que aparecen en los documentos. Las páginas web, sin embargo, contienen hiperenlaces, lo que sugiere la posibilidad de representar dichas páginas en función de tales hiperenlaces, en lugar de términos. Se describe un experimento exploratorio efectuado para estimar la viabilidad de esta hipótesis. Los resultados de dicho experimento sugieren que los hiperenlaces pueden ser aplicados ventajosamente en la representación de las páginas web, pero también que es preciso definir y aplicar mecanismos correctores.

Palabras clave: Recuperación de la información. World Wide Web. Internet. Modelo Vectorial.

0.2 Abstract

The most habitual systems in Information Retrieval are based both on getting a homogeneous and easily processed representation of documents and searches, and on the subsequent calculation of some function that expresses the degree of similarity between a given search and each one of the documents in a collection. Generally, such representations take the words or terms that appear in the documents as base. Web pages, however, contain hyperlinks, which suggests the possibility of representing such pages as a function of hyperlinks, instead of terms. We describe an exploratory experiment carry out to estimate the viability of this hypothesis. The results of this experiment suggest that hyperlinks can be applied advantageously in the representation of web pages, but it is also necessary to define and apply corrective mechanisms.

Keywords: Information Retrieval. WWW. Internet. Vectorial Model.

1. Introducción

La base de los diversos sistemas de Recuperación de Información, independientemente del modelo teórico subyacente, consiste en la implementación de algún formalismo que permita representar cada uno de los documentos y las posibles consultas que los usuarios puedan formular al sistema. La resolución de esas consultas consiste en la computación de alguna función de similaridad que compare la representación de una consulta dada con las representaciones de los documentos, y establezca el grado de adecuación entre ambos (Salton, 1987).

De hecho, esto es lo que se hace incluso en sistemas manuales o semimanuales, como puede ser el típico programa de gestión del catálogo de una biblioteca. La operación de catalogar un libro no es más que la elaboración de una representación del mismo, aplicando (manualmente) un formalismo determinado.

Naturalmente, la bondad de unos sistemas frente a otros –en lo referente a efectividad en la recuperación– depende de la mayor o menor capacidad del formalismo utilizado para representar adecuadamente cada documento, así como de las características de la función de similaridad utilizada, aunque ésta última vienen determinada por las características del formalismo de representación.

En cualquier caso, en los sistemas de recuperación automáticos (Luc, 1998) dicha representación suele basarse en las palabras o términos que aparecen en los documentos y/o en las consultas, formuladas en lenguaje natural. Dichos términos pueden seleccionarse en función de diversos planteamientos, y valorarse o pesarse basándose en diferentes mecanismos o criterios; pero son dichos términos los elementos básicos utilizados para representar los documentos (Salton, 1983; Rijsbergen, 1979).

En este sentido, parece evidente que cualquier página web puede ser considerada un documento, y que puede ser representada aplicando cualquiera de los modelos de recuperación existentes, tomando como base el texto que forma parte de dicha página.

Ahora bien, en las páginas web no sólo hay texto; además de imágenes, sonido, elementos de captación de datos (por ejemplo, formularios) y otras diversiones, existen hipervínculos o enlaces con otras páginas o, en general, con otros recursos disponibles en la red. La existencia de tales enlaces es precisamente lo que confiere su particular carácter a cada página web, en el sentido de que la hace diferente de un documento convencional.

A partir de estos enlaces el espacio web puede ser considerado como un grafo dirigido, en el cual los nodos serían las diferentes páginas existentes, y los arcos los hipervínculos que enlazan un nodo con otro (Ellis, 1994). Al ser un grafo dirigido (un hipervínculo se activa en un nodo determinado y nos dirige hacia otro nodo concreto), podemos distinguir entre enlaces o arcos entrantes y salientes. Así, si hacemos abstracción del contenido interno de cada nodo o página, podríamos definir cada uno de ellos en función de su situación en el grafo, es decir, sobre la base de los enlaces que mantiene hacia otros nodos y a los que otros nodos mantienen con él.

En consecuencia, podría plantearse representar una página web –desde el punto de vista de su posible recuperación– basándose en los enlaces de dicha página, en lugar de hacerlo a partir del texto de la misma, como habitualmente hacen la mayor parte de los buscadores tipo Lycos, Altavista y otros (Almind, 1997; Larson, 1996;Woodruff, 1996). Naturalmente, esto elimina información importante (la que aparece en forma de texto) que no sería utilizada en la recuperación. Sin embargo, dado que los enlaces no suelen apuntar de forma caprichosa, podríamos pensar que dos páginas que apuntan hacia los mismos nodos deben tratar de temas similares (Joachims y otros, 1995).

Las ventajas de tales planteamientos, en caso de ser viables, parecen claras: de un lado, tendríamos una reducción importante de los recursos de máquina necesarios para la recuperación, dado que en general las páginas suelen tener bastante menos enlaces que términos. Por otro, permitiría recuperar la información de manera independiente del idioma, tanto de las páginas o documentos como del propio usuario que formula la consulta. Asimismo, dado que se representan enlaces y no el texto, se evitaría la picaresca de muchas páginas web que repiten intencionadamente una o varias palabras, y que muchos buscadores interpretan como más relevantes.

2. Descripción del experimento

A partir de estos presupuestos hemos llevado a cabo un experimento tendente a sondear la viabilidad y las posibilidades de tal planteamiento, así como los posibles problemas que pudieran sobrevenir.

Para dicho experimento hemos utilizado una colección documental (nuestro espacio de búsqueda) constituido por 99.546 páginas web, recogidas de forma automática por un pequeño

robot a partir de dominios de instituciones académicas y de investigación españolas (Alonso Berrocal, 1997). De esta colección se seleccionaron 200 páginas cuya misión fue la de servir como consultas, es decir, como modelos de los cuales era necesario recuperar las páginas más similares.

A la colección así formada se le aplicó el modelo vectorial clásico (Salton, 1983), constituyendo vectores de cada una de las 99.546 páginas con los enlaces salientes de las mismas. Los elementos de cada vector o enlaces se pesaron utilizando el esquema estándar (Salton y Buckley, 1988) de

$$Fe \times IDP$$

Considerando IDF como

$$log_2 \frac{N}{ne} + 1$$

donde,

Fe es la frecuencia del enlace en la página,
N es el número total de páginas en la colección y
ne es número de páginas en que aparece el enlace

A su vez, la función de similaridad aplicada es la típica del coseno, utilizada ampliamente en operaciones de recuperación de información (Harman, 1992):

$$SIM(X,Y) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \times \sum_{i=1}^{n} Y_i^2}}$$

donde,

X es el vector de la consulta,
Y es el vector del documento,
X_i es el elmento i de X,
Y_i es el elemento i de Y,
N es el número de elementos o términos en los vectores.

Las 200 consultas fueron realizadas mediante la utilización de una versión adaptada del software experimental *Karpanta* (Figuerola, 1999).

3. Evaluación de los resultados

Por lo que se refiere a la evaluación de los resultados, hay que indicar que se trabajó exclusivamente sobre la precisión, toda vez que resulta imposible conocer el número total de páginas relevantes para cada una de las 200 consultas en todo el espacio de búsqueda considerado. Además, se tuvieron en cuenta los primeros 50 documentos recuperados para cada consulta.

Para determinar las relevancias de esos primero 50 documento recuperados se aplicaron las estimaciones de 5 personas, las cuales examinaron de forma independiente los resultados de todas las consultas. Finalmente, se consideraron como relevantes aquellos documentos recuperados que obtuvieron al menos 3 votos.

Los resultados globales quedan reflejados en el gráfico 1. Aunque es difícil hacer una estimación rigurosa de dicho gráfico al carecer de puntos de referencia homologables, parece, sin embargo, que tales resultados podrían considerarse como muy aceptables: valores iniciales superiores a 0.8, que se mantienen por encima del 0.5 durante toda la curva, es decir, valores altos y relativamente constantes.

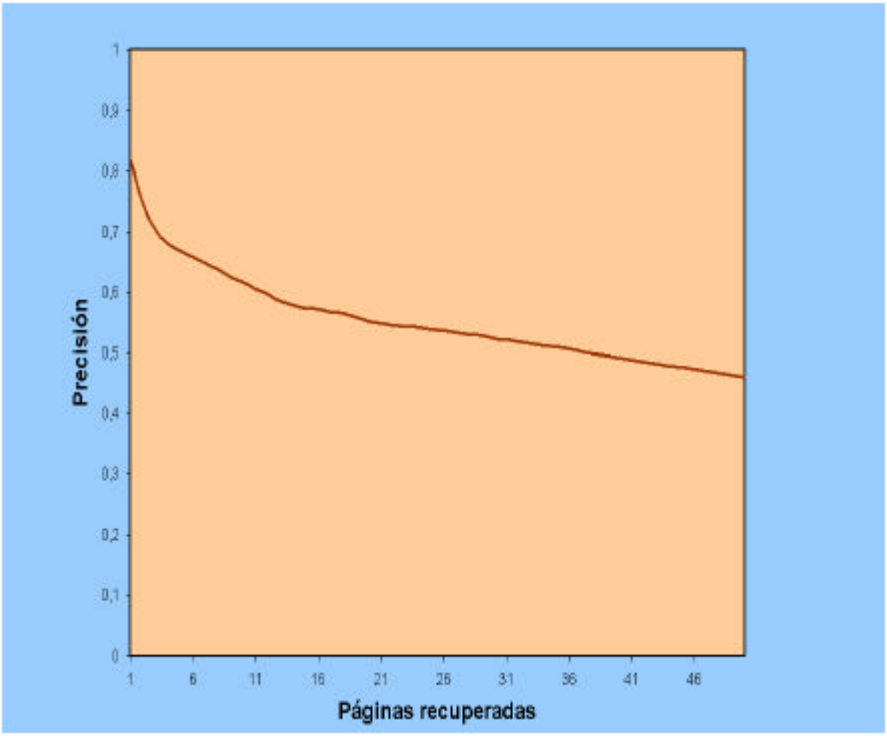

Figura 1. Precisión de los resultados de las búsquedas.

Un examen un poco más a fondo de las páginas recuperadas, muestra no obstante, factores que deben tenerse en cuenta y que matizan considerablemente la bondad de los resultados obtenidos en precisión. En efecto, el 83.2% de las 50 primeras páginas relevantes recuperadas pertenecen al mismo dominio que la utilizada como consulta (figura 2). Esto no quiere decir que no se recuperen páginas relevantes de otros dominios, pero se hace en menor cuantía y en puestos más avanzados, es decir, con un índice de similaridad menor.

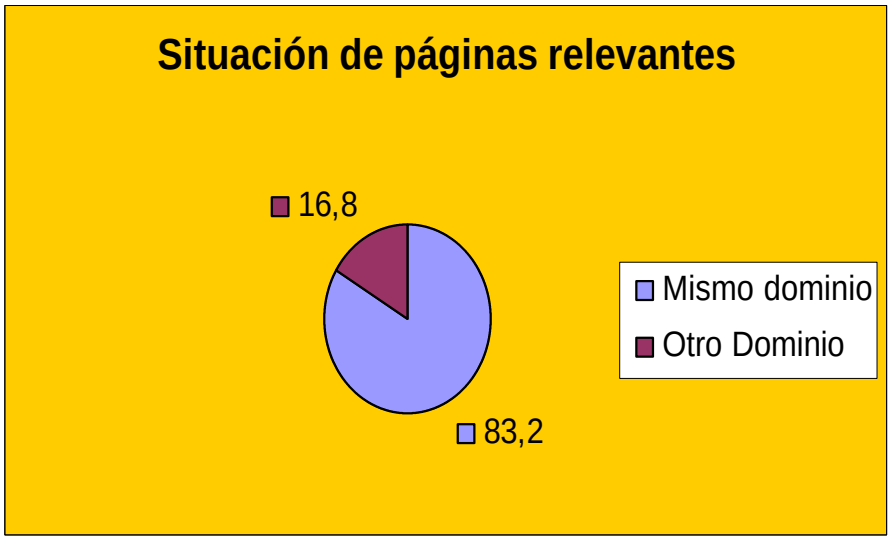

Figura 2.

Así pues, parece que el sistema da mayor importancia a las páginas cercanas a la utilizada como consulta. Parece razonable pensar que de todas formas es más probable que haya un mayor número de páginas similares o fuertemente relacionadas en el mismo entorno. Sin embargo, y sin perjuicio de efectuar análisis más detallados, el hecho es que una aplicación de este modo de representación tal como la hemos efectuado devuelve, en efecto, páginas relevantes, pero predominantemente aquéllas que se encuentran en las cercanías de la página de partida.

El problema puede verse de forma más clara si decimos que el 77% de las páginas relevantes recuperadas se encuentran a una distancia igual o inferior a 2 de la página utilizada como consulta. Esto significa que buena parte de las páginas recuperadas, aún siendo relevantes, son accesibles directa o casi directamente activando los hiperenlaces de la página de partida.

4. Conclusiones

El uso de enlaces salientes para representar contenidos de páginas web con vistas a su posterior recuperación puede ser una vía eficaz que puede aportar ventajas importantes sobre la utilización de términos. Entre éstas están la que supone reducir en varios ordenes de magnitud los cálculos de máquina necesarios para la recuperación, y la de obviar el problema de las diferencias idiomáticas entre documentos y personas que efectúen las búsquedas.

Sin embargo, es preciso incluir en los métodos de cálculo de pesos de los mencionados enlaces algunos mecanismos correctores, que limiten la influencia en los resultados de aquéllas páginas que son cercanas a la página utilizada como consulta.

5. Referencias:

Alonso Berrocal, José Luis (1997) Herramienta software para el análisis de la documentación WEB : rastreo de dominios, estudio de etiquetas, tipología de ficheros, evolución de los enlaces. Salamanca : Universidad de Salamanca, Facultad de Traducción y Documentación, 1997.

Almind, Tomas C. y Ingwersen, Peter (1997). Informetric analyses on the World Wide Web: methodological approaches to ‘webometrics’.// Journal of Documentation, 53:4, 404-426

Ellis, D.; Furner-Hines, J. y Willet, P. (1994). On the creation of hypertext links in full-text documents: measurement of inter-linker consistency.// Journal of Documentation, 50:2, 67-98.

Figuerola, C.G. (1999). Karpanta, URL: <http://milano.usal.es/karpanta>. Consulta: 1999.

Harman, D. (1992). Ranking Algorithms. // Information Retrieval: Data Structures and Algorithms: Prentice Hall, 1992. P. 363-392.

Joachims, T.; Mitchell, T.; Freitag, D. y Armstrong, R. (1995): WebWatcher: Machine Learning and Hypertext. URL <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/mltagung-e.ps.Z>.

Larson, Ray R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. URL: <http://sherlock.berkeley.edu/asis96/asis96.html>

Goffinet, Luc y Noirhomme-Fraiture, Monique (1998). Automatic hypertext link generation. URL: <http://www.info.fundp.ac.be/~lgo/Hypertext/semantic_links.html>

Rijsbergen, C.J. van (1979): Information Retrieval, Butterwoths, London, 1979.

Salton, G. y McGill, M. (1983): Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

Salton, G. (1987): On the relationships between theoretical retrieval models. Informetrics 87/88, Diepenbeeck (Bélgica), 1987, pp. 263-270.

Salton, G. y Buckley, C. (1988): Term-Weighting Approaches in Automatic Text Retrieval, Information Processing & Management, 24(5), 513-523.

Woodruff, Allison y otros (1996). An investigation of documents from the World Wide Web.// Fifth International World Wide Web Conference, May 6-10, Paris, France.