Título:

El contenido semántico de los enlaces de las páginas web desde el punto de vista de la recuperación de la información

Autores:

Carlos G. Figuerola, Universidad de Salamanca, Facultad de Traducción y Documentación
C/ Francisco de Vitoria, 6-16, 37008 SALAMANCA
Tf. + 34 923 29 45 80 ext. 3099 Fax +34 923 29 45 82
e-mail: figue@gugu.usal.es
Profesor del Departamento de Informática y Automática, imparte docencia en la Licenciatura en Documentación sobre Recuperación Automatizada de la Información

José Luis Alonso Berrocal, Universidad de Salamanca, Facultad de Traducción y Documentación
C/ Francisco de Vitoria, 6-16, 37008 SALAMANCA
Tf. + 34 923 29 45 80 ext. 4595 Fax +34 923 29 45 82
e-mail: berrocal@gugu.usal.es
Profesor del Departamento de Informática y Automática, imparte docencia en la Diplomatura en Biblioteconomía sobre Teledocumentación y en la Licenciatura en Documentación sobre Sistemas Hipermedia

Angel Fco. Zazo Rodríguez, Universidad de Salamanca, Facultad de Traducción y Documentación
C/ Francisco de Vitoria, 6-16, 37008 SALAMANCA
Tf. + 34 923 29 45 80 ext. 3092 Fax +34 923 29 45 82
e-mail: afzazo@gugu.usal.es
Profesor del Departamento de Informática y Automática, imparte docencia en la Licenciatura en Documentación sobre Redes Informáticas

Resumen:

Las páginas web pueden ser representadas mediante los formalismos aplicados habitualmente a cualquier otro tipo de documento textual, a efectos de su posterior recuperación. La mayor parte de dichos formalismos operan con términos. El uso de términos, sin embargo, presenta algunos problemas, aún mal resueltos con la tecnología actual; entre ellos está el de la normalización, y también el de la sustitución por sus equivalentes correctos en otros idiomas cuando se trabaja en entornos multilingües. De otro lado, las páginas web contienen otros elementos que pueden ser tenidos en cuenta, y que permitirían soslayar los problemas que encontramos con los términos. Así, es posible plantearse la representación y recuperación basándose en la utilización exclusiva de los enlaces contenidos en dichas páginas. Efectivamente, una página web puede ser caracterizada a través de sus enlaces, ya sean éstos salientes o entrantes. En este sentido, es posible plantear la utilización de formalismos y técnicas que habitualmente trabajan con términos, pero sustituyendo éstos por los enlaces. En este trabajo de definen algunos de estos formalismos y su aplicación práctica, y se explora su implementación, exponiendo la realización de algunos experimentos y proporcionando un avance de los resultados de éstos.

1. Introducción.

El principal problema con el que deben enfrentarse los sistemas de Recuperación de la Información es el de tener que trabajar con información no estructurada (al menos de una forma explícita). De hecho, el fundamento de los diferentes modelos teóricos que se han planteado, y de sus correspondientes implementaciones operativas, consiste en la aplicación de algún formalismo que permita representar adecuadamente cada uno de los documentos almacenados en la base de datos, así como las consultas que puedan generar los usuarios de la misma. La resolución de una consulta requiere la computación de alguna función de similaridad que permita establecer el grado de adecuación entre una consulta y cada uno de los documentos [SALTON87].

Naturalmente, la efectividad en la resolución de las consultas depende directamente de la bondad del formalismo empleado para representar los documentos. En sistemas manuales o semiautomáticos esta representación se elabora mediante el uso (manual) de lenguajes controlados, tales como tesauros, encabezamientos de materias y similares. En sistemas automáticos se aplican diversos modelos, pero la mayor parte de ellos basan la representación que construyen en las palabras o términos que contienen los documentos. Dichos términos pueden seleccionarse o valorarse en función de diversos criterios, pero son dichos términos los elementos básicos utilizados para representar los documentos [SALTON83] [RIJSBERGEN79].

El uso de términos como elemento básico de la representación de un documento se ha demostrado eficaz, pero plantea algunos problemas que con la tecnología actual no están bien resueltos. Entre ellos, el de la normalización de dichos términos, es decir, la reducción a una forma común de las distintas variantes (tanto flexivas como derivativas) que puedan aparecer en los documentos [GOMEZ98]. Pero además, en el caso de entornos multilingües (de lo que es un buen ejemplo el Web) tenemos la cuestión de la conversión de términos en una lengua determinada a sus equivalentes correctos (en función del contexto) en otra u otras lenguas.

Por estas razones, ha habido diversos intentos de representar documentos atendiendo a otros elementos. Un caso notable es el aplicado desde diversas instancias a la literatura científica. Los trabajos científicos se caracterizan, entre otras muchas cosas, por ir acompañados de un aparato bibliográfico más o menos importante: cualquier artículo científico contiene varias citas o referencias, con la intención de indicar al lector fuentes adicionales de conocimiento, o para apoyar las propias tesis en los trabajos o descubrimientos publicados en otros lugares.

Así, cuando operamos con colecciones documentales constituidas por artículos científicos es planteable representar dichos documentos a través de las referencias que contienen a otros artículos. Dicho de una forma simple: si dos artículos contienen las mismas citas o referencias, deben ser muy similares en cuanto a contenidos y temas que traten. Así pues, el grado de coincidencia en referencias o citas puede utilizarse para calibrar la semejanza entre dos artículos científicos. De esta forma, dado un artículo como punto de partida, es posible obtener aquéllos dentro de la colección que son parecidos en cuanto a temática o contenido [LAWRENCE99].

Obviamente, esta técnica no es excluyente y puede utilizarse en conjunción con los esquemas habituales basados en términos. De su aplicación práctica hay buenos ejemplos, del que puede destacarse ResearchIndex, del NEC Research Institute [NEC2000].

2. Aplicación en el web.

Este tipo de planteamientos podría ser extrapolado al Web, considerado éste como una colección de documentos. En este sentido, parece evidente que cualquier página web puede ser considerada como un documento, y podrá ser representada aplicando cualquiera de los modelos existentes, tomando como base el texto de dicha página. De hecho, esto es lo que hacen la mayor parte de los buscadores tipo Lycos, Altavista y otros [ALMIND97], [LARSON96], [WOODRUFF96].

Sin embargo, las páginas web poseen una característica que las hace especiales. Efectivamente, prescindiendo de imágenes, sonido, elementos de captación de datos (formularios) y otras maravillas, las páginas web tienen hipervínculos o enlaces con otras páginas (aunque en general diríamos que con otros recursos disponibles en la red).

Estos enlaces son lo que confieren su particular carácter a las páginas web, haciéndolas diferentes de los documentos convencionales. A partir de esos 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 [ELLIS94]. Consiguientemente, y dado que un hipervínculo se activa en un nodo determinado y nos dirige hacia otro nodo concreto, debemos distinguir entre enlaces entrantes y salientes. De esta forma, haciendo abstracción del contenido interno de cada nodo (página web, documento), 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 los que otros nodos mantienen con él.

Se trataría, entonces, de aplicar los mismos planteamientos indicados para la literatura científica, asumiendo para los enlaces de una página el papel de las referencias en los artículos científicos. Así, podríamos asumir que si dos páginas apuntan o enlazan a los mismos sitios, deben ser más o menos similares en cuanto a sus contenidos. Igualmente, si dos páginas son apuntadas desde los mismos lugares, sus contenidos deben guardar una relación más o menos estrecha.

Este enfoque ha sido planteado en varios trabajos, entre los cuales cabe destacar los de Kleinberg et al. [KLEINBERG98], [CHAKRABARTI99], de Bharat y Henzinger [BHARAT98], así como los de Joachims et al. [JOACHIMS95]. De hecho, estos trabajos, al menos como punto de partida, toman la metodología y los algoritmos del análisis de citas [EGGHE90].

Más allá de los modelos teóricos, parece interesante explorar la efectividad práctica de estos planteamientos; de hecho, resultados preliminares parecen indicar su viabilidad. Como se describe más pormenorizadamente en [FIGUEROLA99], la recuperación simple basada exclusivamente en los enlaces de las páginas web parecen tener una efectividad digna de tenerse en cuenta.

Para probar esto, se constituyó una colección documental (el espacio de búsqueda) a partir de 99.546 páginas web recogidas de forma automática por un pequeño robot experimental a partir de dominios de instituciones académicas y de investigación españolas [ALONSO97]. A esta colección se aplicó el clásico modelo vectorial [SALTON83] constituyendo los correspondientes vectores con los enlaces salientes de cada una de las 99.546 páginas. Los elementos de cada vector se pesaron utilizando el esquema estándar [SALTON88] de

$$F_e \; ? \; IDF_e$$

calculando IDF_e (*Inverse Document Frequency*) como

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

donde

F_e es la frecuencia del enlace en la página

N es el número total de páginas en la colección

ne es el número de páginas en que aparece el enlace

De esta colección se seleccionaron 200 páginas cuya misión fue la de servir como modelos para las consultas. Dicho de otro modo, el experimento consistió en obtener las páginas más similares a cada uno de esos 200 modelos. La similaridad se midió mediante la función típica del coseno [HARMAN92]:

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

donde

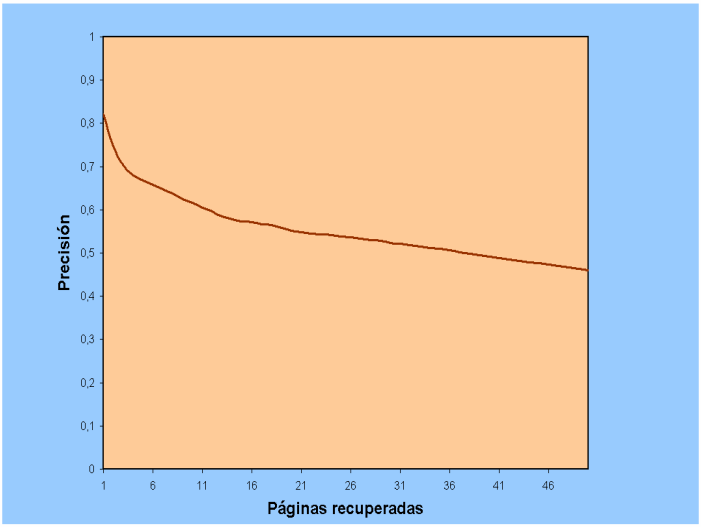
X es el vector de la consulta o modelo

Y es el vector del documento

X_i es el elemento *i* de ***X***

Y_i es el elemento *i* de ***Y***

n es el número de elementos o enlaces en los vectores



Los resultados de precisión fueron los reflejados en el gráfico adjunto, francamente positivos a pesar de la falta de puntos de referencia; aún cuando era apreciable un claro sesgo a la recuperación de páginas en el mismo dominio que la página modelo o consulta, o, en todo caso, a una distancia igual o menor a 2.

3. Tesauros de similaridad con enlaces.

La construcción de tesauros de similaridad es una técnica propuesta para la expansión de consultas cuando se trabaja, como sucede habitualmente, con los términos

de los documentos [QIU93]. Básicamente se apoya en la idea de que del mismo modo que un documento puede definirse o caracterizarse basándose en los términos que lo constituyen, un término podría ser también caracterizado en función de los documentos en que aparece [SCHAUBLE92], [JING94]. De alguna manera, se trataría de invertir el mecanismo habitual de representación de documentos a partir de sus términos, y así lograr una representación de los términos a partir de sus documentos (aquéllos en los que aparece).

De este modo, si operamos con el modelo vectorial clásico, sería posible construir un vector para cada uno de los términos presentes en la colección de documentos o base de datos. Cada vector de cada término tendría tantos elementos como documentos hay en la colección, y, para cada término, los elementos de su vector tendrían un valor numérico correspondiente a la importancia o peso del documento en relación con ese término. Este peso puede calcularse usando los esquemas habituales, pero cambiando los papeles de términos y documentos.

Así, al igual que sucede cuando representamos documentos a partir de sus términos, es posible calcular similaridades, de manera que, dado un término, es posible hallar aquellos otros que son más similares a él. Queda claro que tal similaridad no debe entenderse estrictamente semántica, sino que está calculada en función de que esos términos coincidan en más o menos documentos. En todo caso no parece desatinado pensar que si dos términos son comunes a muchos documentos probablemente existe algún tipo de relación entre ambos.

De la misma forma que se ha hecho con las recuperaciones simples, cabe preguntarse si es posible extrapolar esta técnica sustituyendo términos por enlaces de las páginas web. Es decir, conseguir una representación útil de un enlace determinado a partir de las páginas en que aparece, y a partir de ahí, por simple cálculo de similaridad, obtener los enlaces más similares o relacionados con uno dado.

Nuevamente, hemos efectuado algún experimento preliminar tendente a explorar las posibilidades de tales planteamientos. Se ha elaborado una colección de páginas, obtenidas recopilando de forma automática la totalidad del dominio *usal.es*. Esto ha producido un total de 5.992 páginas, con una media de 12,1 enlaces por página. En este punto es preciso matizar algunas cuestiones. En primer lugar, se han considerado la totalidad de enlaces, debidamente homogeneizados en su formulación, pero considerando los enlaces dirigidos a anclas distintas de una misma página como diferentes ocurrencias del mismo enlace. En segundo lugar, se ha observado una

tremenda dispersión en cuanto al número de enlaces por página (desviación típica de 26,9), sobre todo considerando que casi el 64% de las páginas contienen entre 1 y 10 enlaces.

Por otro lado, se han obtenido 20.830 enlaces diferentes, cada uno de los cuales aparece en una media de 2,27 páginas (10,87 de desviación típica). La distribución de sus frecuencias (entendidas como el número de páginas en que aparecen) es claramente zipfiana, al igual que ocurre con los términos en los documentos: se ha observado que hay 15 enlaces que aparecen en más de 100 páginas, mientras que 19.845 enlaces lo hacen sólo en 5 ó menos páginas, de ellos 13.169 únicamente en una.

Con esta colección de enlaces y páginas hemos construido un tesaurus de similaridad aplicando el modelo vectorial clásico en la forma ya comentada. Se elaboraron vectores para cada uno de los enlaces, cada uno con 5.992 elementos, pues éste es el número de páginas web en nuestra colección. El peso de cada uno de estos elementos se calculó mediante los esquemas habituales, pero cambiando los papeles, tal como se ha comentado anteriormente. Quedaría de la siguiente forma:

$$F_p \,?\, IDF_p$$

calculando IDF_p (*Inverse Document Frequency*) como

$$\log_2 \frac{N}{np} ? 1$$

donde

F_p es la frecuencia de la página P respecto a ese enlace
N es el número total de enlaces en la colección
np es el número de enlaces en que aparece la página P

$$Similaridad(X,Y) ? \frac{\sum_{i?1}^{n} X_i Y_i}{\sqrt{\sum_{i?1}^{n} X_i^2 ? \sum_{i?1}^{n} Y_i^2}}$$

donde

$\boldsymbol{X}$ es el vector en el que aparece el enlace consulta

$\boldsymbol{Y}$ es el vector con las páginas en que aparece el enlace a comparar

X_i es el elemento i de $\boldsymbol{X}$

Y_i es el elemento i de $\boldsymbol{Y}$

n es el número de páginas en la colección

Una vez que tenemos los vectores, dado un enlace concreto podemos localizar aquellos que presuntamente son más similares a ese enlace. Dado que un enlace apunta a una página concreta, parece razonable intercambiar sin mayor problema el enlace por la página a la cual apunta, es decir, la que se encuentra en la dirección contenida por el enlace. Así, partiendo de una página determinada, si en nuestra colección de páginas tenemos la que contiene la dirección o apunta a dicha página, podríamos encontrar otros enlaces (las direcciones de otras páginas) similares a la citada página de partida. La condición indispensable es que en nuestra colección de páginas web haya alguna que enlace con la página de partida, de la cual queremos encontrar semejantes; pero no es preciso que dicha página de partida forme parte de nuestra colección. De la misma forma, obtendremos una serie de enlaces a páginas similares, pero no necesariamente éstas han de formar parte de nuestra colección; si no simplemente ser apuntadas o enlazadas por alguna de las que sí que forman parte.

En el momento actual, una vez construido el tesauro de similaridad con los enlaces, estamos efectuando pruebas preliminares, tendentes a explorar la viabilidad práctica de un sistema basado en estos planteamientos. Las respuestas obtenidas parecen satisfactorias, aunque no podemos cuantificar su efectividad todavía: una métrica medianamente rigurosa requiere la elaboración previa de un conjunto de enlaces modelo que puedan ser representativos de diferentes situaciones, así como el examen de las repuestas obtenidas, realizado preferiblemente por un número amplio de personas, tendente a determinar la relevancia de los enlaces devueltos por el sistema en cada una de las consultas. Igualmente, se necesita el análisis cuidadoso de las respuestas halladas erróneas, y también los comportamientos anómalos o inesperados. No obstante, aunque sin cuantificar, parece claro que un sistema basado en estos principios es capaz de recuperar enlaces (páginas) similares a un enlace proporcionado como punto de partida.

4. Conclusiones.

Los enlaces contenidos en las páginas web pueden ser utilizados para describir o representar dichas páginas desde el punto de vista del contenido temático de las mismas. En este sentido, es posible aplicar con ellos mecanismos semejantes a los empleados habitualmente con los términos de los documentos, tendentes a obtener su recuperación. Así, cabe atribuir a estos enlaces una cierta carga semántica, en el sentido de que

describen el contenido de las páginas en que aparecen. El aprovechamiento de esta carga semántica puede permitir soslayar o aminorar algunos de los problemas que surgen cuando los documentos se representan a partir de los términos que contienen, como el de la normalización de esos términos o las diferencias idiomáticas.

Sin embargo, es difícil valorar el alcance de esa carga semántica. En este trabajo se han descrito algunos de los planteamientos teóricos aplicables, así como algunos experimentos cuyos resultados aún deben medirse y analizarse adecuadamente. Estimaciones preliminares, no obstante, parecen indicar la viabilidad de sistemas basados en el aprovechamiento de los enlaces.

5. Bibliografía.

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*, 1997, 53, 4, p. 404-426

Bharat, K. & Henzinger, M.R. (1998). Improved algorithms for topic distillation in a hyperlinked environment, *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998

Chakrabarti, S.; Dom, B.E.; Gibson, D.; Kleinberg, J.; Kumar, R.; Raghavan, P.; Rajagopalan, S. y Tomkins, A. (1999): Mining the Link Structure of the World Wide Web, *IEEE Computer*, August 1999

Egghe, L. & Rousseau, R (1990). Introduction to Informetrics, *Elsevier*, 1990

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*, 1994, 50, 2, p. 67-98

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

Gómez Díaz, R. (1998). La Recuperación de Información en español: evaluación del efecto de sus peculiaridades lingüísticas (Documento sin publicar en la Universidad de Salamanca, Salamanca, 1998)

Jing, Y. y Croft, W.B. (1994). An association thesaurus for information retrieval, *Proceedings of RIAO 94*, p. 146-160

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

Kleinberg, J. (1998). Authoritative sources in a hiperlinked environment, *Proceedings ACM-SIAM Symposium on Discrete Algorithms*, January 1998, San Francisco, California, 25-27, p. 668-677

Laffling, J. (1992). On Constructin a Transfer Dictionary for Man and Machine. *Target*, 1992, 4 (1), p. 17-31

Larson, Ray R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. *Annual meeting of the American Society for Information Science*, 1996.
Notes: http://sherlock.berkeley.edu/asis96/asis96.html

Lawrence, Steve; Bollacker, Kurt; Lee Giles, C. (1999). Indexing and retrieval of scientific literature. *Eight International Conference on Information and Knowledge Management, CIKM99*, Kansas City, Missouri, November 1999, 2-6, p. 139-146

NEC (2000). ResearchIndex: The NECI Scientific Literature Digital Library, http://citeseer.nj.nec.com/

Qiu, Y. y Frei, H.P. (1993). Concept Based Query Expansion. *ACM SIGIR*, 1993, p. 160-169

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, p. 263-270

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

Schauble, P. y Knauss, D. (1992). The various roles of information structures, *16 Jahrestagung der Gesellschaft für Klassification*, Dortmund, 1992

Sheridan, P. y Ballerini, J.P.(1996). Experiments in Multilingual Information Retrieval using SPIDER system. *SIGIR 96*, 1996, p. 58-65

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