Nuevos puntos de vista en la Recuperación de Información en el Web

Carlos G. Figuerola
José Luis Alonso Berrocal
Angel Francisco Zazo Rodríguez
Universidad de Salamanca. Facultad de Documentación
e-mail: [figue | berrocal | afzazo]@gugu.usal.es

Resumen:

El web, como conjunto de páginas html relacionadas mediante enlaces o hipervínculos, puede ser representado como un grafo, en el que cada página constituye un nodo, y cada enlace puede considerarse un arco. De otro lado, los sistemas más habituales de Recuperación de la Información se basan en conseguir una representación procesable y homogénea de documentos y consultas, y en el cálculo subsiguiente de alguna función de similaridad entre la representación de una consulta dada y los documentos de una colección. Se explora la posibilidad de aplicar este tipo de técnicas, intentando la representación de las páginas web en función de los enlaces, en lugar de las palabras del texto de las mismas.

El web, como conjunto de páginas html relacionadas mediante enlaces o hipervínculos, puede ser representado como un grafo, en el que cada página constituye un nodo, y cada enlace puede considerarse un arco. Dado que los enlaces o hipervínculos parten de páginas o nodos concretos y apuntan a otra página concreta, podemos hablar de grafo dirigido, de manera que los arcos tienen un sentido o dirección determinada, independientemente de que los navegadores puedan tener utilidades de vuelta atrás o mecanismos similares [ELLIS94].

Aunque las páginas html contienen, desde luego, más cosas, además de los enlaces o arcos, éstos constituyen elementos informativos de importancia, que pueden ayudar, además, a caracterizar dichas páginas. En efecto, los hipervínculos establecen una relación entre unas páginas y otras. De una manera intuitiva, podemos pensar que dos páginas que reciben enlaces desde los mismos nodos deben tratar acerca de los mismos o parecidos temas. Páginas que apuntan o enlazan con los mismos nodos podrían ser más o menos similares en su temática o contenido.

De otro lado, los sistemas más habituales de Recuperación de la Información se basan en conseguir una representación procesable y homogénea de documentos y consultas, y en el cálculo subsiguiente de alguna función de similaridad entre la representación de una consulta dada y los documentos de una colección. Aquellos documentos con un índice de similaridad más alto, respecto de una consulta dada, son los que, presuntamente, se ajustan mejor a las necesidades informativas expresadas en la consulta. La bondad de un sistema de recuperación (referente a su efectividad) descansa fundamentalmente sobre el sistema elegido para representar o modelar documentos y consultas. Depende también, obviamente, de la función de similaridad utilizada, pero éste es un aspecto directamente dependiente del sistema de representación, de manera que éste impone de partida una serie de limitaciones y de pautas a las cuales deben ajustarse los cálculos de similaridad. Dicho de otro modo, estas funciones sólo pueden operar con los elementos elegidos para representar documentos y consultas.

Precisamente, uno de los modelos de Recuperación de la Información más conocido es el llamado modelo vectorial, o del espacio vectorial, propuesto por G. Salton a principios de los años 70 [SALTON83], [SALTON87] y ampliamente difundido desde entonces, tanto a nivel experimental como en implementaciones operativas. La base de dicho modelo consiste en la representación de los documentos a través de vectores, en los que cada elemento sería una característica observable en los documentos. Una visión elemental del concepto 'característica' es la equivalente a palabra, y de hecho esto es con lo que con más frecuencia se ha trabajado, aunque podrían contemplarse características de otro tipo. De manera que, según este esquema, cada palabra posible sería una característica, y, en consecuencia, un elemento de los vectores de los documentos.

Un documento concreto, naturalmente, contiene algunas palabras, y otras no. De manera que ese documento podría representarse mediante un vector que, en los elementos correspondientes a las palabras que forman dicho documento, contenga un valor determinado; mientras que para los elementos correspondientes a las palabras que no forman parte de dicho documento, se asigne otro valor distinto que recoja esta circunstancia. En un esquema binario –el más simple, pero también el más ineficaz- podríamos asignar a los elementos del vector un 1 si la palabra forma parte de documento y un 0 en caso contrario.

Naturalmente, no tenemos porqué considerar características observables o significativas todas las palabras posibles. De hecho, es normal no tener en cuenta o descartar las palabras llamadas vacías; como también normalizar las palabras reduciéndolas –en la medida de los posible- a su raíz y unificando derivados y formas flexionadas.

De la misma manera, una consulta o interrogación a la colección de documentos puede expresarse en lenguaje natural, mediante una frase, o varias, o incluso varios párrafos si se desea. De esta manera, puede ser tratada de la misma forma que un documento, y representada también mediante un vector igual, con los mismos elementos, pero con los valores correspondientes en cada uno de ellos, en función de las palabras que formen parte de la consulta.

Reducidos documentos y consultas al mismo tipo de representación, pueden ser comparados fácilmente recurriendo a alguna de las muchas funciones existentes para comparar vectores. Un ejemplo elemental, cuando se trabaja con vectores binarios podría ser el mero producto de vectores.

Sin embargo, los valores binarios no resultan lo suficientemente expresivos, pues parece claro que unas palabras pueden resultar más significativas que otras. Intuitivamente, podemos pensar que no debe representar igual el contenido de un documento una palabra que sólo aparece una vez que otra que aparece varias en el mismo documento. Así pues, se suele utilizar algún tipo de coeficiente o peso que intente expresar el valor o importancia de cada palabra en cada uno de los documentos. Este coeficiente puede calcularse de muchas formas [SALTON88], y en función de diversos parámetros y criterios. Uno de los modos más habituales es hacerlo teniendo en cuenta la frecuencia de aparición de la palabra (o, de un modo más general, la característica). En líneas generales, suele partirse de la idea de que el peso de un término en un documento dado es inversamente proporcional a su frecuencia en la colección de documentos y directamente proporcional a su frecuencia en el documento en cuestión. Hay una buena cantidad de fórmulas propuestas para calcular el peso, basadas en estas ideas. Por poner un ejemplo, una de las más utilizadas es:

$$\frac{\left(ftd * \log(N / n) \right)}{\left[\sum_{i=1}^{nt} ftdi * \log/ (N^2 / ndi) \right]^{1/2}}$$

donde

ftd es la frecuencia del término *t* en el documento *d*

nt es el número de términos en el documento *d*

N es el número de documentos en toda la colección

n es el número de documentos en que aparece el término *t*

nd_i es el número de documentos en que aparece el término *i*

Para el caso de utilización de vectores con pesos, hay también unas cuantas formas de calcular la similaridad entre dos vectores. Una de las más aplicadas, prácticamente un estándar, es la del coseno [HARMAN92b]:

$$SIM(P_x, D_y) = \frac{\sum_{i=1}^{n} p_{xi} d_{yi}}{\sqrt{\sum_{i=1}^{n} p_{xi}^2 \cdot \sum_{i=1}^{n} d_{yi}^2}}$$

donde

P_x es el vector del documento *x*

D_y es el vector del documento *y*

p_{xi} es el elemento *i* de P_x

d_{yi} es el elemento *i* de D_y

n es el número de elementos o términos en los vectores

Las páginas web, dado que contienen, entre otras cosas, texto (palabras) pueden ser recuperadas aplicando el modelo vectorial u otros modelos utilizados en Recuperación de la Información. La cuestión es que tales modelos se utilizan, habitualmente, sobre texto o palabras. Ello conlleva algunos problemas; entre ellos, el de la normalización de las palabras, entendida ésta, fundamentalmente, como la reducción a raíces o lemas originales, agrupando derivados con similar contenido semántico y formas flexionadas [PAICE96][KRAAIJ96]; esta cuestión está lejos de haber sido resuelta de forma satisfactoria, en especial para lenguas (como el español) de gran complejidad morfológica. Otro tanto cabe decir sobre la identificación de formas gramaticales o de expresiones complejas, utilizadas por muchos sistemas de recuperación de forma paralela a los términos individuales.

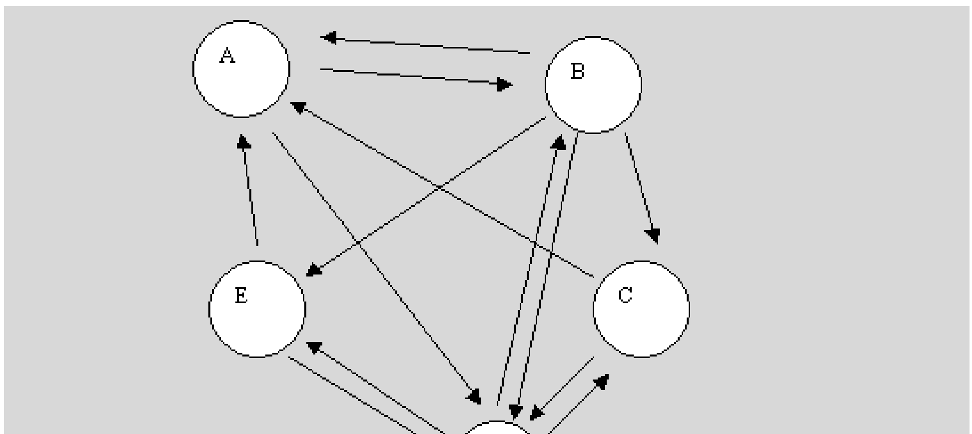
De otro lado, en entornos claramente multilingües, como es el web, la recuperación se torna más difícil, debido a las diferencias idiomáticas entre consultas y documentos. La recuperación multilingüe se encuentra, en la actualidad, en mantillas, y sus resultados son significativamente más pobres que los obtenidos en entornos monolingües [OARD96]. De otro lado, la importante carga en términos de capacidad de proceso que supone manejar un alto número de vectores (tantos cuantas páginas web o documentos conformen la colección en la que efectuar recuperaciones), cada uno de ellos con un número muy elevado de elementos (las posibles palabras que puedan aparecer en las páginas), supone en la práctica un obstáculo importante. Aunque es bien cierto que se han propuesto algunas soluciones de cara a reducir el tamaño de los vectores [DEERWESTER90], esto es, a reducir el número de características (palabras) a tener en cuenta para representar cada documento; el número de páginas web es tan alto y con un crecimiento tan sostenido, que seguimos necesitando una gran capacidad de proceso y almacenamiento para operar con ellas.

Ahora bien, teniendo en cuenta que las páginas html contienen más elementos informativos, además del texto, cabe plantearse la posibilidad de utilizar dichos elementos para la recuperación de la información, bien sea de forma autónoma, bien sea de forma complementaria a la utilización del propio texto. En este sentido, los hipervínculos o enlaces de unas páginas hacia otras pueden jugar un papel importante. Tales enlaces, que, como se ha dicho, cabe representar como los arcos de un grafo dirigido, podrían considerarse como características en mayor o menor medida definitorias de las páginas o documentos. Así, podemos considerar que aquellas páginas hacia las que convergen o apuntan enlaces que parten de los mismos nodos, deben estar, de algún modo, relacionadas. De igual forma, aquellas páginas que tienen enlaces apuntando a los mismos nodos, deben versar sobre temas próximos en mayor o menor medida.

De esta forma, puede plantearse la construcción de vectores representativos de páginas web (documentos) utilizando en lugar de las palabras del texto los hipervínculos o enlaces. Un documento web sería representable tomando como características clave los hipervínculos o enlaces relacionados con dicho documento. Si esta hipótesis se conforma como cierta –al menos en una medida aceptable- podríamos solventar algunos de los problemas apuntados antes. De entrada, los enlaces son independientes del idioma y podríamos sortear la cuestión de la recuperación multilingüe. Pero tampoco habría que abordar la cuestión de la normalización o lematización, como también, al tratarse de un número considerablemente menor de características por documento, tanto capacidad de proceso como de almacenamiento verían reducidas sus necesidades.

La cuestión está, en primer lugar, en examinar cómo se puede efectuar el tratamiento de los hipervínculos, aplicando parecidas técnicas a las empleadas con las palabras. En este sentido, tomando como punto de referencia una página o documento concreto, podemos, en primer lugar, descartar los enlaces internos o anclas, toda vez que apuntan hacia el mismo documento, aunque a una parte concreta del mismo; son los enlaces externos, hacia otras páginas, los que nos interesan. Entre éstos, hay que distinguir entre los que, por así decir, se reciben; esto es, los que, partiendo de otras páginas o nodos apuntan hacia la página en cuestión. Y los que, desde la página tomada como referencia, apuntan o enlazan con otras diferentes. Es decir, enlaces que, desde otras páginas apuntan hacia la que nos interesa, y enlaces que, partiendo de ésta, apuntan hacia otras.

La figura muestra una representación ideal, con 5 nodos o páginas (A, B, C, D y E) y sus enlaces dirigidos entre sí.



El *web* representado a través de un grafo dirigido. Cada página es un nodo con hipervínculos que salen o llegan

El primer tipo o categoría de enlaces parecen tener una mayor carga expresiva acerca de la posible similaridad entre páginas [JOACHIMS95]. Presentan, sin embargo, el problema de la imposibilidad de conocer a ciencia cierta cuántas páginas o enlaces apuntan hacia una dada. Desde un punto de vista teórico, esta cuestión podría aproximarse tomando en cuenta la mayor parte posible del web. Operar con la totalidad de éste es descartable de antemano; debido al tamaño y al dinamismo de éste, probablemente después de explorar un dominio algo grande y de recoger datos acerca de sus enlaces, dicho dominio habría sufrido ya variaciones con nuevas páginas, otras que desaparecen o cambian etc.. Pero parece claro que, al trabajar con el mayor número de páginas posible, la probabilidad de no tomar en consideración algún enlace no conocido disminuye, y su impacto también.

	A	B	C	D	E
A	-	1	1	0	1
B	1	-	0	1	0
C	0	1	-	1	0
D	1	1	1	-	1
E	0	1	0	1	-

Matriz de nodos que son apuntados. Los nodos o páginas de las filas son apuntados por enlaces desde los nodos de las columnas

Obsérvese que este problema sigue presente aún cuando se desease limitar la recuperación a páginas pertenecientes a una parte concreta y delimitada del web. En efecto, supóngase que deseamos recuperar las páginas similares a una dada, pero sólo las pertenecientes a determinada parte del web. Deberíamos encontrar aquellos enlaces que apuntan hacia la página de partida, y luego los enlaces que apuntan hacia cada una de las páginas que componen la parte del web a la que queremos limitar la recuperación. Tanto la página origen como las demás pueden recibir enlaces desde cualquier otra parte del web, no sólo desde dentro de la parte acotada para la recuperación.

Otra cuestión importante es la referente al cálculo de pesos en los enlaces. Cuando tratamos con enlaces que se reciben, tal vez los pesos pierdan parte de su significado. Parece difícil, desde el punto de vista del receptor del enlace o apuntado por éste, considerar que se es más referenciado por un enlace que por otro. De hecho, algunos experimentos relacionados en la literatura en base a estas técnicas han manejado vectores binarios. El cálculo de la similaridad podría reducirse al mero producto entre los vectores o, lo que es lo mismo, a la simple suma de enlaces coincidentes. Algunos esquemas más complejos podrían tener en cuenta otro matices, como el número total de enlaces tenidos en cuenta, etc. [MITCHELL94].

El otro tipo o categoría de enlaces comentado son los hipervínculos que parten de la página origen hacia otras páginas. Aquí parece haber menos dificultades teóricas. En efecto, el número de enlaces que contiene una página hacia se conoce puesto que todos están incluídos en esa página. Así, aún cuando sólo dispongamos datos de una parte del web, podemos tener la seguridad de que operamos con todas las características (enlaces) de las páginas con las que tratamos. Lo más que puede pasar (que no es poco) es que, al quedar parte del web fuera de nuestro trabajo, no recuperemos la totalidad de páginas similares a la de partida.

	A	B	C	D	E
A	-	1	0	1	0
B	1	-	1	1	1
C	1	0	-	1	0
D	0	1	1	-	1
E	1	0	0	1	-

Matriz de nodos que apuntan. Los nodos o páginas de las filas apuntan mediante enlaces a los nodos de las columnas

Este tipo de enlaces, sin embargo, plantea otros problemas. Por ejemplo, el ruido introducido por los típicos enlaces de 'Se ve mejor con NetEscape', www.microsoft.com y similares. Este tipo de enlaces, que conducen a lugares comunes y que tienen escaso valor a la hora de definir el tema o contenido de una página, podrían recibir un tratamiento parecido al de las llamadas palabras vacías en la recuperación de información basada en términos. Aquí suele abordarse mediante la aplicación de listas de palabras vacías construidas a priori, basándose en el conocimiento del lenguaje: preposiciones, conjunciones, artículos, etc. [SALTON83]. No es difícil elaborar listas similares de 'enlaces vacíos'.

Otro enfoque, compatible con el anterior, incluye un tratamiento más agresivo a la hora de localizar enlaces inútiles desde el punto de vista de identificar el contenido de una página. La base de dichos tratamientos reside en un estudio de las frecuencias de aparición de términos [WILBUR92] –enlaces en nuestro caso- que puede llegar a ser bastante radical en la eliminación o exclusión de términos que pueden considerarse carentes de valor discriminatorio[YANG96].

Esto nos lleva directamente a la cuestión de la aplicación de pesos en la descripción de las páginas. Básicamente, puede aplicarse el mismo criterio para calcular pesos de enlaces que el utilizado cuando se trabaja con vectores de términos: IDF (inverse document frequency) [RIJSBERGEN79] y número de veces que un término (enlace) aparece en un documento (página html). El IDF es una función inversa de la distribución del término en la colección de documentos [HARMAN92b], y puede aplicarse sin problemas a nuestro caso, sustituyendo términos por enlaces y documentos por páginas web. Por lo que se refiere al número de veces que un enlace aparece en una página web o documento, parece que ese número será 1 en la mayor parte de los casos: no debe ser habitual que en una misma página el mismo enlace aparezca varias veces, si exceptuamos enlaces en imágenes y versiones alternativas en texto y situaciones similares. Sin embargo, sí que parece tener más importancia la siguiente situación que se produce con cierta frecuencia: enlaces diferentes, pero que se refieren todos o apuntan a la misma página, sólo que a lugares distintos de la misma. Fácilmente identificables, podría ser razonable tratarlos como el mismo enlace, con varias ocurrencias o apariciones en el mismo documento.

En la misma línea podrían estar enlaces a páginas diferentes, pero dependientes o colgando del mismo 'home'; tal vez hubiera que afinar aquí algo más y dar entrada a la distancia del 'home' a la página enlazada. Otro caso, más difícil de identificar, es el de los enlaces a direcciones diferentes pero que en realidad contienen lo mismo, por tratarse de un 'mirror', o debido a la utilización de 'alias', etc.

Conclusión

La utilización de los hipervínculos como características definitorias de las páginas web podrían ser usadas –al menos desde un punto de vista teórico- en la recuperación de dichas páginas, aplicando las mismas técnicas que se emplean habitualmente en Recuperación de Información basada en palabras o términos. Ello permitiría sortear las diferencias idiomáticas, dado el carácter multilingüe del web, y reducir significativamente la capacidad de proceso y almacenamiento necesarias. Este planteamiento, sin embargo, precisa de comprobación experimental y de una evaluación de sus posibilidades. Sin embargo, parece que, bien como procedimiento de recuperación autosuficiente, bien como complemento o parte de esquemas de recuperación más complejos, puede resultar útil e interesante.

REFERENCIAS:

[DEERWESTER90]	DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. & HARSHMAN, R. (1990): "Indexing by latent semantic analysis", JASIS, 41(6), 391-407
[ELLIS94]	ELLIS, D.; FURNER-HINES, J. & 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
[HARMAN92]	HARMAN, D. (1992): "Relevance Feedback and Others Query Modification Techniques", en Information retrieval: data structures and algorithms, Prentice-Hall, New Jersey, 1992
[JOACHIMS95]	JOACHIMS, T., MITCHELL, T., FREITAG, D. & ARMSTRONG, R. (1995): "WebWatcher: Machine Learning and Hypertext", http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/mltagung-e.ps.Z (consultado el 13-03-1998)
[KRAAIJ96]	KRAAIJ, W. & POHLMANN, R. (1996): "Viewing Stemming as Recall Enhancement", SIGIR 96, Konstanz (DE), 1996, 40-48
[MITCHELL94]	MITCHELL, T.M., CARUANA, R., FREITAG, D., MCDERMOTT, J. & ZABOWSKI, D. (1994):"Experience with a Learning Personal Assistant", CACM, 37(7), 81-91, http://www.cs.cmu.edu/afs/cs/user/mitchell/ftp/cacm.ps.Z
[OARD96]	OARD, D.W. & MARCHIONINI, G. (1996): "A Conceptual Framework for Text Filtering", en http://www.ee.umd.edu/medlab/filter/papers/filter/filter.html [consulta 02 mar 1998]
[PAICE96]	PAICE, C.D.(1996): "Method for Evaluation of Stemming Algorithms Based on Error Counting", JASIS, 47(8), 632-649
[RIJSBERGEN79]	Information Retrieval, London, 1979
[SALTON83]	SALTON, G. & MCGILL, M. (1983): Introduction to Modern Information Retrieval, New York, McGraw-Hill, 1983
[SALTON87]	SALTON, G.(1987): "On the relationship between theoretical retrieval models", Informetrics 87/88,. Diepenbeeck (Bélgica), 1987, pp. 263-270.
[SALTON88]	SALTON, G. & BUCKLEY, C. (1988): "Term-Weighting Approaches in Automatic Text Retrieval", Information Processing & Management, 24(5), 513-523
[WILBUR92]	WILBUR, J. Y SIROTKIN, K.(1992): "The automatic indentification of stop wordsd", Journal of Information Science, 18, 45-55
[YANG96]	YANG, Y. Y WILBUR, J.: "Using Corpus Statistics to Remove Redundant Words", JASIS 47(5), 1996, pp. 357-369