Navegación Autónoma y Recuperación de la Información en el web

C. G. Figuerola, J. L. Alonso Berrocal, A. F. Zazo Rodríguez, E. Rodríguez Vázquez de Aldana

Universidad de Salamanca

Abstract. El problema de la recuperación de información en el Web se puede plantear desde diferentes puntos de vista, con mecanismos como la realimentación por relevancia, la utilización de tesauros, el estudio de los hiperenlaces, o la aplicación de redes neuronales, entre otros. Todos estos mecanismos se aplican sobre grandes bases de datos construidas a partir de la exploración previa de sectores más o menos amplios del Web. La experiencia ha demostrado que la precisión de estos sistemas es baja, y la exhaustividad está relativizada al sector explorado. Existe sin embargo otra aproximación al problema que pretende obtener resultados mucho más precisos, aunque sin perseguir altas tasas de exhaustividad, basándose en el uso de agentes inteligentes que rastreen la red según las necesidades informativas del usuario. Se indican las características de los agentes y se analizan algunas de las propiedades y habilidades deseables para aquellos agentes dedicados a la recuperación de información en el Web.

El número de páginas *web* crece exponencialmente y afecta a todos los ámbitos del conocimiento [1]. Este hecho pone de relieve el problema de la Recuperación de Información en la red; básicamente, los sistemas de Recuperación en el *Web* utilizan dos mecanismos, no excluyentes entre sí: búsquedas mediante palabras clave o clasificación en clases o categorías de páginas *www* [2]. Ambas técnicas pueden ser combinadas entre sí.

Sin embargo, cualquiera de estos sistemas parte de la existencia de una base de datos de páginas web, la cual suele ser de gran tamaño. La experiencia de los usuarios muestra claramente que, en general, estos sistemas producen respuestas de muy baja precisión [3]. En cuanto a la exhaustividad, se percibe en términos brutos como muy alta (la típica respuesta de un buscador con cientos o miles de páginas encontradas). Aunque esta exhaustividad debe ser relativizada, puesto que es conocido que incluso los buscadores más importantes cubren sólo una parte de todo el espacio *web*[4].

Existe, sin embargo, otra aproximación al problema que, sin perseguir altas tasas de exhaustividad, pretende obtener resultados mucho más precisos, basada en la navegación autónoma.

1 Navegación Autónoma

La idea básica es la exploración automática del *www*, a fin de recuperar las páginas relevantes para unas necesidades informativas determinadas. El software encargado de esta tarea tomaría la formulación de tales necesidades informativas como parte de las especificaciones iniciales. El programa exploraría la red, eligiendo los enlaces más prometedores, accediendo a nuevas páginas, recopilando las que pudiesen satisfacer las especificaciones iniciales, y así sucesivamente. Todo ello de forma muy parecida a como lo hacemos manualmente las personas, pero de forma automática, mientras trabajamos en otras tareas o, simplemente, descansamos.

Puesto que la mera exploración del *web*, requiere grandes cantidades de tiempo, un enfoque de este tipo tiene algunas limitaciones. No es esperable una respuesta inmediata, ni siquiera probablemente con la agilidad suficiente para plantear una dinámica interactiva con el usuario. Se espera que, en un plazo razonable (el propio usuario podría establecer plazos máximos) el software entregue las páginas *www* encontradas útiles para satisfacer las necesidades de información del usuario.

La otra limitación importante es la renuncia implícita a la exhaustividad. Parece claro que, dado el tamaño del *Web* la exploración completa, incluso de una parte significativa de él, resulta implanteable; mediante estos planteamientos se exploraría tan sólo una pequeña parte del *web*. Se espera, en contrapartida, que los resultados obtenidos alcancen una notable precisión; de esta forma, podría obviarse el conicido efecto de la *sobrecarga de información*. Aceptando estas limitaciones, este tipo de planteamientos debe resolver una serie de cuestiones, para lo cual se han propuesto diversas soluciones, la mayor parte no excluyentes entre sí. Examinaremos a continuación los problemas más importantes.

2 La elección de los puntos de partida

Puesto que se trata de explorar el *web*, es preciso determinar algún punto de partida. Podemos ver el *web* como un grafo dirigido, en el que las páginas son los nodos y los enlaces son los arcos. El proceso de exploración parte de un nodo y, utilizando los enlaces, pasa a otros nodos, y así sucesivamente. Como la distancia entre el nodo por el que se empieza y cualquiera de los nodos relevantes puede ser muy grande, es crítico localizar previamente puntos de partida que puedan estar lo más cercanos posible a las relevantes para las necesidades de información del usuario.

Una vía utilizada frecuentemente para elegir buenos puntos de partida es comenzar con una búsqueda al estilo clásico en diferentes buscadores convencionales. En estos casos tales búsquedas previas suelen enviarse a servicios *metabuscadores*[5], los cuales tratan con los diferentes buscadores, recogen los resultados de cada uno de ellos, los organizan y los devuelven a quien hizo la consulta. Las páginas devueltas son las candidatas a ser puntos de entrada. Los puntos de entrada pueden manejarse de forma secuencial o en paralelo; esto puede permitir

utilizar procesamiento paralelo o varios ordenadores [6], pero también presenta la ventaja de obviar en alguna medida problemas derivados de las comunicaciones, como cuellos de botella, etc.

De otro lado, disponer de varios puntos de entrada implica la selección de parte de ellos (en un número razonable), así como su priorización. Hay diversas estrategias para abordar esta cuestión; desde tomar simplemente los n primeros, hasta aplicar medidas de similitud entre las especificaciones del usario y el contenido de las páginas, pasando por cosas como el número de enlaces de cada punto de entrada o incluso prospecciones de tiempos de respuesta. También es posible una realimentación por parte del usario, seleccionado éste los puntos de entrada. Naturalmente, estos diversos enfoques son combinables entre sí.

3 Activando enlaces

Dada una página de partida, el software debe extraer los enlaces que esa página contenga y guardarlos en una lista. Posteriormente, irá tomando enlaces de esa lista, recuperando las páginas a las que apuntan y así sucesivamente. El seguimiento de todos los enlaces en la lista llevaría, teóricamente, a la exploración de todo el *web*. Sin embargo, como solemos tener limitaciones de recursos y, especialmente, de tiempo, se hace preciso establecer un orden de prioridad en dicha lista. Este orden atiende a dos premisas fundamentales: en primer lugar, la relevancia de los enlaces respecto de las necesidades informativas del usuario. En segundo lugar, las posibilidades de acceder a mayores espacios del *web* desde unos enlaces que desde otros.

Se ha propuesto diversos sistemas para seleccionar aquellos enlaces más prometedores desde este segundo punto de vista. Una posibilidad es utilizar los *backlinks* de una página, esto es, las páginas que tienen enlaces hacia esa página [7]. El mecanismo más simple es contar el número de *backlinks*; sin embargo, el problema es disponer de dicha información. En este sentido, cabe mencionar el proyecto Compaq's Connectivity Center Server [8], en estrecha relación con *Altavista*.

Más sofisticado es el algoritmo conocido como *PageRank* [9]. La idea básica es que la importancia de una página es directamente proporcional al número de *backlinks* que éste tiene; pero no todos los *backlinks* pesan lo mismo, sino que su valor está en función de la importancia de la página de la que procedan. Y la página de procedencia tiene, a su vez, una importancia que viene determinada por los *backlinks* que recibe, y así sucesivamente. El cálculo del *PageRank* ha de hacerse de forma iterativa, y es costoso en tiempo de proceso. Éste es el mismo problema que encontramos para calcular otro tipo de coeficientes, cuya finalidad es también estimar la importancia de unos determinados nodos frente a otros [10]. Estos índices son utilizados también por algunos buscadores covencionales para ordenar los resultados obtenidos en una búsqueda [11] (como *Google*).

4 Selección de páginas por contenido

En un exploración de este tipo es preciso disponer de medios para estimar la proximidad de un nodo a las necesidades informativas del usuario; esto debe permitir seleccionar páginas para que sean entregadas como resultado. Pero también, en conjunción con la estimación de importancia vista antes, para determinar cuáles son los enlaces más prometedores para proseguir la exploración. En esta línea, diversos mecanismos pueden ser utilizados, y muchos de ellos pueden combinarse entre sí.

4.1 Similaridad documental

Si consideramos cada página *web* un documento, podemos aplicar las técnicas utilizadas habitualmente en Recuperación de la Información para estimar la semejanza entre una página explorada y las necesidades informativas. En realidad, lo que se compara con estas técnicas es sólo el *texto*, con lo que toda la carga informativa no textual presente en las páginas *web* no se toma en consideración. Otra limitación importante es la *multilingualidad*. Aunque la lengua mayoritaria en el *web* es el inglés, es obvio que no es la única. La Recuperación de Información Multilingüe viene siendo objeto de investigación desde hace varios años. Así, es objeto preferente de las conocidas conferencias *TREC* o *CLEF*. Una revisión amplia del tema puede verse en [12] y [13].

4.2 Estudio de enlaces

La similitud documental también puede abordarse desde el punto de vista de los enlaces, consiguiendo con ello eliminar el problema de la multilingualidad. Además la recuperación basada exclusivamente en los enlaces parece tener una efectividad notable, como se puede deducir de los trabajos de [14] y [15], entre otros. La similitud dependiente de los enlaces ha sido definida en [16] como

$$sim_{ij}^{link} = \frac{link_{ij}}{\sum_{k=1}^{N} link_{ij}}$$

donde $link_{ij}$ es el número de enlaces desde el documento D_i a D_j en una colección de N documentos del Web. [17] aplica técnicas del análisis de cocitas para la recuperación basada en los enlaces.[15] propone algoritmos específicos que tratan este aspecto y que intentan encontrar soluciones operativas aplicables a la similitud documental.

5 Conclusiones

La mejora de los sistemas de recuperación en el *web* pasa por la consecución de resultados más precisos. Una posible vía es la Navegación Autónoma, la cual debe enfrentarse a diversos problemas, como la elección de las mejores rutas de exploración y la selección de páginas relevantes. Mecanismos como el análisis de los enlaces hipertextuales o la exploración de grafos pueden ser de gran utilidad para abordar estos problemas.

References

1. Hubermann, B., Adamic, L.: Evolutionary dynamics of the world wide web. Technical report, Xerox Palo Alto Research Center (1999)
2. Chen, H., Zhang, Y., Houston, A.: Semantic indexing and searching using a hopfield net. Technical report, Dep. of MIS, College of Business and Public Administration, Univ. of Arizona, Tucson, AZ (1997)
3. Chen, H. , Houston, A.S.R., Schatz, B.: Internet browsing and searching: User evaluations of category map and concept space techniques. JASIS **49** (1998) 582–603
4. Lawrence, S., Giles, C.: Searching the world wide web. Science (1998) 98–100
5. Chowdhury, G.G.: The internet and information retrieval research: a brief overview. Journal of Documentation **55** (1999) 209–225
6. Wooldridge, M., Jennings, N.R.: Intelligent agents: Theory and practice. Knowledge Engineering Review **10** (1995) 115–152
7. Cho, J. , García-Molina, H., Page, L.: Efficient crawling through url ordering. In: Procs. of the 7 WWW Conference. (1998)
8. Bharat, K. , Broder, A. , Henzinger, M. , Kumar, P., Venkatasubramanian, S.: The connectivity server: fast access to linkage information on the web. In: Procs. of the 7 Internet. WWW conference, Brisbane, Australia (1998)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report (1998)
10. Ellis, D., Furne-Hines, J., Willet, P.: On the creation of hypertext links in full-text documents: Measurement of inter-linker consistency. Journal of Documentation (1994) 67–98
11. Page, L., Brin, S.: Page rank, an eigenvector based ranking approach for hypertext. In: SIGIR 98, Melbourne, Australia (1998)
12. Oard, D., Dorr, B.: A survey of multilingual text retrieval. Technical report UMIACS-TR-9619 (1996)
13. Oard, D.W., Dorr, B.J., Hackett, P.G., Katsova, M.: A comparative study of knowledge-based approaches for cross-language information retrieval. Technical Report CS-TR-3897 (1998)
14. J.L. Alonso Berrocal, C.F., Rodríguez, A.Z.: Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de la información. In: IV Encuentros Internacionales sobre Sistemas de Información y Documentación: IBERSID 99, Zaragoza (1999) 15–18
15. Dean, J., Henzinger, M.: Finding related pages in the world wide web. WWW8 / Computer Networks **31** (1999) 1467–1479
16. Chen, C.: Structuring and visualizing the www by generalized similarity analysis. In: Proceedings of Hypertext'97, Southampton, UK (1997) 177–186
17. Cui, L.: Rating health web sites using the principles of citation analysis: a bibliometric approach. Journal of Medical Internet Research **1** (1999)