Agentes Inteligentes: Recuperación Autónoma de Información en el Web.

Intelligent Agents: Autonomous Information Retrieval in the Web.

José L. Berrocal, Carlos G. Figuerola, Ángel F. Zazo y Emilio Rodríguez

Grupo de Recuperación de Información
Departamento de Informática y Automática
Facultad de Documentación, Universidad de Salamanca
C/ Francisco Vitoria, 6-16, 37008-Salamanca (ESPAÑA)
e-mail: [berrocal|figue|afzazo|aldana]@usal.es

Resumen: 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.

Palabras clave: Recuperación de información, World Wide Web, Agentes.

Abstract: The problem of the information retrieval in the Web can be raised from different points of view, with mechanisms like the feedback by relevance, the use of thesauri, the study of the hyperconnections, or the application of neuronal networks, among others.

All these mechanisms are applied on great data bases constructed from the previous

exploration of more or less ample sectors of the Web.The experience has demonstrated that the precision of these systems is low, and the recall is relativized to the explored sector. Another approach to the problem that it tries to obtain precise results much more, although without persecuting discharges rates of recall exists nevertheless, being based on the use of intelligent agents who track the network according to the informative necessities of the user. The characteristics of the agents are indicated and some of the properties and desirable abilities for those agents dedicated to the information retrieval in the Web are analyzed.

Keywords: Information Retrieval, World Wide Web, Agents.

1. Introducción

Uno de los fenómenos más importantes de los últimos años en el campo de la Información es el desarrollo y espectacular crecimiento de Internet, especialmente de lo que conocemos como Web. El número de páginas crece exponencialmente y afecta a todos los ámbitos del conocimiento (1) (2). En el terreno científico, por ejemplo, muchas de las fuentes de información tradicionales como revistas, actas de congresos, etc. se encuentran ya en la red. Estamos observando como la red favorece la aparición de nuevos tipos de fuentes de información (3). Incluso algunas de ellos ya sólo se encuentran allí, habiendo abandonado los soportes tradicionales.

Este hecho pone de relieve el problema de la Recuperación de Información en la red, y más específicamente en el Web. Básicamente, los sistemas de recuperación en el Web utilizan dos mecanismos, que no son excluyentes entre sí y que pueden utilizarse de forma combinada. Uno es la búsqueda mediante palabras clave. El otro es la clasificación en clases o categorías de páginas web (4).

En el caso conocido de las búsquedas por palabras clave se pueden aplicar diversas técnicas que mejoren los resultados, tendentes fundamentalmente a superar o aminorar lo que algunos autores han denominado como la *barrera semántica* (5): la manera de expresar una misma idea o concepto

difiere de unas personas a otras. Buena parte de tales técnicas tienen que ver con la expansión de la consulta o adición de nuevos términos a las palabras clave de la búsqueda, en especial la realimentación por relevancia (6) (una revisión de estos métodos puede encontrarse en (4)). Otra posibilidad es el uso de tesauros. Éstos pueden ser elaborados previamente de forma manual por lo general, y estar especializados en algún dominio del conocimiento concreto, o bien pueden construirse de forma más o menos automática. Para esto último se aplican mecanismos de análisis de similitud o distancia entre términos (7). Entre las técnicas utilizadas podemos citar los denominados *tesauros de similitud* (8) (puede verse una aplicación de los mismos en (9)), el análisis de *cluster* para agrupar automáticamente palabras relacionadas (10), o la utilización de redes neuronales para la obtención de términos cercanos o relacionados (11).

Por lo que se refiere a las búsquedas mediante categorización previa, dicha clasificación suele efectuarse manualmente, aunque hay experiencias interesantes de clasificación automática de páginas web, como por ejemplo el proyecto WEBSOM (http://websom.hut.fi/websom/), en el que se utilizan mapas autoorganizativos (12) para establecer categorías de términos. Esas categorías se emplean para definir o representar vectores de las páginas web, con los cuales se puede construir, aplicando el mismo mecanismo, un mapa visual de las propias páginas colocadas en función de su similitud. El usuario puede utilizar dicho mapa para, una vez seleccionada alguna de las páginas, obtener las más cercanas o más relacionadas con ella (13).

Otros proyectos de categorización utilizan modelos basados en las características de los enlaces hipertextuales, incluso algunas veces en combinación con técnicas de redes neuronales. En (14, sección 9) puede encontrar un resumen de estas técnicas. También hay que destacar el programa Inxight Star y su aplicación a la producción de mapas temáticos (http://www.inxght.com/products/st_studio). Sin embargo, bien se trate de búsquedas mediante palabras clave o a través de categorización previa de páginas, o de una combinación de ambos métodos, cualquiera de estos sistemas parte de la existencia de una base de datos que contenga la

colección de páginas web. Dejando de lado los sistemas especializados y circunscritos a ámbitos muy concretos del conocimiento, los sistemas generales de búsqueda tienden a manejar bases de datos muy grandes, puesto que, como es bien sabido, el Web crece día a día. En general, la experiencia de los usuarios de tales sistemas de búsqueda muestra claramente que se producen respuestas de muy baja precisión (15). 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 (16), el hecho es que respuestas con un número tan alto de páginas encontradas producen en el usuario el fenómeno bien conocido de la *sobrecarga de información (desbordamiento cognitivo)*.

Existe sin embargo otra aproximación al problema que, sin perseguir altas tasas de exhaustividad, pretende obtener resultados mucho más precisos, basándose para ello en el uso de *agentes inteligentes*.

2. Agentes Inteligentes

Los agentes inteligentes son programas que han sido definidos por diversos autores; en muchos casos de manera lo suficientemente poco precisa como para que se produzca discusión acerca de si tal o cual sistema es o no un agente (17). Las clasificaciones que podemos hacer de los agentes son variadas y dependen de las diferentes características que se tengan en cuenta. En la siguiente figura se presenta una clasificación de los mismos:

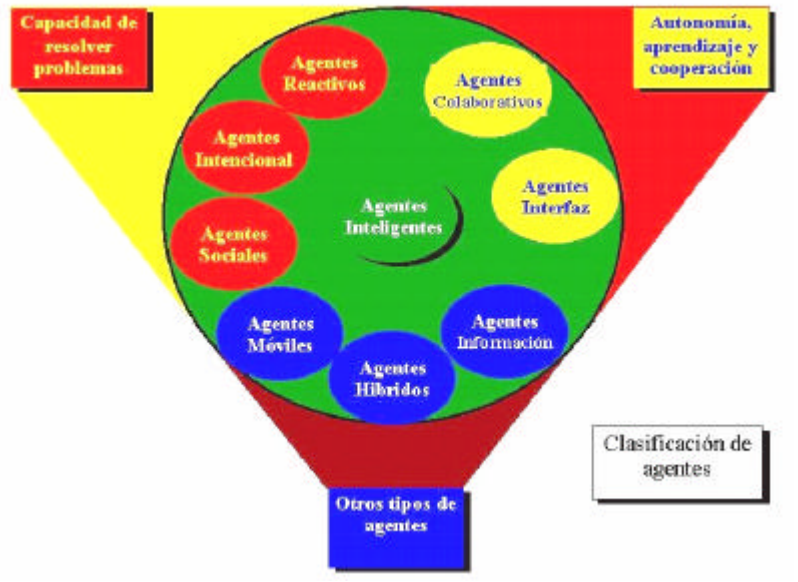

Figura 1

Además hay que tener presente la existencia del internet invisible o Infranet (18, 19) entendiendo que es el conjunto de recursos accesibles solamente a través de pasarelas o formularios o bien contenidos de creación dinámica (por ejemplo consultas a bases de datos) y que por ello están inaccesibles para los robots (20). Los agentes intentan ofrecer una solución a este tipo de situaciones, fundamental pues en muchos casos son de vital importancia en la recuperación de información por la calidad de sus contenidos.

A pesar de ello, casi todo el mundo está de acuerdo en que los agentes inteligentes tienen una serie de características que los definen como tales. Tal vez el problema esté en que, de un lado, no siempre están presentes todas esas características, y de otro, buena parte de dichas características son lo suficientemente ambiguas como para admitir diferentes interpretaciones acerca de su contenido real y concreto (21). En adición, el estado todavía embrionario de muchos proyectos, y los intereses comerciales de muchas empresas que proclaman de forma abusiva producir y vender auténticos agentes inteligentes, son factores que añaden más confusión.

A grandes rasgos las características típicas de un agente inteligente son:

- Autonomía. Los agentes deben trabajar sin supervisión humana, al contrario que los

programas que operan en base a interfaces de manipulación directa por parte del usuario. Así, una vez fijadas las condiciones y restricciones necesarias por parte del usuario, se espera que el agente intente cubrir o conseguir sus objetivos, dejando ocultos los detalles para dicho usuario (22).

- Inteligencia. En (23) se describe esta característica y se mencionan distintos escenarios y modos de concretar esta inteligencia, pero parece claro que el concepto de *inteligencia* puede entenderse de maneras muy distintas.

- Cooperación. Un agente debería ser capaz de colaborar con otros agentes, intercambiando informaciones y resultados de acciones propias. La negociación puede hacerse con agentes que persigan los mismos objetivos (de manera que, por ejemplo, objetivos ya logrados por un agente podrían ser cedidos a otro agente que persigue lo mismo), o con agentes de objetivos diferentes, pero necesarios o al menos útiles para lograr las metas del primero. La capacidad de cooperación requiere disponer de algún mecanismo que permita la negociación entre agentes, aún cuando éstos sean heterogéneos. Se han propuesto diversos estándares, siendo probablemente el más difundido el conocido como KQML (24) (25).

- Comunicación. Esto no sólo implica la simple capacidad para comunicarse con el usuario, sino también la necesidad de tener conocimiento sobre el mundo o dominio sobre el cual opera el agente. Habitualmente este conocimiento se implementa mediante ontologías y reglas de inferencia (26).

- Reactividad. Un agente debería poder responder ante eventos, tomando sus propias decisiones, incluso modificando su manera de operar, siempre en vista a la consecución de sus metas.

- Adaptitividad. Un agente debería poder aprender de experiencias pasadas (y de la experiencia de otros agentes), así como de las reacciones del usuario ante resultados

previos. Esto está directamente relacionado con el *aprendizaje de máquina* o *aprendizaje automático*.

3. Agentes inteligentes y recuperación de información en el Web

Una de las tareas obvias que podría ser abordada mediante agentes es la exploración automática del WWW, con el fin de recuperar las páginas relevantes para unas necesidades informativas determinadas. De esta forma, la formulación de tales necesidades sería parte de las especificaciones iniciales proporcionadas al agente; éste 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, sin requerir de nuestra presencia.

Puesto que la propia exploración del Web, manual o automática, requiere grandes cantidades de tiempo, un enfoque de este tipo tiene de entrada algunas limitaciones. No es esperable una respuesta inmediata, ni siquiera probablemente con la agilidad suficiente para plantear una dinámica especialmente interactiva con el usuario. Antes bien, y muy en la línea de lo que entendemos por agentes inteligentes, de alguna forma el usuario *delega* en el agente, después de haberle facilitado algunas instrucciones (por ejemplo, indicándole qué clase de información se desea). Se deja al agente hacer su trabajo de forma autónoma y tomándose su tiempo, en espera de que en un plazo razonable (el propio usuario podría establecer plazos máximos) entregue el resultado de su trabajo, esto es, las páginas web encontradas útiles para satisfacer las necesidades de información expresadas por el usuario.

La otra limitación importante de este enfoque es la renuncia implícita a la exhaustividad. Dado el tamaño del Web, parece claro que la exploración completa, o incluso de una parte significativa de él, resulta implanteable; antes al contrario, agentes de este tipo trabajando para usuarios individuales o personales, por ejemplo, explorarían tan sólo una pequeña parte del Web. Se espera, en contrapartida,

que los resultados obtenidos alcancen una notable precisión. Esta clase de agentes permitirían obviar el efecto de *sobrecarga de información* aludido más arriba.

Aceptando estas limitaciones, los agentes a que nos referimos deben resolver una serie de cuestiones, para las que se han propuesto diversas soluciones, la mayor parte no excluyentes entre sí. Examinaremos a continuación los problemas más importantes.

3.1 La elección de los puntos de partida

Puesto que un agente de este tipo debe explorar gran cantidad de páginas, es preciso determinar algún punto de partida. Habitualmente se suele representar el Web como un grafo dirigido, en el que las diferentes páginas son los nodos, y los enlaces son los arcos de tal grafo. El proceso de exploración parte de un nodo y, utilizando los arcos o enlaces, conduce y explora otros nodos, y así sucesivamente. Como la distancia entre el nodo por el que se empieza a explorar y cualquiera de los nodos relevantes puede ser muy grande, es crítico localizar previamente nodos o puntos de partida que puedan estar lo más cercanos posible a nodos o páginas relevantes para las necesidades de información del usuario.

La distancia a recorrer (el número de nodos por los que hay que pasar) no sólo depende del tamaño del Web, sino que incluso podemos encontrar nodos con vías muertas que se extinguen sin permitir proseguir con la exploración.

Un enfoque utilizado frecuentemente para elegir buenos puntos de partida es comenzar el trabajo del agente con una búsqueda al estilo clásico en las bases de datos de diferentes buscadores convencionales. En estos casos tales búsquedas previas suelen enviarse a servicios *metabuscadores* (27), 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. En este caso sería el propio agente quien enviaría la consulta a esos *metabuscadores*, recogiendo las páginas devueltas por éstos. Tales páginas son las candidatas a ser puntos de entrada o de comienzo de exploración.

Dichos puntos de entrada pueden manejarse de forma secuencial, empezando la exploración

por uno de ellos, hasta una determinado distancia prefijada de antemano, o en paralelo, utilizando varios agentes para ello. En este caso los agentes deben hacer uso de sus capacidades cooperativas, no sólo para compartir criterios de selección de páginas relevantes, sino también para evitar exploraciones de los mismos nodos.

La exploración de la red con varios agentes tomando diferentes puntos de entrada ofrece el atractivo de permitir utilizar procesamiento paralelo o varios ordenadores para el proceso (22), pero incluso sin ello presenta la ventaja de obviar en alguna medida problemas derivados de las comunicaciones, como cuellos de botella, líneas o servidores lentos, etc., redundando en una mejora en el tiempo de respuesta.

De otro lado, el hecho de disponer de varios puntos de entrada puede implicar la selección de parte de ellos (en un número razonable), así como posiblemente la priorización. Hay diversas estrategias automáticas para abordar esta cuestión, desde tomar simplemente los *n* primeros, hasta aplicar medidas de similitud (que trataremos luego) entre las especificaciones del usuario y el contenido de las páginas, pasando por el análisis de aspectos como el número de enlaces de cada punto de entrada, o incluso prospecciones de tiempos de respuesta. Del mismo modo, es posible una realimentación por parte del usuario, dejando que sea éste quien seleccione los que estime como mejores puntos de entrada. Naturalmente, estos diversos enfoques son combinables entre sí.

3.2 Activación de enlaces

Dado un punto o página de partida, un agente que pretenda explorar el Web debe extraer los enlaces (direcciones URL) 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. Si la exploración se ha de llevar a cabo por varios agentes de forma cooperativa, esa lista debería ser compartida en alguna forma, a fin de no duplicar exploraciones de los mismos nodos.

El almacenamiento y posterior seguimiento de todos los enlaces en la lista llevaría,

teóricamente, a la exploración de todo el Web. Sin embargo, como suele tenerse limitaciones de recursos de almacenamiento, capacidad de proceso o comunicaciones, etc., y especialmente de tiempo, se hace preciso establecer un orden de prioridad para los elementos de la lista. Este orden atiende a dos premisas fundamentales: en primer lugar, la relevancia de los enlaces (o su presunción) 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.

Empezando por este último aspecto, se han propuesto diversos sistemas para seleccionar aquellos enlaces más prometedores desde ese punto de vista. Para determinar la importancia de una página, una posibilidad consiste en utilizar los *backlinks* de la misma, esto es, las páginas que tienen enlaces hacia la página en cuestión (28). El mecanismo más simple es contar el número de *backlinks*, pero el problema es disponer de dicha información. En este sentido, cabe mencionar el proyecto *Compaq's Connectivity Center Server* (29) en estrecha relación con *Altavista* (http://www.altavista.com), o el también muy conocido buscador *Google* (http://www.google.com) (30).

Más sofisticado que el simple recuento de *backlinks* es el algoritmo conocido como *PageRank* (31). La idea básica es que la importancia de un nodo o 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. Según este algoritmo, el cálculo del *PageRank* ha de hacerse de forma iterativa, asignando de antemano pesos a determinados nodos o páginas, ya sea de forma aleatoria o en función de algún otro criterio, y asumiendo que, de una forma u otra, en algún momento de la computación se llega a esos nodos. Se trata de una visión muy genérica, en la que hay que resolver otros detalles, pero lo que importa resaltar aquí es que se trata de un cálculo costoso en términos de 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. (32). Parece que tales índices no son aplicables en una exploración directa del Web, aunque algunos buscadores –basados en búsquedas en bases de datos de páginas web previamente recopiladas– los utilizan para ordenar los resultados obtenidos en una búsqueda de este tipo (30).

3.3 Selección de páginas por contenido

Más allá de la mayor o menor importancia de una página (en el sentido de la mayor o menor facilidad de exploración del Web a partir de la misma), lo que realmente nos interesa es disponer de medios para estimar la proximidad de un nodo a las necesidades informativas del usuario. Esto debe permitir, naturalmente, seleccionar páginas para que el agente las entregue al usuario 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 o compaginarse entre sí.

3.3.1 Técnicas de recuperación de información

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 de información expresadas por el usuario. En realidad, a través de estas técnicas lo que comparamos es el *texto* de las páginas web. Entre los procedimientos más conocidos está el llamado modelo vectorial (33), el cual es también representativo de las limitaciones que presenta la aplicación de estos métodos.

El modelo vectorial opera con palabras o términos y calcula para cada uno de éstos un peso o índice que trata de expresar la importancia de la palabra en cuestión. Este cálculo se efectúa en base a las frecuencias de las palabras, estimando que el peso es directamente proporcional a la frecuencia de aparición de la palabra en el documento, pero inversamente proporcional a la frecuencia de aparición

de la palabra en toda la colección de documentos (34). Ahora bien, en una exploración directa del Web el segundo factor es desconocido. Ello conlleva a que se opere con vectores binarios sin pesar las palabras, o estimar el peso de las mismas exclusivamente en función de la frecuencia en cada página. En cualquiera de estos casos, la eficiencia del sistema se resiente.

Otra de las limitaciones importantes es la que deriva de la *multilingualidad*. Aunque parece claro que la lengua mayoritaria en el Web es el inglés, es obvio que no es la única. Además, por otra parte, si el usuario no es angloparlante preferirá expresar sus necesidades informativas en su propia lengua, aunque en muchos casos, a pesar de ello, puede aceptar también como relevantes páginas en otro idioma. La Recuperación de Información Multilingüe viene siendo objeto de investigación desde hace varios años. Así, es objeto preferente de las conocidas conferencias *Text REtrieval Conference* (TREC), o más recientemente de las *Cross Lingual European Forum* (CLEF). Una revisión amplia del tema, aunque tal vez algo antigua, puede encontrarse en (35) y, más parcial, pero también algo más reciente, en (36).

3.3.2 Estudio de enlaces

La similitud documental también puede abordarse, como ya se ha indicado, desde el punto de vista de los enlaces, obviando el contenido de las páginas a las que apuntan y consiguiendo con ello eliminar los problemas relacionados con la *multilingualidad.* Además la recuperación basada exclusivamente en los enlaces de las páginas web parecen tener una efectividad digna de tener en cuenta como se puede deducir de los trabajos de (37) y (38), entre otros.

La similitud dependiente de los enlaces ha sido definida en (39) como

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

donde $link_{ij}$ es el número de enlaces desde el documento D_i a D_j en una colección de N documentos del Web.

En (40) se aplican las técnicas del análisis de cocitas para la recuperación de información basada en los enlaces y ha originado la creación de algoritmos (38) específicos que tratan este aspecto y que intentan encontrar aplicaciones operativas aplicables a la similitud documental.

4. Conclusiones.

La mejora de los mecanismos de trabajo actuales y su aplicación en la recuperación de información deberían integrarse en el empleo de los agentes, con el fin de conseguir aplicaciones muy precisas en sistemas de usuario. A pesar del problema de los relativamente largos tiempos de espera, que podrían mejorarse con la cooperación entre agentes, se conseguirían tasas de precisión muy elevadas.

Los mecanismos actuales como el análisis de los enlaces hipertextuales, que han demostrado su efectividad, permitirían, entre otros resultados, obviar en gran medida la situación multilingüe del Web, sin olvidar los mecanismos clásicos como la exploración de grafos, o la aplicación de modelos habituales muy usados en recuperación de información.

Para la utilización de agentes en la recuperación de información en el Web se necesita la determinación clara y precisa de aspectos tan importantes como la designación de los puntos (páginas) de partida, y la selección de enlaces (páginas, de nuevo) en el camino de la selección de los mejores candidatos a ser explorados. Ambos aspectos se han tratado en este artículo. Todo ello redundaría en conseguir no sólo valores altos de precisión, sino también, dentro de la porción de Web explorada, mejores definiciones para conseguir valores altos de exhaustividad.

5. Bibliografía.

1. HUBERMAN, B.A. y ADAMIC, L.A.. Evolutionary dynamics of the World Wide Web. *Tech. Rep. Xeros Palo Alto Research Center*, (February, 1999).

2. HOBBES ZAKON, R.. Hobbe's Internet Timeline v5.2., 2000.
URL: http://www.zakon.org/robert/internet/timeline

3. SHIRI, A.A.. Cybermetrics: a new horizon in information research. *49th FID Conference and Congress*, 1998, 11-17 october , New Delhi, India, 1998

4. CHEN, H.; ZHANG, Y. y HOUSTON, A.L.. Semantic indexing and searching using a hopfield net. Technical report, Dep. of MIS, College of Business and Public Administration, Tucson, AZ, 1997.

5. NADIS, S.. Computation cracks semantic barrier between databases. *Science*, 1997, vol. 272, p. 1419

6. SALTON, G.. On the relationship between theoretical retrieval models. In *Informetrics 87/88*, Diepenbeeck (Bélgica), 1987, p. 263-270

7. CROUCH, C.J.. An approach to the automatic construction of global thesauri. *Information Processing & Management*, 1990, vol. 26, p. 629-640

8. QIU, Y.; FREI, H. Concept based query expansion, *SIGIR 93*, 1993, p. 160-169

9. FIGUEROLA, C.G.; ALONSO BERROCAL, J.L. y ZAZO RODRÍGUEZ, A.F.. El contenido semántico de los enlaces de las páginas web desde el punto de vista de la recuperación de información. *I Jornada de Terminología i Documentació*, 2000, Barcelona, Maig de 2000.

10. RASMUSSEN, E.. Clustering algorithms. En: FRAKES, W.B. y BAEZA-YATES, R. (editors). *Information Retrieval: Data structures and algorithms*. Prentice-Hall, 1992.

11. CHEN, H. y otros. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of American Society for Information Science*, 1998, vol. 49, nº 7, p. 582--603

12. KOHONEN, T.. Self-Organizing Maps, Springer Series in Information Sciences, vol. 30

13. KASKI, S. y otros. WEBSOM--self-organizing maps of document collections, *Neurocomputing*, 1998, vol. 21, p. 101-117

14. SEBASTIANI, F. Machine Learning in Automated Text Categorisation. *EDCL 2000*. 2000, Lisboa, Portugal, Septiembre, 2000

15. CHEN, H. y otros. An intelligent personal spider (agent) for dynamic internet/intranet searching. *Decision Support Systems*, 1998, vol. 23, nº 1, p. 41-58

16. LAWRENCE, S. y GILES, C.L. Searching the world wide web. *Science*, 1998, vol. 280, p.: 98-100

17. PETRIE, C.J. Agent-based engineering, the web, and intelligence. *IEEE Expert*, 1996, vol. 11, nº 6, p. 24-29

18. AGUILLO, I.F. Infranet. Buscar en los escondrijos de Internet. *Net Conexión*, 1997, 15, p. 55-56

19. CORNELLA, A. La infranet: ¿dónde está el valor?. *El profesional de la información*, 1999, vol. 8, nº 5, p. 3

20. AGUILLO, I.F. Internet invisible o Infranet: definición, clasificación y evaluación, *VII Jornadas Españolas de* Documentación, 2000 (Bilbao, Octubre 19-21), p. 249-269

21. FRANKLIN, S. y GRAESSER, A. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures and Languages*, Springer-Verlag, 1996, p. 21-35

22. WOOLDRIDGE, N.R. y JENNINGS, M.. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 1995, vol. 10, nº 2, p. 115-152

23. ROSELER, M. y HAWKINS, D.. Get agents: Software servants for an electronic information world (and more!). *ONLINE*, 1994, July, p. 19-32

24. FINIM, T. Umbc kqml web.
URL: http://www.cs.umbc.edu/kqml/. (Consultado el 30/11/2001)

25. LABROU, Y. y FININ, T. A proposal for a new kqml specification. Technical report, Computer Science and Electrical Engineering Department, University of Maryland, Baltimore, MD 21250, 1997.
URL: http://www.cs.umbc.edu/ jklabrou/publications/tr9703.ps

26. HENDLER, J. Is there an intelligent agent in your future? *Nature*, 1999, 11 Macrh

27. CHOWDHURY, G.G.. The internet and information retrieval research: a brief overview. *Journal of Documentation*, 1999, vol. 55, nº 2, p. 209-225

28. CHO, J.; GARCÍA-MOLINA, H. y PAGE, L. Efficient crawling thorugh url ordeirng.
URL: www-db.stanford.edu/pub/papers/efficient-crawling.ps.gz/cho98efficient.ps

29. BHARAT, K. y otros. The connectivity server: fast access to linkage information on the web. In *Procs. of the 7 Internet. WWW conference*, 1998, Brisbane, Australia

URL: http://www7.scu.edu.au/programme/fullpapers/1938/com1938.htm

30. BRIN, S. y PAGE, L. The anatomy of a large-scale hypertextual (Web) search engine. *Computer Networks and ISDN Systems*, 1998, vol. 30, nº 1-7, p. 107-117

URL: citeseer.nj.nec.com/brin98anatomy.html

31. PAGE, L. y otros. The pagerank citation ranking: Bringing order to the web. Technical report, 1998.

URL: citeseer.nj.nec.com/page98pagerank.html

32. ELLIS, D.; FURNER-HINES, J. y WILLETT, P. On the creation of hypertext links in full text documents: measurements of inter-linker consistency. *Journal of Documentation*, 1994, vol. 50, nº 2, p. 67-98

33. SALTON, G. y MCGILL, M.. *Introduction to Modern Information Retrieval,* New York: McGraw-Hill, 1983

34. HARMAN, D.. Ranking *Algorithms, en Information Retrieval. Data Structures and Algorithms*, NJ: Prentice Hall, Upper Saddle River, 1992, p. 363-392

35. OARD, D. y DORR, B.J. A survey of multilingual text retrieval, Technical report UMIACS-TR-9619, Univ. of Maryland, 1996.

URL: http://www.ee.umd.edu/medlab/mlir/mlir.html

36. OARD, D. y otros. A comparative study of knowledge-based approaches for cross-language information retrieval. Technical Report CS-TR-3897, 1998.

URL: http://citeseer.nj.nec.com/44309.html

37. ALONSO BERROCAL, J.L.; FIGUEROLA, C.G. y ZAZO RODRÍGUEZ, A.F.. Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información. *IV Encuentros Internacionales sobre Sistemas de Información y Documentación*: IBERSID 99, Zaragoza, 15-18 de Marzo de 1999.

38. DEAN, J. y HENZINGER, M.R. Finding related pages in the World Wide Web, *WWW8 / Computer Networks*, 1999, vol. 31, nº 11-16, p. 1467-1479

URL: http://citeseer.nj.nec.com/dean99finding.html

39. CHEN, C. Structuring and visualizing the WWW by generalized similarity analysis. *Proceedings of Hypertext'97*. 1997, Southampton, UK, p. 177-186

40. CUI, L. Rating Health Wen sites using the principles of citation análisis: a bibliometric approach. *Journal of Medical Internet Research*, 1999, vol. 1, nº 1