Análisis de conectividad en la recuperación de información web

Cristian Merlino-Santesteban
Centro de Documentación. Facultad de Ciencias Económicas y Sociales. Universidad Nacional de Mar del Plata. Argentina.
E-mail: csantest@mdp.edu.ar

Resumen

En este trabajo se presenta una introducción a los patrones de enlace, brindados por la estructura de red de la World Wide Web, como una nueva fuente de información para la recuperación de información efectiva y eficiente.
Se describen sus características y tres tipos de algoritmos de ordenación por relevancia basados en el análisis de conectividad.

Palabras clave

Análisis de enlaces; Recuperación de información; Topología de red; World Wide Web.

Connectivity analysis in web information retrieval

Abstract

This paper deals with an introduction to link patterns, presented by the World Wide Web hyperlink structure, as a new source of information for efficient and effective information retrieval. It describes the main characteristics and three types of connectivity-based ranking algorithms.

Keywords

Link analysis; Information retrieval; Network topology; World Wide Web.

INTRODUCCIÓN

La World Wide Web (WWW, W3 o Web) es una red hipertextual de enorme complejidad que crece a un ritmo acelerado. Esta inmensa estructura provee patrones de conectividad que pueden mejorar sustancialmente los métodos de búsqueda y ordenación por relevancia de los sistemas de búsqueda tradicionales.

Inspirado en el estudio de redes sociales y en el análisis de citaciones de la literatura científica, el uso de la estructura de enlaces o hipervínculos ha emergido recientemente como un nuevo y promisorio acercamiento a la recuperación de información efectiva y eficiente[1, 2, 3, 4, 5].

Una citación provee una relación entre dos o más documentos y, con frecuencia, es la única manera de localizar textos pertinentes relacionados con el tema del documento citante. Un enlace de una página web sirve para un propósito similar, pero hay importantes diferencias entre una citación y un enlace web:

– Una citación en la literatura científica es estática y unidireccional. Una vez publicado un trabajo no hay manera de incorporar nuevas referencias. Por esta razón, es raro encontrar artículos que se citen recíprocamente. En cambio, en las páginas web se pueden agregar, modificar, actualizar o remover enlaces posteriormente a su creación.

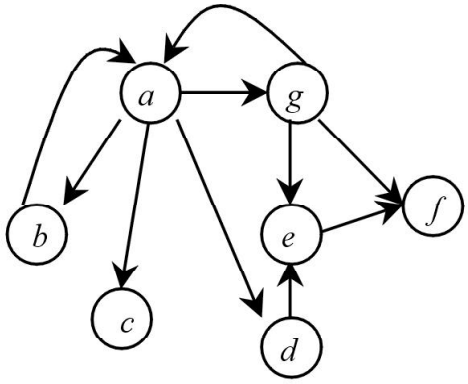
– La referencia web generalmente es más subjetiva y menos relevante que en la literatura científica[6, 7]. Los creadores de documentos web (comúnmente los mismos usuarios) muchas veces no toman en cuenta la calidad, relevancia u objetividad de la información. Se podría decir que la creación de enlaces es un fenómeno esencialmente anárquico.

– Mientras algunos enlaces en una página web pueden dirigir a documentos relacionados semánticamente (o no relacionados), otros son creados por razones que no tienen que ver con este asunto. Muchos vínculos existen por propósitos puramente navegacionales (retroceder, ir al inicio) o publicitarios/comerciales (gane dinero ya, compre autos baratos, descargue la nueva versión de ...).

PATRONES DE ENLACE

La WWW puede ser vista estructuralmente como un grafo dirigido o digrafo, donde cada página web es un nodo o vértice y cada enlace es un arco o arista (figura 1). Aunque los documentos web ofrecen información textual, la conectividad del digrafo, además de permitir la navegación, puede ayudar a caracterizarlos. Esa caracterización se logra mediante el procesamiento de su conectividad en matrices de adyacencia, en las que se indica con un valor 1 la existencia de un arco y con un 0 la inexistencia del mismo, como muestra la figura 2.

FIGURA 1
La Web representada a través de un digrafo



Un hipervínculo en una página web que conecta un documento con otro y representa un aval implícito con la página destino. Si un vínculo de la página p_a enlaza a la página p_b, el autor de p_a (por suposición) está recomendando a p_b. Si este enlace no estuviera presente la probabilidad que ambas páginas tratasen el mismo o parecido tema sería mucho menor. Esta presunción deriva principalmente de la teoría hipertextual donde el contenido externo introducido por los hiperenlaces tiende a ser de alta calidad y utilidad*.

FIGURA 2
Matrices de adyacencia del grafo de la figura 1

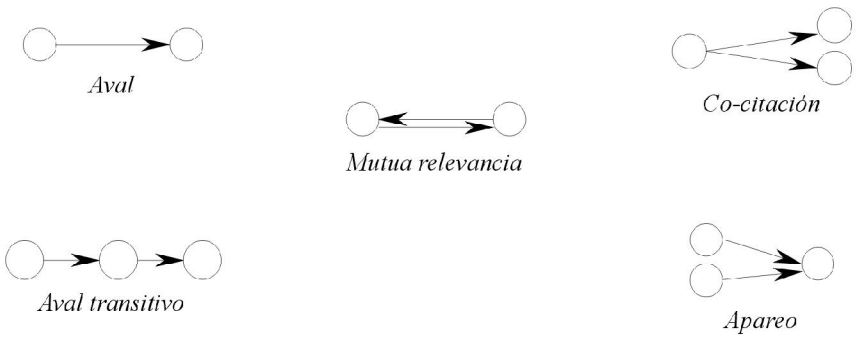
	a	*b*	*c*	*d*	*e*	*f*	*g*
a	-	1	1	1	0	0	1
b	1	-	0	0	0	0	0
c	0	0	-	0	0	0	0
d	0	0	0	-	1	0	0
e	0	0	0	0	-	1	0
f	0	0	0	0	0	-	0
g	1	0	0	0	1	1	-

Arcos salientes

	a	*b*	*c*	*d*	*e*	*f*	*g*
a	-	1	0	0	0	0	1
b	1	-	0	0	0	0	0
c	1	0	-	0	0	0	0
d	1	0	0	-	0	0	0
e	0	0	0	1	-	0	1
f	0	0	0	0	1	-	1
g	1	0	0	0	0	0	-

Arcos entrantes

FIGURA 3
Patrones básicos formados por dos enlaces



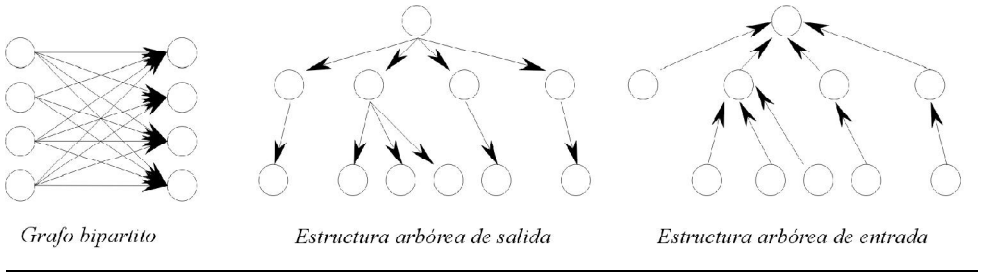
Cuando consideramos dos vínculos, obtenemos un variado número de patrones básicos (figura 3)[8]. Dos páginas apuntándose entre sí refuerzan nuestra intuición acerca de su mutua relevancia. Una página referenciando dos páginas distintas (co-citación) sugiere la probabilidad de que estén relacionadas en contenido. Dos documentos vinculando a un tercero producen un apareo (*coupling*), es decir, dos páginas están relacionadas mutuamente sin estar enlazadas entre sí. Un aval transitivo ocurre cuando la página p_g vincula a p_u, la cual a su vez enlaza a p_v. Transitivamente, p_g es considerada como aval de p_v aunque este sea un aval débil. Y por consiguiente estas estructuras pueden combinarse unas con otras formando patrones muy complejos que fortalecen la relación entre las páginas (figura 4).

De las múltiples combinaciones posibles, la generalización de la co-citación con la estructura arbórea de salida (*out degree*) y la generalización del apareo con la estructura arbórea de entrada (*in degree*) han sido de

* La colección, en el contexto de esta teoría, está constituida por un conjunto de documentos homogéneos que versan sobre un tópico determinado, enlazados entre sí con el propósito de referenciar información relevante.

particular interés para el campo de la recuperación de información, especialmente aquellas estructuras con grandes grados de conectividad a un nodo raíz.

FIGURA 4
Patrones complejos formados por numerosos nodos y vínculos



La estructura arbórea de entrada nos indica que si muchas páginas diferentes apuntan directamente o transitivamente a un nodo, es probable que éste sea un recurso valioso (autoridad) sobre algún tema o de interés compartido por las otras páginas. Esto es análogo a la medición del impacto de un artículo científico por el número de citaciones que recibe. El interés en la estructura arbórea de salida está dado por la suposición de que si un nodo vincula a muchas páginas web valiosas en un tópico, denominado *hub*, luego podemos considerarlo como una buena fuente para buscar información. En este sentido, un *hub*, es análogo a un artículo de revisión bibliográfica. Partiendo de estas interpretaciones podemos esperar que los documentos con alto valor de autoridad tengan contenido relevante mientras que los documentos con alto valor de *hub* tengan hiperenlaces a contenidos relevantes.

En este acercamiento al estudio de patrones, los algoritmos de conectividad, también incluyen en su análisis el criterio de cantidad e importancia de los enlaces entrantes y salientes dado que no todos los vínculos tienen el mismo peso, y la simple cuenta de citaciones a un documento no siempre implica importancia. En otras palabras, un documento web con un alto grado de citación y/o luminosidad podría verse como un documento muy importante, pero la calidad de los enlaces que apunta a otras páginas o recibe no es siempre la misma. En muchos casos, una página con unos pocos enlaces entrantes de calidad se juzga más visible que una página con centenares de vínculos de sitios menos valiosos.

ALGORITMOS DE ORDENACIÓN POR RELEVANCIA BASADOS EN LA CONECTIVIDAD

Básicamente, los algoritmos de ordenación por relevancia basados en la conectividad pueden dividirse en tres tipos: a) dependientes de la consulta del usuario, los cuales asignan una puntuación (valor de relevancia) a una página en el contexto de un proceso de equiparación; b) cuasi-dependientes de la consulta del usuario, los cuales otorgan una puntuación a una página en el entorno de un *corpus* documental (URL) dado; y c) independientes de la consulta del usuario, los cuales asignan una puntuación a una página independientemente de la búsqueda proporcionada. A diferencia de lo expuesto por Henzinger[9], se incorporó el tipo cuasi-dependiente porque un URL (Uniform Resource Locator) también constituye un punto de partida y brinda un contexto de búsqueda, su contenido.

Algoritmos dependientes de la consulta del usuario

Este esquema de ordenamiento tiene su implementación más popular en el algoritmo HITS (*Hyperlink-Induced Topic Search*) desarrollado por Kleinberg[5] como mejora al trabajo realizado por Carriere y Kazman[2]. La idea básica es colectar un conjunto de páginas iniciales sobre un tópico a partir de los resultados obtenidos de la búsqueda de una o más palabras en un sistema de recuperación de información, ampliar el mismo con la adición de las páginas que le enlazan y las páginas que son referenciadas por él, construir una matriz de adyacencia W y calcular el vector representativo de $A=W^TW$ y $H=WW^T$. Estos vectores corresponden a los pesos de los nodos autoridad y *hub* respectivamente.

Chakrabarti *et al.*[10] extendieron HITS en el ARC (*Automatic Resource Compiler*) incorporando el texto próximo a los anclajes en la computación de los pesos de los nodos autoridad y *hub*. La utilidad del texto alrededor de los anclajes había sido mostrada años antes por McBryan[11], pero fusionarlo con HITS fue un paso innovador. La creación del ARC fue ideada como un mecanismo para compilar y mantener automáticamente listas de recursos de alta "calidad". Chakrabarti *et al.*[12]continuaron mejorando el ARC en el proyecto CLEVER, el cual conjuga varios algoritmos de conectividad para potenciar su funcionalidad[13].

Un algoritmo alternativo en esta línea de investigación, SALSA (*Stochastic Approach for Link-structure Analysis*), fue

propuesto por Lempel y Moran[14]. SALSA emplea el mismo meta-algoritmo que HITS, pero asigna a cada nodo un peso autoridad proporcional a la suma de los pesos de sus vínculos entrantes y un peso *hub* proporcional a la suma de los pesos de sus vínculos salientes, es decir, SALSA calcula, para todas las páginas, la suma de los pesos de sus enlaces salientes y entrantes, y normaliza estos dos vectores. Este algoritmo resalta básicamente la "calidad" de una página a través de la ponderación de su conectividad dentro del subgrafo generado.

Chen *et al.*[15], yendo más allá del análisis de vínculos explícito, combinaron la información extraída de la estructura de enlaces implícita, generada por la interacción de los usuarios con las páginas web (*browsing logs*), y HITS en un nuevo algoritmo donde la importancia de las páginas y los usuarios se refuerza entre sí. La presunción subyacente del análisis de enlaces implícito es que cuanto más visitada sea una página por los usuarios, más importante es la página aunque como señalan Schapira[16] y Joachims[17] este mecanismo está sujeto a ciertas suposiciones en la conducta del usuario. En la figura 5 se observa la comparación de precisión @10 de este algoritmo, indicado como Chen, con los algoritmos de AltaVista*, DirectHit* * y HITS. La mejora promedio de precisión sobre DirectHit es de 25,3% y de 11,8% con HITS.

Borodin *et al.*[18], por su parte, testearon experimentalmente nueve algoritmos diferentes sobre el mismo conjunto de páginas iniciales y determinaron que ninguno es el mejor indiscutible, por el contrario, hay búsquedas donde el rendimiento de algunos algoritmos no es bueno y hay búsquedas donde el rendimiento de todos los algoritmos es bueno. También hallaron que ciertos algoritmos parecen ser más "equilibrados", mientras que otros más "enfocados". La figura 6 muestra la comparación de precisión @10 para 5 preguntas y refleja el comportamiento mencionado.

FIGURA 5
Comparación de precisión @10 de 4 algoritmos (Chen *et al.*[15])

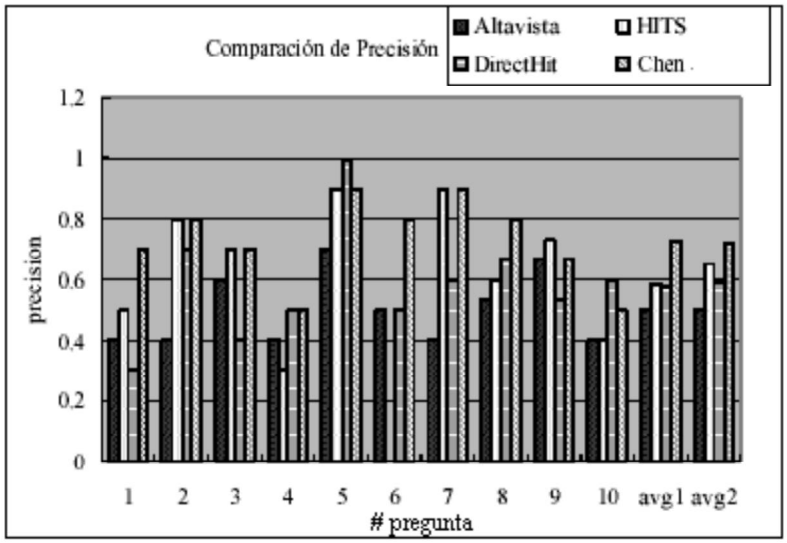
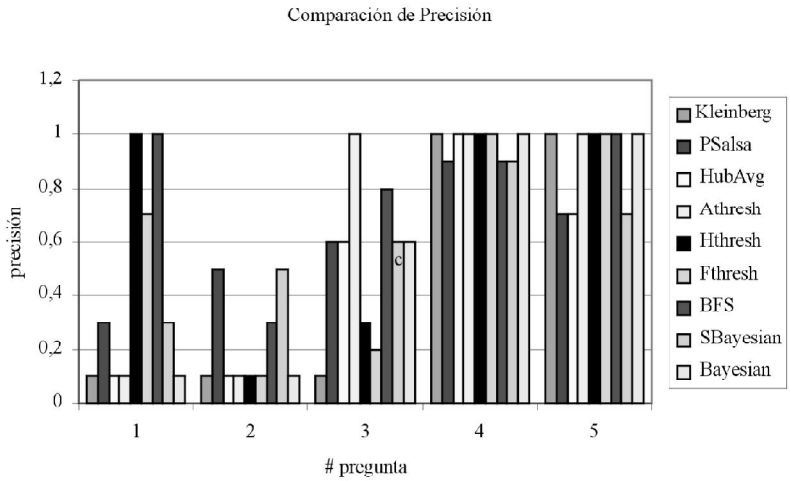


FIGURA 6
Comparación de precisión @10 de 9 algoritmos
(Elaboración propia con datos de Borodin *et al.*[18])



Referencias:
Kleinberg: HITS
PSalsa: SALSA modificado
HubAvg (Hub-Averaging Kleinberg algorithm): híbrido entre SALSA y HITS
Athresh (Authority-Threshold algorithm): HITS modificado
Hthresh (Hub-Threshold algorithm): HITS modificado
Fthresh (Full-Threshold algorithm): HITS modificado
BFS (Breadth-First-Search algorithm): generalización de SALSA y con elementos de HITS
SBayesian (Simplified Bayesian algorithm): algoritmo Bayesiano simplificado
Bayesian: algoritmo Bayesiano

* http://www.altavista.com

** El motor de búsqueda DirectHit, como tal, dejó de existir a principios del año 2002. Este sistema de recuperación de información ponderaba la "popularidad" de un documento en el contexto de búsquedas individuales que el motor recibía. DirectHit monitoreaba la selección de ítems de la lista de resultados durante al menos un año y luego utilizaba la información para retroalimentar su algoritmo.

El principal problema provocado (en mayor o menor medida) por estos enfoques que se está tratando de superar es la tendencia temática*. Si la mayoría de las páginas que conforman el subgrafo inducido son de un tópico disimilar al tema de consulta, los principales nodos autoridad y *hub* pueden ser de una temática diferente. Una posible explicación es que en estos casos el tópico no está bien representado en la Web o no hay una comunidad (conjunto de páginas relacionadas por el mismo tema) fuertemente interconectada. Esto conlleva a decir que ningún motor de búsqueda puede funcionar solamente en base a la información de la conectividad, por el contrario, deben mejorar notablemente su búsqueda por palabras y/o conceptos**.

Algoritmos cuasi-dependientes de la consulta del usuario

En estos algoritmos la entrada del proceso de búsqueda no son los términos de una consulta, sino el URL de una página (o un sitio web) y la salida es un conjunto ordenado de páginas web relacionadas***. Una página relacionada es aquella que contiene el mismo tópico que la página original, pero no son necesariamente idénticas semánticamente.

El algoritmo más implementado es el de co-citación (también lo podemos encontrar como opción - what's related - en el navegador Mozilla de Mozilla Foundation y Navigator de Netscape). Dean y Henzinger[20] implementaron esta técnica con buenos resultados. El método usado se basó en encontrar páginas que vinculasen al URL modelo y después determinar las páginas enlazadas por él, obteniendo como respuesta una lista ordenada por autoridad. Rafiei y Mendelzon[21] construyeron un prototipo híbrido entre PageRank y SALSA, denominado TOPIC (TOronto PageInfluence Computation), que dado un URL inicial identifica los temas donde la página tiene una buena reputación y genera una lista ordenada por autoridad de estos tópicos (asociados a URLs).[8 9]

Algoritmos independientes de la consulta del usuario

Los algoritmos independientes de la consulta del usuario producen un ranking independientemente de la similitud consulta-documento puesto que su objetivo principal es medir la "calidad" intrínseca de una página. El mejor ejemplo de esta implementación es el algoritmo PageRank (PR) usado por el motor de búsqueda Google* como uno de sus componentes para ayudar a ordenar los resultados retornados por una búsqueda[4, 22]. Partiendo de un grafo (*off-line*) construido a priori, PR calcula la importancia de una página otorgando a cada hipervínculo que le apunta un peso proporcional a la autoridad de la página que lo contiene**. Para determinar la autoridad de la página citante, el PR es utilizado recursivamente unas 100 veces (hasta que los valores converjan). El PR de una página web no es influenciado por la página en sí misma o alguna consulta potencial, pero alcanza a una media objetiva universal basándose solamente en determinaciones subjetivas de importancia aportadas por los hiperenlaces***.

ANÁLISIS DE CONECTIVIDAD EN NUEVOS MOTORES DE BÚSQUEDA DE ACCESO PÚBLICO Y GRATUITO

Luego del éxito y popularidad alcanzado por Google y su algoritmo PageRank en el año 2000, dos nuevos motores de búsqueda, con gran énfasis en el análisis de enlaces al momento de buscar y ordenar por relevancia sus resultados, tuvieron su debut a mediados de 2001: WiseNut**** y Teom*****. Poco se conoce de su real funcionamiento, más allá de la poca información divulgativa brindada en su sitio y en boletines especializados, pero se tratará de ofrecer una idea del mismo******.

---

* Ver la investigación de Borodin *et al.*[18], para una discusión más extensa sobre éste y otros problemas.

** Otro argumento a favor de esta proposición surge del estudio realizado por Broder *et al.*[19] donde analizaron la macroestructura de hiperenlaces de la W3 y determinaron que fuera de un componente principal fuertemente conexo, la conectividad es relativamente baja entre los componentes que le integran.

*** En cierta manera, el URL de una página inicial puede tener una "carga semántica" un tanto más ambigua o precisa que las palabras de búsqueda, puesto que depende de su contenido y la "carga expresiva" de sus enlaces entrantes y salientes para determinar su tópico predominante.

* http://www.google.com

** Es importante señalar que los vínculos entrantes en una pagina web siguen una distribución exponencial[19]. En tal distribución, el valor de la mediana es mucho más bajo que el valor promedio. Esto significa que muchas páginas tienen pocos enlaces entrantes mientras pocas páginas tienen un número elevado. PR determinará que haya páginas dominantes y páginas poco influyentes.

*** PageRank no hace distinción entre nodos autoridades y nodos *hubs*.

**** http://www.wisenut.com

***** http://www.teoma.com

****** Estos sistemas, al igual que Google, utilizan para ordenar sus resultados según su relevancia en forma definitiva otra gran cantidad de criterios (frecuencia de las palabras, localización, etc.) que no se mgencionarán aquí porque exceden los propósitos del presente trabajo.

WiseNut posee un sistema de análisis de conectividad sensitivo al contexto, llamado WiseRank[23]. Este algoritmo se basa en el principio de que, dada una consulta, hay dos tipos de información diferente para determinar la relevancia de una página web: información intrínseca e información extrínseca. La información intrínseca proviene de la página en sí misma (título, cuerpo del documento y meta etiquetas) y la extrínseca proviene de las páginas que están conectadas a ella (información del anclaje y texto que lo rodea, meta etiquetas y título). WiseRank determina cuanto peso debe otorgar a la perspectiva objetiva (información extrínseca) y a la perspectiva subjetiva (información intrínseca) para ponderar los resultados. Según lo expresa Peter Adams, CTO de LookSmart, WiseNut está enfocado a valorar la calidad de cada vínculo y su contexto de conexión más que otros sistemas[24].

Teoma emplea el algoritmo SSP (Subject-Specific Popularity) que muestra una página importante en la parte superior del ranking, no sólo por su popularidad general, sino también en base al número de páginas del mismo tópico que le apuntan[25]. Primeramente, el sistema identifica y organiza las comunidades que versan sobre el tópico de la consulta, luego utiliza el SSP para analizar la relación existente entre las páginas de la comunidad y finalmente ordena cada resultado por autoridad basándose en el número de páginas temáticamente relacionadas que le referencian.

CONCLUSIÓN

Este artículo presentó una revisión sobre tres tipos de algoritmos de ordenación por relevancia basados en la conectividad de la World Wide Web.

La topología de red de la W3 provee patrones de enlace que permiten caracterizarla y ayudan a mejorar sustancialmente la recuperación de información de los sistemas de búsqueda. En este último aspecto, el análisis de enlaces cumple un rol fundamental en los algoritmos de ranking, especialmente en aquellos enfocados al reordenamiento de resultados. Los avances logrados en los últimos años dan muestra de ello. Sin embargo, la investigación sobre la estructura de hiperenlaces en su perspectiva objetiva y subjetiva recién está comenzando. Los futuros trabajos de investigación debieran considerar desarrollos de algoritmos multidimensionales a partir de la sinergia de diferentes métricas, ya que, cómo se determinó en diversos estudios, la conectividad por sí misma no es suficiente para recuperar los ítems más relevantes en un entorno tan anárquico como es la Web.

REFERENCIAS

1. PIROLLI, P.; PITKOW, J.; RAO, R. Silk from a sow's ear: extracting usable structures from the web. *In:* ACM SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING, 1996. ***Proceedings***... [S. l. : s. n.], 1996. Disponível em: <http://arbor.ee.ntu.edu.tw/~chyun/dmpaper/pirosf96.pdf >. Acceso em: 1 set. 2002.

2. CARRIERE, J; KAZMAN, R. Webquery: searching and visualizing the web through connectivity. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 6., 1997. ***Proceedings***... [S. l. : s. n.], 1997. Disponível em: <http://decweb.ethz.ch/WWW6/Technical/Paper096/Paper96.html >. Acceso em: 20 fev. 2002.

3. MARCHIORI, M. The quest for correct information on the web: hyper search engines. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 6.,1997. ***Proceedings***... [S. l. : s. n.], 1997. Disponível em: <http://decweb.ethz.ch/WWW6/Technical/Paper222/PAPER222.html >. Acceso em: 20 fev. 2002.

4. PAGE, L. *et al.*. *The PageRank citation ranking: bringing order to the web*, 1998. Disponível em: <http://dbpubs.stanford.edu:8090/pub/1999-66>. Acceso em: 20 abr. 2001.

5. KLEINBERG, J. M. Authoritative sources in hyperlinked environment. *In:* ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS, 9th, 1998. ***Proceedings***... [S. l. : s. n.], 1998. Disponível en: <http://simon.cs.cornell.edu/home/kleinber/auth.ps>. Acceso em: 31 maio 2001.

6. THELWALL, M. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, v. 8, n. 3, 2003. Disponível em: <http://informationr.net/ir/8-3/paper151.html>. Acceso em: 10 jun. 2003.

7. KIM, H. J. Motivations for hyper-linking in scholarly electronic articles: a qualitative study. *Journal of the American Society for Information Science*, v. 51, n. 10, p. 887-899, 2000.

8. BJÖRNEBORN, L. *Link patterns on the world wide web.* [S. l. : s. n.], 1999. Disponível en: <http://ix.db.dk/lb/linkpatterns/page.htm>. Acceso em: 26 nov. 2002.

9. HENZINGER, R. Hyperlink analysis for the web. *IEEE Internet Computing*, v. 5, n. 1, p. 45-50, Jan./Feb. 2001.

10. CHAKRABARTI, S. *et al.* Automatic resource list compilation by analyzing hyperlink structure and associated text. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 7., 1998. ***Proceedings***... [S. l. : s. n.], 1998. Disponível em:<http://www7.scu.edu.au/programme/fullpapers/1898/com1898.html>. Acceso em: 20 abr. 2002.

11. MCBRYAN, O. A. GENVL. *WWWW*: tools for taming the web. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 1., 1994. ***Proceedings***... [S. l. : s. n.], 1994. Disponível em: <http://www94.web.cern.ch/PapersWWW94/mcbryan.ps>. Acceso em: 8 nov. 1999.

12. CHAKRABARTI, S. *et al.* Experiments in topic distillation. *In:* ACM SIGIR WORKSHOP ON HYPERTEXT INFORMATION RETRIEVAL ON THE WEB, 1998. ***Proceedings***... [S. l. : s. n.], 1998. Disponible em: <http://www.almaden.ibm.com/cs/k53/abstract.html>. Acceso em: 26 nov. 2002.

13. The CLEVER project. Disponible em: <http://www.almaden.ibm.com/cs/k53/clever.html>. Acceso em: 26 nov. 2002.

14. LEMPEL, R.; MORAN S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 9., 2000. ***Proceedings***... [S. l. : s. n.], 2000. Disponível em: <http://www9.org/w9cdrom/175/175.html>. Acceso em: 29 abr. 2002.

15. CHEN, Z. *et al.* A unified framework for web link analysis. *In:* INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS ENGINEERING, 3., 2002. ***Proceedings***... [S. l. : s. n.], 2002. p. 63- 70, 2002.

16. SCHAPIRA, A. *Collaboratively searching the web: an initial study*. [S. l. : s. n.], 1999. Disponível em: <http://www-ccs.cs.umass.edu/~schapira/thesis/report/>. Acceso em: 21 jul. 2002.

17. JOACHIMS, T. Evaluating retrieval performance using clickthrough data. *In:* SIGIR WORKSHOP ON MATHEMATICAL/FORMAL METHODS IN INFORMATION RETRIEVAL, 2002. ***Proceedings***... [S. l. : s. n.], 2002. Disponível em: <http://www.cs.cornell.edu/People/tj/publications/joachims_02b.ps.gz>. Acceso em: 21 jul. 2002.

18. BORODIN, A *et al*. Finding authorities and hubs from link structures on the world wide web. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 10., 2001. ***Proceedings***... [S. l. : s. n.], 2001. Disponível em: <ftp://markov.utstat.toronto.edu/jeff/www10-314.ps.Z>. Acceso em: 29 abr. 2002.

19. BRODER, A. *et al.* Graph structure in the web. *Computers Networks*, v. 33, n. 1-6, p. 309–320, 2000.

20. DEAN, J.; HENZINGER, M. R. Finding related pages on the web. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 8., 1999. ***Proceedings***... [S. l. : s. n.], 1999. Disponível em: <http://decweb.ethz.ch/WWW8/data/2148/html/index.htm>. Acceso em: 20 abr. 2001.

21. RAFIEI, D.; MENDELZON, A. O. What is this page known for?: computing web page reputations. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 9., 2000. ***Proceedings***... [S. l. : s. n.], 2000. Disponível em: <http://www9.org/w9cdrom/368/368.html>. Acceso enm 29 abr. 2002.

22. BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *In:* INTERNATIONAL WORLD WIDE WEB CONFERENCE, 7., 1998. ***Proceedings***... [S. l. : s. n.], 1998. Disponível em: <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>. Acceso em: 25 jun. 2001.

23. WISENUT SE. WiseNut search engine. *White Paper*. [S. l. : s. n.], 2001. Disponível em: <http://www.wisenut.com/pdf/WISEnutWhitePaper.pdf>. Acceso em: 11 dez. 2001.

24. SHERMAN, C. *LookSmart revives WiseNut*. [S. l. : s. n.], 2002. Disponível em: <http://siliconvalley.internet.com/news/article.php/1472411>. Acceso em: 20 jun. 2003.

25. TEOMA. *Search with authority: the Teoma difference*. Disponível em: <http://sp.teoma.com/docs/teoma/about/searchwithauthority.html>. Acceso em: 24 jul. 2003.