Búsqueda de Sitios Web con Autoridad en un Tema

Fernando R.A. Bordignon, Pablo J. Lavallén y Gabriel H. Tolosa
{bordi, plavallen, tolosoft}@unlu.edu.ar

Universidad Nacional de Luján
Laboratorio de Redes de Datos

Resumen

La búsqueda de recursos – páginas o sitios web – que son referentes (o autoridades) en un tema particular es una tarea básica que ayuda a construir o mejorar distintos servicios de información. No obstante, es posible plantear el concepto de autoridad desde un nivel de abstracción mayor teniendo en cuenta el contenido de las páginas, para evaluar desde otro punto de vista si éstas son relevantes en un tópico particular.

En este trabajo se propone un método simple que utiliza las capacidades de los motores de consulta existentes a los efectos de obtener – de manera automática – listas de sitios web que son autoridades temáticas.

Los resultados muestran que el método es eficiente, obteniendo una precisión entre 0.66 y 1.00 para los diferentes experimentos. Si bien aún se encuentra en una primera etapa, se propone su utilización como un filtro a incorporar a los motores de consultas, donde existan listas predefinidas de sitios a incluir o excluir de una consulta.

Palabras clave: recuperación de información, espacio web, motores de consulta, búsqueda de autoridades, búsquedas de precisión.

1 – Introducción

La búsqueda de recursos – páginas o sitios web – que son referentes (o autoridades) en un tema particular es una tarea básica que ayuda a construir o mejorar distintos servicios de información. El concepto de autoridad en la web fue propuesto por Kleinberg [8] como una parte del algoritmo de rankeo de páginas web denominado HITS. Este algoritmo permite establecer la importancia de una página en función de los enlaces que posee y los que recibe.

No obstante, es posible plantear el concepto de autoridad desde un nivel de abstracción mayor teniendo en cuenta su contenido, de manera de evaluar desde otro punto de vista si éstas son relevantes en un tópico particular.

Por otro lado, la estructura de enlaces de la web es explotada por los motores de búsqueda más importantes como Yahoo![1] y Google[2], los cuales están preparados para ser eficientes manejando altos volúmenes de datos y poseen una amplia cobertura del espacio web indexable [14] [12].

En este trabajo se plantea la siguiente cuestión: ¿Es posible utilizar las capacidades de los motores de consulta existentes a los efectos de obtener – de manera automática – listas de sitios web que son autoridades temáticas? Esta hipótesis se basa en la idea de incorporar valor agregado al proceso de armado de listas de respuestas de los buscadores pudiendo filtrar aquellos sitios respecto de su autoridad temática, aunque su contenido sea relevante a una determinada consulta. Suponga las siguientes necesidades de información:

a) "información comercial sobre venta de un celular marca Motorola", la cual puede ser mapeada en una consulta de usuario utilizando los siguientes términos: *celular motorola venta*.

b) "información sobre como conectar un celular Motorola con una PC". En este caso la consulta puede ser: *celular motorola "conexión a PC"*.

Luego de realizadas las consultas se obtendrán listas de sitios relacionados donde – en ambos casos – se encontrarán referencias a algunos no relevantes al tema central, las cuales introducen ruido en la lista de respuestas. Si se tiene en cuenta que los usuarios – en general – revisan unas pocas páginas de respuestas, la calidad de la salida se verá afectada sensiblemente.

Se propone – entonces – una metodología simple destinada a obtener listas de autoridades sobre un tema a partir de consultas a un motor de búsquedas. El método es económico en términos de recursos informáticos involucrados y –además – requiere poco tiempo de procesamiento.

Alternativamente, este método puede ser utilizado para la asistencia en la construcción de directorios temáticos que ayudan a soportar la navegación mediante browsing como Open Directory[3] (Dmoz) a nivel mundial o Todoar[4] para

[1] http://www.yahoo.com
[2] http://www.google.com
[3] http://www.dmoz.org/

el dominio argentino. McCallum et al. [9] plantean que obtener de forma fácil listas de sitios altamente relacionados con un tema ayuda a alimentar motores de consulta especializados en una temática particular.

2 – Trabajos Relacionados

Existen diferentes enfoques para identificar páginas de alta calidad relacionadas con una consulta. En [1] se aplica el estudio de contenidos como un agregado al análisis de conectividad para obtener listas de autoridades sobre un dominio específico del conocimiento que satisfagan una consulta de usuario. Esta información, es un atributo más para calificar a los documentos, donde si el sitio de donde provienen es autoridad en el tema dado podría ayudar a mejorar su ponderación.

El concepto de autoridad – aplicado a documentos – es la base del algoritmo Hilltop [2], el cual tiene por finalidad asignar un puntaje de importancia a páginas web. En su trabajo, Bharat utiliza el concepto de documentos con autoridad sobre un tópico particular a los efectos de ponderar en mayor grado a una página si posee enlaces que provengan de esta clase de recursos. Hilltop solo considera como páginas referentes a aquellos documentos que sean expertos – es decir – que fueron creados con el propósito de auxiliar a los usuarios en la búsqueda de un tema particular y en primera instancia arma una lista de ellos sobre cada tema.

Kimbrough et al. [7] proponen un enfoque alternativo para el ranking de autoridades basado en el uso de la web. En su trabajo, definen que la autoridad de un sitio está dada por los usuarios que lo visitan, por lo que sus métricas son derivadas de un enfoque centrado en el usuario (*user-centric*) en vez de la información de navegación de un sitio (*site-centric*), almacenada en sus archivos de *logs*.

Por el lado de los motores de búsqueda, los desarrolladores de Teoma[5] proponen una nueva tecnología que denominan Subject-Specific Popularity [13], la cual permite incorporar otro nivel de análisis de la autoridad de los resultados de búsqueda. De acuerdo a ésta, Teoma mantiene un mapa de la web organizado por comunidades temáticas. Para refinar el proceso de ranking de resultados y la autoridad de un sitio, se analizan la cantidad de enlaces desde páginas de su mismo tema que apuntan a éstos. Si bien no se presentan estudios de eficiencia, la idea detrás de este modelo reconoce la necesidad de contar con sitios que posean autoridad sobre un tema como complemento del análisis de conectividad.

[4] http://ww.todoar.com.ar
[5] http://www.teoma.com

3 – El Grafo de la Web y los Algoritmos de Ranking

La web puede ser modelada como un grafo dirigido donde los nodos corresponden a las páginas HTML y los enlaces o hipervínculos entre éstas son las aristas [4]. Formalmente, este grafo consiste en un conjunto de nodos, denotado como P y un conjunto de aristas, A. Cada arista (denotada como q→p) es un par ordenado (q, p) donde q, p ∈ P y representan un enlace o vínculo entre las páginas (nodos) q y p, situación que se da solo con algunos pares. En este caso, q es un enlace entrante de p y éste uno saliente de q.

A partir de este modelo, se cuenta con un grafo con millones de nodos y miles de millones de aristas cuya estructura y complejidad requiere de métodos de análisis que permitan determinar sus propiedades. Esta información puede ser utilizada en diversas aplicaciones como búsquedas, ranking, recuperación y minería en la web. Un ejemplo concreto lo presentan los algoritmos de rankeo de páginas web utilizados por los buscadores, como HITS [8] y PageRank [10].

En el algoritmo HITS se proponen los conceptos de centros y autoridades (*Hubs* y *Authorities*) con la finalidad de clasificar la importancia de una página web. Para un conjunto P de páginas web, con su correspondiente conjunto A de aristas, las cuales determinan el grafo, HITS rankea cada página p ∈ P a partir de su calidad como Authority x_p y Hub y_p. Los valores de se calculan de la siguiente manera:

$$\text{Authority-Score}_p = x_p = \sum_{q|q\to p} y_q$$

$$\text{Hub-Score}_p = y_p = \sum_{q|q\to p} x_q$$

Los valores iniciales de x_p e y_p se inicializan en 1 y el algoritmo itera varias veces hasta estabilizar los valores. Una página *Hub* posee una importante cantidad de enlaces a páginas con contenidos relevantes, como por ejemplo las del directorio de Yahoo!. En cambio, una página *Authority* es aquella que recibe una importante cantidad de enlaces y pocos salen de ella. Esta característica presupone que tales páginas son referentes en un tema específico. Originalmente, el algoritmo HITS se aplica al subconjunto de páginas obtenidas como respuesta a una consulta más un subconjunto anexado desde los enlaces de éstas para presentar la lista de respuestas

final al usuario. Sobre esta base, se han introducido algunas variantes a HITS como las presentadas en [1] y [6]. En [5] se aplica HITS a sitios completos, de manera estática.

Por otro lado, el algoritmo PageRank asigna un valor de importancia a todas las páginas del grafo web, independientemente de alguna consulta realizada. Sus autores lo proponen como un modelo del comportamiento de los usuarios, donde éstos pueden seleccionar al azar un enlace dentro de una página (navegante aleatorio), cuya probabilidad depende de la cantidad de enlaces salientes de ésta. Entonces, la probabilidad de que un determinado navegante aleatorio alcance una página es la suma de las probabilidades de que siga los vínculos hacia ésta. En su versión original, el cálculo del valor de PageRank para una página p ∈ P resulta:

$$PR_p = (1-d) + d \times \sum_{pi \to p} \frac{PR(p_i)}{CT(p_i)}$$

Donde

PRp es el PageRank de la página p.

PR(pi) es el PageRank de las páginas pi que poseen enlace a la página p.

CT(pi) es la cantidad de enlaces salientes de pi.

d es un factor de damping que representa la probabilidad de que el navegante aleatorio no se detenga (0-1).

Para obtener el PageRank final de una página el algoritmo también itera una serie de veces hasta estabilizar el valor.

3 – Búsqueda de Autoridades

El método propuesto consiste en el envío de una serie de consultas a un buscador, las cuales deben ser cuidadosamente seleccionadas. Luego, se analizan las listas de respuestas obtenidas y se establece la cantidad de ocurrencias de cada sitio web. Con tal información, se construye un ranking, donde aquellos sitios que se encuentren en los primeros lugares tienen mayor autoridad temática respecto de las búsquedas.

El punto más importante es la definición de las consultas a enviar. Éstas deben contener términos específicos del dominio temático buscado, como así también frases representativas y conjunciones de términos relevantes. Debe evitarse la utilización de términos que planteen ambigüedad y – por ende – generen resultados no deseados. La construcción de esta lista – en principio – debería ser realizada por un experto en el tema, aunque se tiene en cuenta, a futuro, el desarrollo de métodos automáticos basados en análisis de documentos relevantes del dominio.

Para determinar la lista de resultados final se procede de la siguiente manera: Dado un conjunto C de consultas realizadas por un experto, existirán R listas de respuestas, donde $C_i \rightarrow R_i$. Luego, el ranking de cada sitio web S_i se obtiene calculando:

$$S_i = \sum_{\forall r \in R_i} q_{ji} \quad (1)$$

Donde

Si es el score final del sitio i

Ri es la lista de respuestas de la consulta i

qji es la cantidad de veces que aparece el sitio i en la lista de respuestas j.

El resto de las operaciones a realizar son: envío de las consultas, *parsing* de los resultados, cálculo de los scores de cada sitio y fusión de las respuestas para armar la lista final. Todas éstas son operaciones simples que se realizan de forma totalmente automática.

4 – Experimentos y Resultados

A los efectos de evaluar el funcionamiento de la metodología propuesta se realizaron una serie de experimentos para obtener sitios con autoridad temática en diversos dominios. En todos los casos, se utilizó Google como motor de búsquedas, solicitándole listas de 100 respuestas por consulta y restringiéndolas a sitios del dominio de Argentina (.ar). Esta última elección se realizó para facilitar la tarea de evaluación.

Como métrica de performance de la recuperación se utilizó la medida de Precisión [11], la cual establece la proporción de respuestas relevantes respecto de las recuperadas:

Precisión = | Relevantes_Recuperados | / | Recuperados |

En este caso, el concepto de relevante se aplica a sitios considerados “Autoridad Temática” y el juicio fue determinado por expertos humanos.

En el primer experimento se intentó detectar sitios que poseen amplitud temática y – por ende – aparecen como respuesta para consultas de distintos tópicos. En general, corresponden a sitios de noticias, bibliotecas, enciclopedias, etc., los cuales no contienen información específica de un tema particular. A éstos los denominamos Sitios de Amplitud Temática (SAT).

Para la identificación de los SAT se utilizó un juego de consultas de un término, el cual se extrajo al azar de una lista de palabras en

español. En total, se ejecutaron 32.000 consultas y se obtuvieron 31.834 listas de resultados no vacías. De la fusión de las mismas se confeccionó el ranking final ordenando – descendentemente – por frecuencia de aparición en las mismas, de acuerdo a (1).

En la tabla 1 se muestran los primeros 15 sitios SAT detectados para el dominio argentino. La tercera columna corresponde al porcentaje de las listas de respuesta en las que participó el sitio en cuestión. Estos sitios no son considerados como Autoridad Temática, aunque su detección resulta importante y se utilizó en experimentos posteriores.

Orden	Sitio	% aparición
1	www.lanacion.com.ar	**54.4**
2	www.pagina12.com.ar	**51.3**
3	Axxon.com.ar	**41.3**
4	Articulo.mercadolibre.com.ar	**39.9**
5	www.lacapital.com.ar	**36.4**
6	www.inta.gov.ar	**29.2**
7	www.paginadigital.com.ar	**29.1**
8	rionegro.com.ar	**27.8**
9	www.buenosaires.gov.ar	**26.4**
10	www.losandes.com.ar	**26.0**
11	www.todoar.com.ar	**25.4**
12	www.fac.org.ar	**25.0**
13	www.laopinion-rafaela.com.ar	**24.6**
14	www.po.org.ar	**24.0**
15	weblog.educ.ar	**22.7**

Tabla 1 – Primeros 15 SATs el dominio .ar

La segunda prueba consistió en detectar sitios dedicados a la promoción y venta de múltiples productos. Entre éstos se consideran los sitios de remates, catálogos de grandes almacenes y los que construyen su negocio publicitario sobre la base de enlaces hacia los primeros. Estos últimos representan un alto porcentaje del conjunto objetivo. Este conjunto de sitios se los considera autoridades en su función específica, la cual puede ser definida como:

"sitios que proporcionan información comercial sobre productos, proveedores y precios, en un amplio espectro"

Este conjunto de sitios los denominamos "Sitios Comerciales Multiventa" (SCM). Se probaron varias formas de construcción de consultas para determinar cuáles son SCM. Finalmente, se determinó que las consultas más adecuadas son aquellas que poseen una combinación de términos que definen marcas de productos junto con palabras del ámbito comercial (compra, venta, cuota, permuta, oferta, etc.). En este experimento se enviaron 500 consultas como las descriptas a Google. En la tabla 2 se muestra un subconjunto de las consultas enviadas.

Renault compra hasbro
Torino disney kyocera
3com renault intel
Alcatel nintendo vendido
Compra peugeot pioneer

Tabla 2 – Subconjunto de las consultas enviadas.

De la fusión de las 500 listas de respuestas (cada una con 100 elementos) se obtuvo el ranking final. En la tabla 3 se presentan los primeros 15 SCM del dominio argentino.

Orden	Sitio
1	www.mercadolibre.com.ar
2	www.buscaturismo.com.ar
3	clasificados.grippo.com.ar
4	www.buscandoenlaweb.com.ar
5	www.ringtones.com.ar
6	foros.3dgames.com.ar
7	alive-net.com.ar
8	www.emoweb.com.ar
9	www.visitaelmundo.com.ar
10	www.bowen.com.ar
11	www.deremate.com.ar
12	www.grandes-ofertas.com.ar
13	www.dotcom.com.ar
14	www.virtualbanking.com.ar
15	www.yupimail.com.ar

Tabla 3 – Primeros 15 sitios de venta del dominio .ar

Del análisis de relevancia de cada uno resultó que en los primeros 50 sitios se obtuvo una precisión de 0.92, un valor 0.90 para los 100 primeros y para los 200 ésta decayó a 0.71.

Se realizó un nuevo experimento a los efectos de identificar sitios referentes en un tópico específico. A estos sitios los denominamos Sitios Autoridad (SA). Para la prueba se tomó el tema "educación" y se definió la lista de consultas utilizando términos del dominio, por ejemplo: universidad, profesorado, "autoridades escolares", "problemática escolar", etc.

Se ejecutaron 100 consultas y se fusionaron las respuestas siguiendo la misma metodología propuesta para obtener los SA_{edu}. En los primeros 50 sitios se obtuvo un 0.76 de precisión, mientras que para los primeros 100 el valor fue 0.69. Aquí, como elemento de juicio se consideraron solamente sitios que exclusivamente están dedicados a la educación.

Luego, se realizó un experimento similar pero tomando como tópico el tema "agricultura". Se construyó una lista con términos del dominio y se

ejecutaron 100 consultas para intentar detectar los SA_{agro}. La performance en este experimento fue de 0.58 de Precisión en los 100 primeros documentos, la cual subió a 0.71 en los 200 documentos. Los resultados obtenidos se presentan en la tabla 4.

	SCM	SA_{edu}	SA_{agro}
P@5	1.00	0.80	0.60
P@10	1.00	0.90	0.80
P@20	1.00	0.75	0.60
P@50	0.92	0.76	0.46
P@100	0.90	0.70	0.58

Tabla 4 – Resumen de resultados

En los resultados anteriores se nota máxima eficiencia en la detección de SCM, mientras que en los demás – aunque el valor es interesante – no se llega a alcanzar. Una posible explicación puede estar en el armado de la lista de consultas, suponiendo que la estrategia utilizada para la primera permite discriminar de mejor manera. Sin embargo, de una inspección manual de las listas de respuestas de SA_{edu} y SA_{agro} encontramos vario sitios que están en la lista de SAT. Teniendo – entonces – reconocidos los SAT en el primer experimento los quitamos de la lista final de SA_{edu} y SA_{agro}, aumentando la calidad de la respuesta. En la tabla 5 se presentan los resultados luego de realizada esta optimización hasta los 50 primeros documentos.

	SA_{edu} – SAT	SA_{agro} – SAT
P@5	1.00	1.00
P@10	0.90	1.00
P@20	0.90	0.95
P@50	0.78	0.88

Tabla 5 – Resultados luego de optimizar

5 – Discusión y Trabajos Futuros

En este artículo se presenta un método simple para obtener un ranking de sitios que son "Autoridades Temáticas", utilizando la infraestructura de los buscadores existentes. El método es eficiente ya que requiere de consultas simples y fusión de los resultados. No obstante, el conjunto de las consultas a enviar debe ser cuidadosamente seleccionado por un experto humano.

Los resultados iniciales indican que el método resulta eficaz y eficiente. En los mismos, se muestra como obtener determinados sitios, por ejemplo, de venta masiva o de información general, los cuales pueden ser considerados como ruido en una consulta por algún tema. Al poder identificarlos, se los puede eliminar de las respuestas de estas últimas, obteniendo una precisión entre 0.78 y 1.00.

El método puede ser utilizado – además – como un filtro a incorporar a los motores de consultas, donde existan listas predefinidas de sitios a incluir o excluir de una consulta. Inclusive, estas listas podrían ser personalizables por usuario.

Finalmente, se considera la posibilidad de definir una metodología de armado automático de las listas de consultas a partir del muestreo de términos de documentos ejemplo del tema.

6 – Bibliografía

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 104-111, 1998.

[2] K. Bharat and G. Mihaila. Hilltop: A search engine based on expert documents. In Poster Proceedings of WWW9, pp. 72-73, 2000.

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30 No. 1–7, pp. 107–117, 1998.

[4] A.Broder, R.Kumar, F. Maghoul, P.Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web. In Proceedings of the WWW9 Conference pp. 309-320, 2000.

[5] C. Castillo and R. Baeza-Yates. WIRE: an Open Source Web Information Retrieval Environment. Workshop on Open Source Web Information Retrieval (OSWIR), 2005.

[6] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan and S. Rajagopalan. Automatic Resource Complation by Analyzing Hyperlink Structure and Associated Text. Proceedings of the WWW7, 30 (1-7), pp. 107-117, 1998.

[7] S. Kimbrough, S. Kimbrough, B. Padmanabhan, Z. Zheng. "On Usage Metrics for Determining Authoritative Sites". Information Technologies and Systems (WITS), 2000.

[8] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Association for Computing Machinery - Journal of the ACM (46:5), pp. 604-632, 1999.

[9] A. McCallum, K. Nigam, J. Rennie and K. Seymore. Building domain-specific search engines with machine learning techniques. In AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999.

[10] L. Page, S. Brin, R. Montwani and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford Digital Library, 1998

[11] G. Salton. The Smart System – Experiments in Automatic Document Processing. Prentice Hall Inc., 1971.

[12] D. Sullivan. Search Engine Sizes. http://searchenginewatch.com/reports/article.php/2156481

[13] Teoma. Adding a New Dimension to Search: The Teoma Difference is Authority. http://sp.teoma.com/docs/teoma/about/searchwithauthority.html

[14] L. Vaughan and M.Thelwall. Search Engine Coverage Bias: Evidence and Possible Causes. Information Processing & Management, Vol. 40, No. 4. pp. 693-707, 2004.

[15] J. Wu and K. Aberer. Using SiteRank for Decentralized Computation of Web Document Ranking. Proceedings of the 3th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 2004.