Análisis de Enlaces en el Espacio Web de las Universidades Argentinas

Gabriel H. Tolosa y Fernando R.A. Bordignon
{Tolosoft, bordi}@unlu.edu.ar

Universidad Nacional de Luján
Laboratorio de Redes

Resumen

En este artículo se presentan los primeros resultados de un trabajo de caracterización y análisis de muestras del espacio web educativo de Argentina.

Particularmente, éste corresponde a las universidades nacionales, en el cual se realizó un estudio a tres niveles, encontrando parámetros similares a otros muestreos del espacio web.

Consideramos que estas tareas, junto con la información de su uso permitirán proponer optimizaciones sobre gestión de los recursos digitales existentes y planificar su desarrollo.

Palabras clave: análisis de enlaces, web educativa, grafo web, universidades argentinas, minería web

1 – Introducción

El estudio de las características de porciones específicas del espacio web permite analizar el comportamiento de un conjunto de entidades respecto de la web global. Diversos esfuerzos se han realizado tomando diferentes muestras de tamaño variado, especialmente relacionadas con dominios nacionales [2] [3] [4] [12] [16], los cuales tienen un buen balance entre diversidad y completitud [5].

La riqueza de la web no está dada solamente por el contenido de las páginas sino también por los hipervínculos que las conectan. Esta estructura de enlaces es armada – en general – por humanos y representa una fuente de información indirecta (respecto del contenido) que es ser de alto valor [10]. En algunos casos, cuando el autor de una página web genera un *link* hacia otra, está dando una recomendación implícita acerca de la calidad del contenido de esta última. Esta información es utilizada en diversas aplicaciones como búsquedas, ranking, recuperación y minería en la web. Un ejemplo clásico son los algoritmos de rankeo de páginas web como HITS [13] y PageRank [17] utilizados por algunos buscadores.

Complementariamente, los enlaces de un subconjunto específico de sitios web relacionados brinda una visión de la estructura social subyacente, permitiendo comprender algunas cuestiones relacionadas con generación y utilización de contenidos, la importancia de algunos de tales sitios y la calidad de ciertas páginas, entre otras [1] [19].

En este trabajo se toma como objeto de estudio el conjunto de las Universidades Argentinas y se realizan estudios a nivel de enlaces. El objetivo principal es caracterizar la distribución de las organizaciones, estado de conectividad, puntajes de los algoritmos de ranking y otros indicadores que permitan definir el estado (*status*) – a nivel de enlaces – de cada una en la red. Se estudió la red como grafo de páginas (*WebGraph*), sitios (*HostGraph*) y – además – se estableció un análisis a nivel de dominios (*DomainGraph*).

2 – El Grafo Web y las Redes Libre de Escala

La web puede ser modelada como un grafo dirigido donde los nodos corresponden a páginas HTML y los enlaces entre éstas son las aristas [8]. Generalmente, se lo denomina *WebGraph*. Formalmente, este grafo consiste en un conjunto de nodos, denotado como P y un conjunto de aristas, A. Cada arista (denotada como $q \rightarrow p$) es un par ordenado (q, p) donde $q, p \in P$ 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.

Kleinberg [14] y Barabasi [6] plantearon que la topología del grafo de la web corresponde a una red libre de escala, en la cual la distribución de los enlaces sigue una ley de potencias de la forma: $P(x = k) \approx k^{-\beta}$, para $\beta > 0$. Esta situación fue luego observada por Broder en un muestreo de la web de gran escala [8], encontrando como propiedad básica del grafo web que la distribución del grado entrante de los vértices sigue una ley de

potencias con exponente β = 2.1. Por otro lado, la distribución del grado saliente sigue una ley de potencias imperfecta con β = 2.72.

En el mismo trabajo, Broder et all proponen una estructura macroscópica para la web a partir del análisis de los enlaces. En éste, se muestra que el grafo de la web está formado por un componente gigante de tres partes: a) CORE, que incluye el SCC (*Strongly Connected Component*) mayor, b) IN, formado por nodos que pueden alcanzar al CORE pero no son alcanzables desde éste y c) OUT, que es un conjunto de nodos alcanzables desde el CORE que no poseen enlaces salientes hacia éste. Complementariamente, se identificaron otros componentes que no pertenecen al gigante y se encuentran desconectados (DISCONECTED), como así también nodos que son alcanzables solo desde porciones de IN o de OUT llamados TENDRILS. Finalmente, los nodos desde IN que alcanzan OUT forman el componente TUBES. A esta estructura se la conoce como "*bowtie*".

3 – El Espacio Web de las Universidades Agentinas

Como se mencionó anteriormente, la colección estudiada (Uni.AR) corresponde a páginas web de universidades Argentinas, tanto públicas como privadas, cuyos dominios fueron obtenidos del sitio oficial del Ministerio de Educación, Ciencia y Tecnología de Argentina[1].

Para la obtención de los datos se tomo una parte de la colección Edu.AR, la cual es una muestra del dominio educativo Argentino obtenida con el *crawler* WIRE [9] durante diciembre de 2005 como parte de otro trabajo de investigación de nuestro grupo.

Para el análisis se construyeron grafos en tres niveles de abstracción. En el primero – el WebGraph – las páginas representan nodos y sus enlaces internos corresponden a las aristas del grafo. En este, se obtuvieron 323.405 nodos con 1.268.933 aristas. En el HostGraph [11], cada nodo corresponde a un sitio web completo y los *links* entre éstos existen si existe al menos una páginas de uno tiene un enlace que apunta hacia una páginas del otro. Se encontraron 1.692 nodos con 4.108 aristas. Finalmente, se generó un nivel más de análisis a nivel de dominios (DomainGraph) que – en este caso – corresponden al tercer nivel y representa a una Universidad completa (nodos). Los enlaces se obtienen de manera similar al HostGraph. Este grafo posee 80 nodos y 751 enlaces.

[1] http://www.me.gov.ar

4 – Experimentos y Resultados

4.1 – *WebGraph*

Se tomó el grafo web y se estudiaron dos propiedades: distribución del grado y distribución de los valores del algoritmo de ranking PageRank. El centro de las distribuciones de grado entrante (in-degree) y saliente (out-degree) son consistentes con una ley de potencias de exponente β = 2.17 ± 0.05 y β = 2.48 ± 0.05, respectivamente. Estos valores son comparables con los encontrados por Broder [8] para su muestra de la web. Las distribuciones del grado se presentan en el gráfico 1.

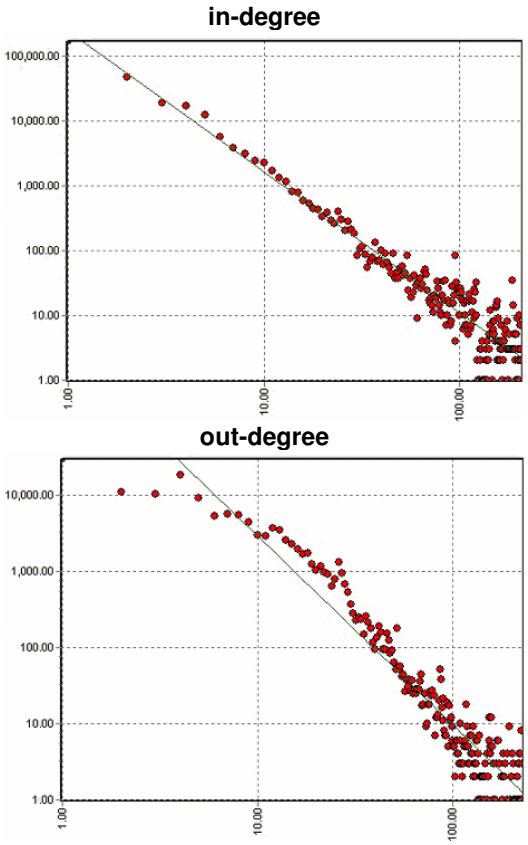

Gráfico 1 – Distribuciones de grado entrante y saliente para el *WebGraph*

El promedio de *links* por página (tanto entrantes como salientes) observado es 4. Aquí hay que recordar que solo se estudiaron los enlaces entre las universidades Argentinas (y no entre otras organizaciones) por lo que el valor no es bajo como aparenta. Complementariamente, se estudió el grado saliente de enlaces hacia páginas que no pertenecen a Universidades. La distribución

(Gráfico 2) también sigue una ley de potencias con parámetro $\beta = 2.38 \pm 0.17$ y la media es 1. Es interesante notar que cada 5 enlaces, 4 son dentro del dominio en cuestión (aquí no se distingue si es dentro del mismo sitio o dominio) y 1 a organizaciones fuera del dominio.

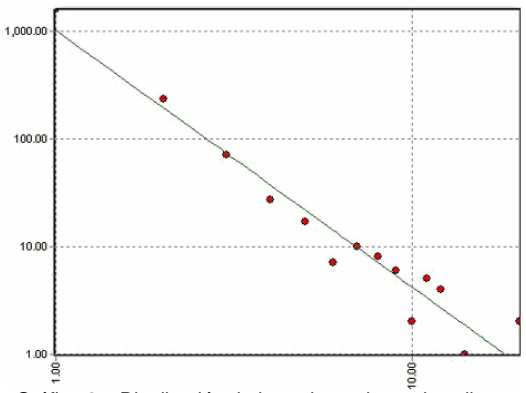

Gráfico 2 – Distribución de los valores de grado saliente fuera del dominio estudiado

A continuación, se estudió la distribución de PageRank, dado que es uno de los algoritmos de ranking basado en enlaces más populares. En ésta se encontró una ley de potencias con parámetro $\beta = 2.09$ (+/- 0.04), el cual es comparable con lo reportado en [5] donde se observó que para dominios nacionales el parámetro de la distribución 1.86 +/- 0.06. La distribución de los valores de PageRank se muestra en el gráfico 3. De acuerdo a Panduragán [18] este exponente debe ser similar al de la distribución del grado entrante, situación observada en esta colección.

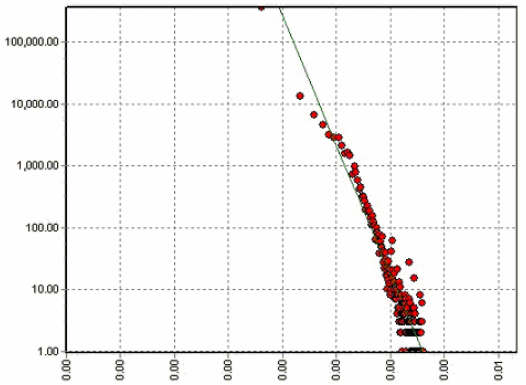

Gráfico 3 – Distribución de los valores de PageRank para el WebGraph

Finalmente, se presenta el análisis de la estructura macroscópica. Para el procesamiento del grafo se utilizó la librería COSIN [15], la cual se diseñó especialmente para manejar grafos web masivos. El gráfico de *bowtie* se presenta en la Figura 2, mientras que en la Tabla 1 se muestran los tamaños de cada región.

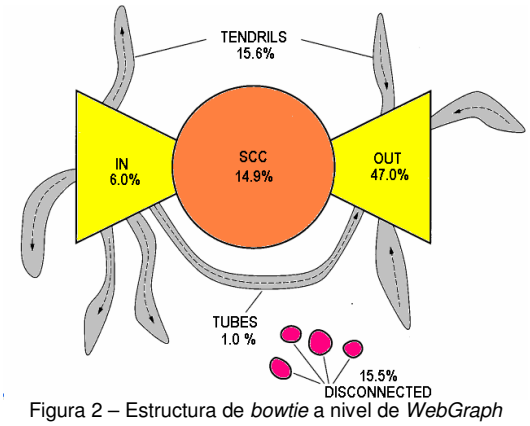

Figura 2 – Estructura de *bowtie* a nivel de *WebGraph*

SCC	IN	OUT	DISC	TENDRILS	TUBES
48.018	19.501	151.964	50.144	50.354	3.424

Tabla 1 – Tamaños de las regiones del *bowtie*

Aquí se puede apreciar que la distribución de los tamaños no se corresponde con la encontrada por Broker. Lo más notorio es el porcentaje de páginas en OUT, lo cual se puede atribuir a una baja tasa de actualización. Es común en el ambiente académico encontrar páginas de determinados cursos o asignaturas pasadas que no son actualizadas y permanecen para consultas o referencias.

4.2 – *Hostgraph*

En el segundo experimento se tomó como unidad de análisis el *HostGraph*. De manera similar al anterior, se estudió la distribución del grado. Las mismas responden a leyes de potencias con parámetro $\beta = 1.28 \pm 0.08$, para el grado entrante y $\beta = 1.07 \pm 0.07$ para el grado saliente. El promedio de *links* (entrantes y salientes) observado es 12. Las distribuciones se presentan en el gráfico 4.

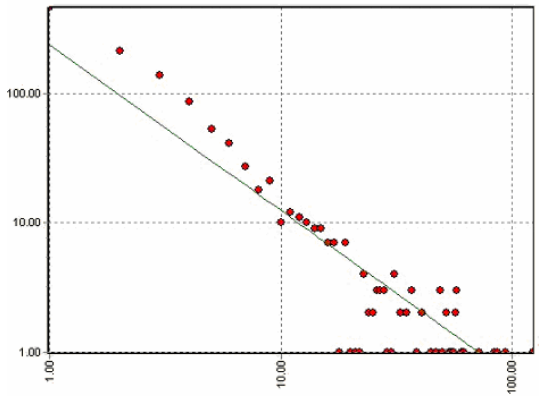

Gráfico 4_a – Distribución de grado entrante

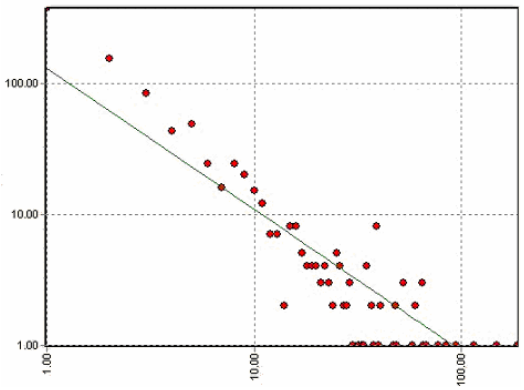

Gráfico 4 $_b$ – Distribución de grado saliente

En la figura 3 se muestra la estructura de bowtie del HostGraph. Se pueden apreciar ciertas diferencias interesantes, como el aumento de la proporción de elementos en el componente DISCONNECTED. De igual manera, crece la región de SCC y disminuye la de OUT.

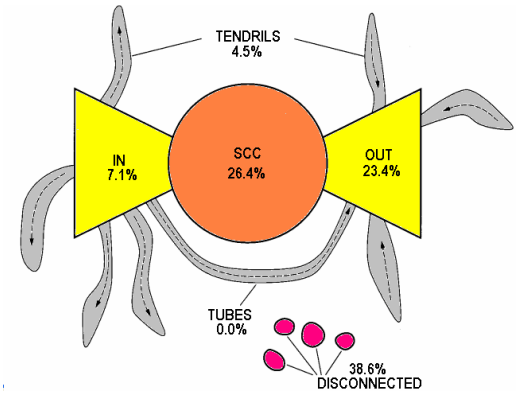

Figura 3 – Estructura de *bowtie* a nivel de *HostGraph*

4.3 – *DomainGraph*

Finalmente, se estudió el espacio a nivel de dominios. Se midió la cantidad de páginas y cantidad de enlaces por dominio. En ambos casos, el ajuste de las distribuciones sigue una ley exponencial de la forma $y \approx a\, e^{\beta}$. Para la cantidad de páginas el valor del exponente observado resulta $\beta = -0.11$, mientras que para la distribución de enlaces es $\beta = -0.13$. En el gráfico 5 se presentan las distribuciones con su curva de ajuste.

Sobre este grafo se realizaron mediciones de parámetros clásicos del análisis de redes sociales [7] [20]. Estos permiten caracterizar aspectos estructurales tanto de la red y como de los nodos participantes (en este caso los actores son las universidades).

En el primer caso, se obtuvo la densidad de la red, que es una métrica calculada entre el número de enlaces existentes respecto del total de enlaces posibles (normalizada en [0-1]). En este caso, la densidad es 0.57, lo cual indica que la red es rica en relaciones entre pares. Con respecto a la cohesión, se calculó el diámetro de la red cuyo valor resultó 1.83, lo cual indica que sus componentes están bien conectados dado cualquier par de nodos se encuentran separados por 2 enlaces promedio.

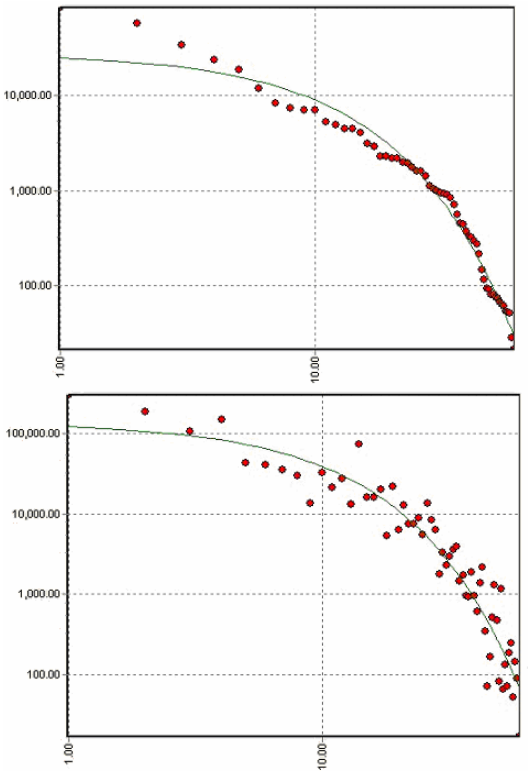

Gráfico 5 – Distribuciones de cantidad de páginas (arriba) y cantidad de enlaces (abajo) para el *DomainGraph*

Respecto de los nodos, se obtuvieron algunas de las medidas de centralidad. Particularmente, se calcularon grado, betweenness y valor de eigenvector. En todos los casos, el orden de los 3 primeros dominios fue el mismo resultando UBA, UNSE y UTN. En la figura 4 se muestra el núcleo de la red de universidades, el cual se construyó con aquellos nodos que poseen 10 o más enlaces con pares, siendo su tamaño proporcional a su grado.

5 – Discusión y Trabajos Futuros

En este trabajo se presenta el análisis de una muestra del espacio web educativo de Argentina que corresponde a las universidades. Se realizó una caracterización a tres niveles, encontrando parámetros similares a otros muestreos del espacio web.

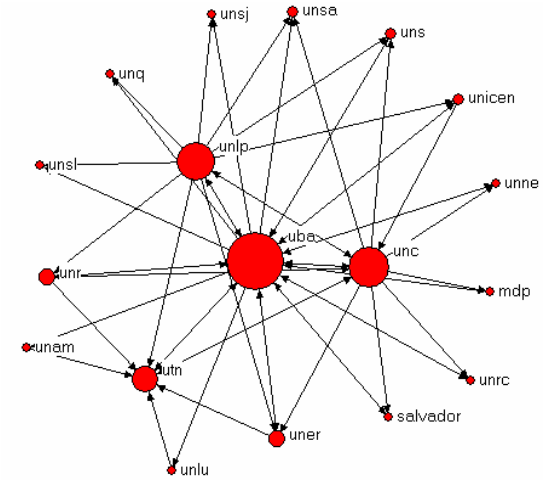

Figura 4 – Núcleo de la red de universidades

Si bien esta es una tarea exploratoria, resulta una primera aproximación válida dado que la web presenta características de invarianza, es decir, posee propiedades similares a varias escalas.

Esta investigación es parte de un proyecto mayor de nuestro grupo orientado al estudio de diferentes espacios web educativos y de la web Argentina completa. Consideramos que su caracterización, junto con la información de su uso (existente en los archivos de auditoria de cada institución) permite proponer optimizaciones sobre gestión de los recursos digitales existentes y planificar su desarrollo. Esto es especialmente interesante si se tiene en cuenta que las herramientas de búsqueda actuales se basan en información de enlaces para realizar los rankings.

6 – Agradecimientos

Agradecemos especialmente a Debora Donato del Dipartimento di Informatica e Sistemistica, Universita' di Roma ``La Sapienza" por su asesoramiento en la utilización de las herramientas de la librería COSIN.

7 – Referencias

[1] Adamic, L., B. Orkut, & E. Adar. A social network caught in the Web. First Monday, Volume 8, Number 6, Junio 2003.

[2] R. Baeza-Yates, and F. Lalanne. Characteristics of the Korean Web. Technical Report, Korea-Chile IT Cooperation Center, ITCC, 2004.

[3] R. Baeza-Yates and C. Castillo. Características de la Web Chilena 2004. Technical Report, Center for Web Research, University of Chile, 2005.

[4] R. Baeza-Yates, C. Castillo and V. Lopez. Characteristics of the Web of Spain. Cybermetrics, Vol. 9, No. 1, 2005.

[5] R. Baeza-Yates, and C. Castillo. Link Analysis in National Web Domains. Workshop on Open Source Web Information Retrieval (OSWIR), pp. 15-18. Compiegne, France, September, 2005.

[6] A. L. Barabasi and A. Albert. Emergence of Scaling in Random Networks. Science, (286): 509-512, 1999.

[7] P. Bonacich. Power and centrality: A family of measures. American Journal of Sociology, 92, pp. 1170-1182. 1987.

[8] 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.

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

[10] S. Chakrabarti, B.E. Dom, D. Gibson, D., and J. Kleinberg. Mining the Link Structure of the World Wide Web. IEEE Computer, Vol. 32, No. 8, pp: 60-67

[11] S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. ACM Transactions on Internet Technology, 2(3):205-223, 2002.

[12] E. Efthimiadis and C. Castillo. Charting the Greek Web. In Proceedings of the Conference of the American Society for Information Science and Technology (ASIST), Providence, Rhode Island, USA, November, 2004.

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

[14] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a Graph: Measurements, Models and Methods. In Proceedings of the International Conference on Combinatorics and Computing, 1999.

[15] L. Laura, S. Leonardi, and S. Millozzi. A software library for generating and measuring massive webgraphs. Technical Report 05-03, Dipartimento di Informatica e Sistemistica, Universita' di Roma ``La Sapienza", 2003.

[16] M. Modesto, A. Pereira, N. Ziviani, C. Castillo and R. Baeza-Yates. Un Novo Retrato da Werb Brasileira. In Proceedings of SEMISH, São Leopoldo, Brazil, 2005.

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

[18] G. Pandurangan, P. Raghavan, and E. Upfal. Using Pagerank to characterize Web structure. In Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON), volume 2387 of Lecture Notes in Computer Science, pages 330--390, Singapore, August 2002. Springer.

[19] B. Wellman, J. Boase, & W. Chen. The networked nature of community online and offline. IT & Society, volume 1, number 1 (Summer), pp. 151-165. 2002.

[20] S. Wasserman and K. Faust. Social Network Analysis. Cambridge University Press, 1994.