Modelos y Algoritmos de Enlaces sobre el Grafo Web del Dominio Educativo Argentino

Pablo J. Lavallen[1], Gabriel H. Tolosa, Fernando R.A. Bordignon
Universidad Nacional de Luján, Argentina
Departamento de Redes
{plavallen,tolosoft,bordi}@unlu.edu.ar

Resumen

El estudio de las características de la Web, su dinamismo y el análisis de los distintos algoritmos que operan sobre ella se centran en modelar la misma como un grafo dirigido *(webgraph*). A partir de esto se pueden realizar diferentes tareas de análisis teniendo en cuenta la información aportada por este enfoque y que puede ser utilizada para mejorar las estrategias y herramientas que permiten la gestión de los recursos que se encuentran distribuidos en el espacio Web.

En este sentido, este trabajo presenta los primeros resultados de un estudio realizado sobre los sitios que conforman el espacio web educativo argentino, el cual permite analizar sus características básicas y diferencias con el resto de la web. El objetivo fundamental es lograr una mejor comprensión de las interacciones que ocurren en este sistema distribuido y dinámico de gran escala, que surge y se desarrolla a partir de las acciones no coordinadas de sus usuarios. Estos resultados permitirán – a priori – mejorar las estrategias de localización, organización y acceso a la información, lo que redundará en un uso más eficiente de los recursos disponibles, optimizando heramientas existentes o creando nuevas adaptadas especialmente al dominio educativo argentino.

1 – Introducción

El estudio de las características de la Web, su estructura, dinamismo y particularidades ha abierto varias líneas de investigación, especialmente en cuanto a las estrategias de recuperación de información. En este sentido, aparecen nuevos desafíos y oportunidades para explorar a partir de estudiar la componente que diferencia a la web de otros conjuntos o corpus de documentos: los enlaces entre las páginas. A partir de éstos, Kumar [Kumar 2000_a] y Broder [Broder 2000] introducen la idea de modelar la web como un grafo dirigido (denominado *webgraph*) en el cual las páginas HTML corresponden con los nodos del grafo y los hipervínculos, sus aristas.

A partir de ello se pueden realizar diferentes tareas de análisis teniendo en cuenta la información aportada por este enfoque y que puede ser utilizada para mejorar las estrategias y herramientas que permiten la gestión de recursos que se encuentran naturalmente distribuidos. La web es – además – una red libre de escala [Barabasi 2000] en la cual algunas de sus variables – entre ellas, el grado entrante y saliente – siguen distribuciones que se modelan mediante una ley de potencias (*power law)* [Kumar 2000_a] [Broder 2000] [Barabasi 2000].

Sobre esta idea, Broder [Broder 2000] planteó una estructura para la web conocida como *bow-tie* en la cual se ubica a cada nodo (página) en una región diferente de acuerdo a su conectividad con el resto. Esta estructura está constituida por: a) un CORE, formado por el SCC (*Strongly Connected Component*) de mayor tamaño. Un SCC es el conjunto de todos los nodos que se pueden alcanzar siguiendo caminos a través de los enlaces dirigidos. b) un conjunto de nodos

1 Actualmente, se encuentra desarrollando su trabajo final, carrera Licenciatura en Sistemas de Información, UNLu.

denominado IN, formado por los nodos que alcanzan al CORE, pero que no pueden ser alcanzados desde éste. c) el conjunto OUT, que esta formado por los nodos que son alcanzables a partir de CORE pero que sus enlaces salientes no apuntan al CORE. d) TENDRILS: esta formado por los nodos que son alcanzables desde el IN, o que alcanzan al OUT pero no pasan por el CORE. e) ISLANDS que son los nodos que no se encuentran en ninguna de las clasificaciones anteriores. Esta estructura no solamente permite determinar la conectividad del grafo web sino que además facilita el estudio de su dinamismo [Baeza-Yates 2006].

El análisis de las características mencionadas anteriormente, conjuntamente con el análisis de los enlaces, brinda información valiosa que es utilizada con diferentes fines, entre los mas destacados se encuentran los algoritmos que utilizan dicha información para determinar la importancia de las páginas, entre otros. Los ejemplos más representativos de este tipo de algoritmos son: HITS [Kleinberg 1999_a] en el cual básicamente se proponen los conceptos: *autoridad* – sitios muy apuntados – y *hub* – sitios que apuntan a muchos - *,* que permiten indicar la "calidad" de una página y dicho valor es utilizado para su rankeo. Por otro lado, se encuentra el algoritmo PageRank [Page 1998] que utiliza los enlaces a una página para calcular su importancia, donde – básicamente – cada enlace a una página se toma como un voto, luego la página mas votada será mejor rankeada.

Otra de las aplicaciones de algoritmos que siguen esta línea de trabajo es la detección de *webspam* [Becchetti 2006] [Gyöngyi 2004] basándose en los enlaces de páginas web. La idea subyacente en el webspam es alterar la manera en la que los usuarios navegan la web generando "granjas de enlaces" que modifican el comportamiento de los algoritmos de rankeo. TrustRank [Gyöngyi 2004] es un ejemplo de utilización de un algoritmo que usa información de los enlaces de páginas web para detectar dichas "granjas de enlaces" y evitar el *webspam*. Por otra parte se han propuesto diferentes algoritmos que – utilizando los enlaces entre páginas – permiten identificar comunidades temáticas en la web. Se trata – básicamente – de identificar páginas que tienden a vincularse entre si pero que no se vinculan con otras comunidades [Kleinberg 1999_b].

En este trabajo se propone un estudio exhaustivo a nivel de enlaces de un subconjunto particular del grafo web consistente en el Dominio Educativo Argentino. En primer lugar, se ha mostrado en [Bordignon 2006] que los espacios web educativos presentan características propias respecto de otros dominios o de la web global, especialmente en su estructura de enlaces y conectividad. Además, si bien se han propuesto modelos de generación de grafos de la web [Albert 1999] [Kumar 2000_b], aún no se han realizado estudios sobre dominios particulares acotados. Para nuestro conocimiento, tampoco se han encontrado trabajos sobre aplicación de algoritmos que operan sobre enlaces a este tipo de dominios. Una mejor comprensión de las interacciones que ocurren en este sistema distribuido y dinámico de gran escala, que surge y se desarrolla a partir de las acciones no coordinadas de sus usuarios permitirá mejorar las estrategias de localización, organización y acceso a la información, especialmente dentro el dominio educativo argentino.

2 – Objetivos

El análisis del grafo web Educativo de Argentina brindará un caso de estudio real donde se podrá determinar cuales son sus propiedades y cuál es el modelo que mejor se ajusta a su estructura. A partir de éste, se podrán simular otros grafos que mantengan las mismas propiedades. Esto es particularmente interesante debido a que los modelos de grafos web deben considerar [Shah, 2005]:

a) Generación en tiempo relativamente corto sobre el equipamiento disponible.
b) Deben ajustarse de la mejor manera posible al grafo real, por ejemplo, en las distribuciones de enlaces.

c) Se debe poder escalar en número de nodos y aristas para poder simular la evolución de la estructura.

Además, este grafo servirá para poder probar el funcionamiento de diversos algoritmos en diferentes escalas, permitiendo describir y ejemplificar algoritmos de tratamiento de enlaces para diferentes tareas como – por ejemplo – ranking, detección de spam y detección de comunidades.

Adicionalmente se codificará una pieza de software que sirva como *front-end* de la librería COSIN[Laura 2003_b], la cual se utilizará para el análisis del grafo web. La misma facilitará la tarea de análisis y la emisión de reportes a partir de automatizar procesos intermedios y de generar estructuras de datos específicas para los diferentes casos.

3 – Metodología

Para el trabajo propuesto es necesaria una muestra representativa del Dominio Educativo Argentino. Para ello se requiere de una etapa de recolección (*crawling*) de páginas, la cual ya ha se ha llevado a cabo. Luego, a partir de diferentes algoritmos que operan sobre enlaces se estudiarán las particularidades del mismo, especialmente en cuanto a los modelos de generación de grafos a partir de distribuciones de grado, grafos aleatorios y grafos con características de mundos pequeños [Watts 1998].

3.1– Recolección de datos

Para la etapa de recolección (*crawling*) se utilizó el software WIRE [Castillo, 2005] configurado para descargar sólo páginas cuyo dominio de segundo nivel sea *.edu.ar*. La semilla inicial de direcciones estaba compuesta por una lista de 3.612 sitios. La tarea se llevó a cabo en el mes de enero de 2007 utilizando un servidor con un procesador de 2.5 Ghz, 1 GB de memoria RAM y Sistema Operativo Linux.

3.2 - Análisis

La tarea de análisis se llevará a cabo utilizando la metodología clasica de área, la cual incluye: estudio de las distribuciones de grado entrante y saliente, tanto a nivel de páginas (webgraph) como de sitios (hostgraph). Se aplicarán los algoritmos de análisis de enlaces para ranking como PageRank [Page 1998] y HITS [Kleinberg 1999_a] y se estudiará el rol de los nodos de acuerdo a diferentes criterios (grado, centralidad, reputación, etc.). Además, se utilizarán métodos de detección de comunidades web [Kleinberg 1999_b] y se estudiará la existencia de *webspam* [Becchetti 2006].

Por otro lado, se aplicarán diferentes modelos de generación de grafos web [Kumar 2000_b] [Shah 2005] a los efectos de determinar el modelo que representa de mejor manera al dominio educativo argentino.

4 – Resultados Preliminares

Los resultados obtenidos a partir de la tarea de *crawling* indican un total de 2.236 sitios recolectados, lo que muestra un crecimiento respecto a los 1.964 sitios hallados en [Bordignon, 2006]. De los sitios recolectados el 37,7% no tiene links salientes, es decir no apuntan a ninguna otra página, y el 47,6% no tiene links entrantes, esto significa que no son apuntados por ninguna página.

Para las distribuciones de grado entrante y saliente se hallaron los ajustes mediante una ley de potencias. En el primer caso, se encontró un ajuste con un exponente ß = 1,99 (Figura 1, izq) mientras que para la distribución de grado saliente se halló de ß = 4,20 (Figura 1, der).

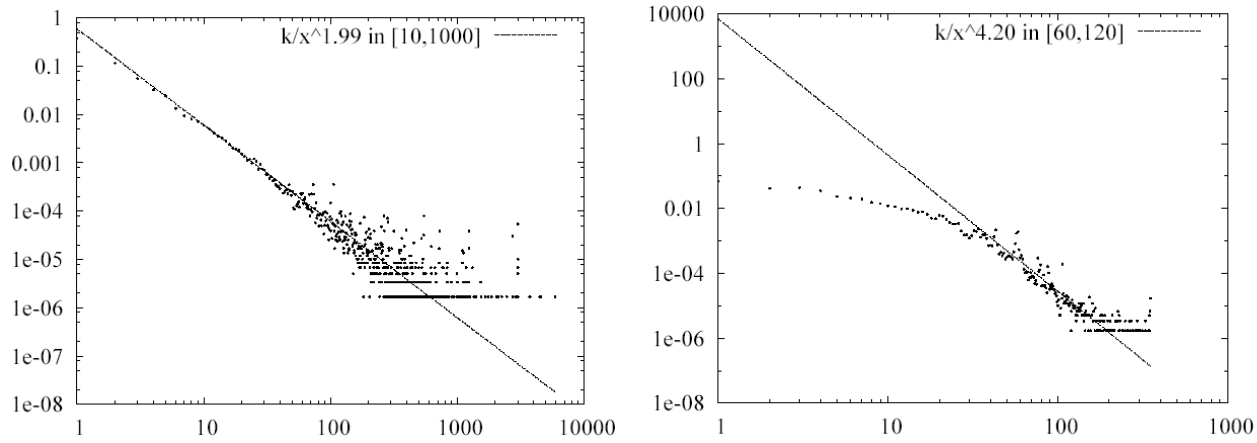

Figura 1. Distribuciones de grado entrante (izquierda) y saliente (derecha) con su recta de ajuste correspondiente. (los ejes x e y se encuentran en escala logarítmica)

Complementariamente, se determinó la estructura de la web de acuerdo al modelo de Broder [Broder 2000]. Como se dijo anteriormente, esta estructura ubica a los nodos de acuerdo a su conectividad con el resto (Figura 2)

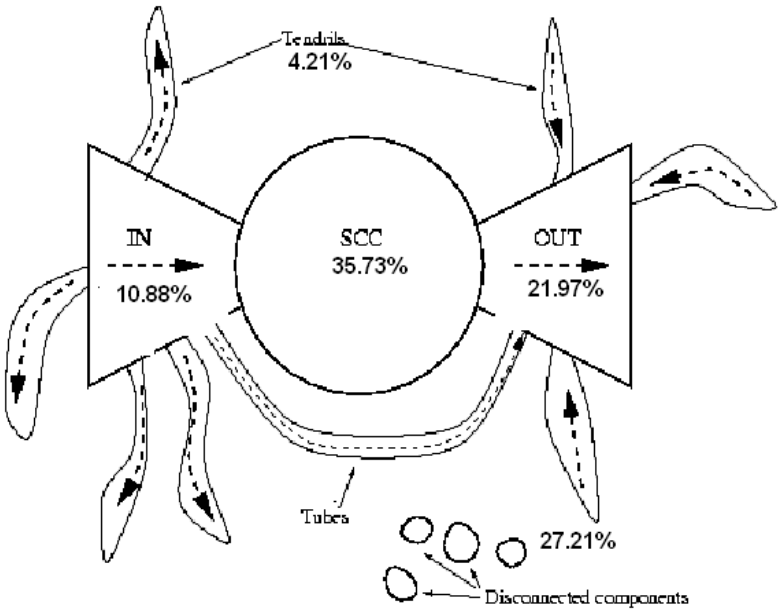

Figura 2. Estructura de *bow-tie* [Broder 2000] del espacio web educativo argentino en 2007

El análisis del modelo de Broder muestra que en el dominio educativo de argentina los tamaños de las regiones se distribuyen de manera diferente que en la web global [Broder 2000]. El tamaño del CORE evidencia buena conectividad, mientras que existe una fracción significativa de elementos desconectados (27,21%). No obstante, el tamaño del CORE ha aumentado respecto de muestreos anteriores [Bordignon 2006], disminuyendo las demás. Esto evidencia una evolución

positiva en la estructura de enlaces. Actualmente, se está trabajando con los modelos de generación de grafos web a partir de secuencias de distribuciones de grado y se están realizando tareas de prueba de los diferentes algoritmos que operan sobre los enlaces.

5 – Referencias

[Albert 1999] R. Albert, H. Jeong, and A.-L. Barabasi. *Diameter of World-Wide Web*. Nature, vol. 410, pp. 130-131, September , 1999.

[Baeza-Yates 2006] R. Baeza-Yates, B. Poblete. *Dynamics of the Chilean Web Structure.* Computer Networks: The International Journal of Computer and Telecommunications Networking Vol. 50, 2006.

[Barabasi 2000] A Barabasi, R Albert, H Jeong. *Scale-free characteristics of random networks: the topology of the world-wide web.* Physica A 281, pp 69-77, 2000.

[Becchetti 2006] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates. *Link-based characterization and detection of web spam.* In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2006.

[Bordignon 2006]]Bordignon, Fernando R. A. and Tolosa, Gabriel H. *Characterization of South American Educational Web Domains* . In Proceedings Congreso Argentino de Ciencias de la Computación. CACIC 2006. 2006.

[Broder 2000] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, A. y J. Wiener. *Graph Structure in the Web*. 9th International World Wide Web Conference / Computer Networks, 33(1-6), pp. 309-320, 2000.

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

[Gyöngyi 2004] Z. Gyöngyi, H. Garcia-Molina, J. Pedersen. *Combating Web Spam with TrustRank.* Thirtieth International Conference on Very Large Data Bases, Canada, August-September, 2004.

[Haveliwala 2002] T. Haveliwala. *TopicSensitive PageRank.* 11th international conference on World Wide Web 2002, May 07 - 11, 2002

[Kleinberg 1999_a] J. Kleinberg. *Authoritative Sources in a Hyperlinked Environment*. ACM, 46(5), pp 604-632, 1999.

[Kleinberg 1999_b] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. *The Web as a graph: measurements, models and methods*. 5th International Computing and combinatorics Conference.1999.

[Kumar 2000_a] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. *The Web as a graph.* 19th ACM Symposium on Principles of Database Systems, 2000.

[Kumar 2000_b] R Kumar, P Raghavan, S Rajagopalan D. Sivakumar. *Stochastic Models for the Web Graph.* IEEE Symposium on Foundations of Computer Science (FOCS.). pp 57-65, 2000.

[Laura 2003_a] L. Laura, S. Leonardi, S. Millozzi, U. Meyer, J. Sibeyin. *Algorithms and Experiments for the Webgraph*. 11th Annual European Symposium on Algorithms , 2003.

[Laura 2003$_b$] L.Laura, S. Leonardi, 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

[Page 1998] S. Brin, L Page. *The anatomy of a large-scale hypertextual Web search engine*. 9th international conference on World Wide Web 7. pp 107-117. 1998

[Shah 2005] Shah, S, *Generating a web graph*. M.Eng. report, Computer Science Department, Cornell University, 2005. http://www.infosci.cornell.edu/SIN/WebLib/papers/Shah2005a.doc.

[Watts 1998] D. Watts, S. Strogatz. *Collective dynamics of small-world networks.* Nature 393, 440. 1998.