Cibermetría del Web: Las leyes de exponenciación.

Autores:

José Luis Alonso Berrocal
Carlos G. Figuerola
Ángel F. Zazo Rodríguez

Universidad de Salamanca. Facultad de Documentación.

E-mail:

[berrocal | figue | afzazo]@gugu.usal.es

Dirección Postal:

Facultad de Documentación

C/ Francisco Vitoria, 6-16, 37008 – Salamanca

Tfno: +34 923 294580 Fax: +34 923 294582

Resumen:

Se realiza una introducción a las leyes de exponenciación, enunciadas por Michalis Faloutsos y que nos permiten realizar una caracterización del Web a través del análisis de su topología. Se describen sus características más importantes y cómo se calculan algunos de los valores de las funciones más interesantes.

Palabras clave:

Leyes de exponenciación, web, internet, grafos, topologías de red, cibermetría.

Abstract:

An introduction to the power laws, enunciated by Michalis Faloutsos, is made and that allows us to make a characterization of the Web through the analysis of their topology. Their most important characteristics are described and how calculate some of the values of the most interesting functions.

Keywords:

Power laws, web, internet, graphs, networks' topologies, cybermetrics.

1. Introducción

El estudio del World Wide Web se está convirtiendo en uno de los campos de investigación más interesantes y como dice (Kleinberg, 1999) pocos eventos de la historia de la computación han tenido tanta influencia en la sociedad como la llegada y crecimiento del World Wide Web. Precisamente este crecimiento (2000 millones de páginas Web en el verano del 2000 según (Aguillo, 2000)) y esta influencia (basada en los contenidos) han creado un sistema de comunicación de información muy potente, pero que al mismo tiempo tiene enormes carencias desde el punto de vista documental y por ello es necesario abordar su estudio.

Para algunos autores (Turnbull, 1996) este estudio debe realizarse con las técnicas bibliométricas clásicas y de análisis de citas, sin embargo es necesario realizar otros estudios y abrir nuevas vías de investigación que nos permitan caracterizar adecuadamente el Web, porque no hay que olvidar que el tipo de información con el que estamos trabajando, por ejemplo, tiene unos niveles de permanencia (Koehler, 1999) concretos que nos obliga a ajustar nuestras técnicas de estudio.

1.1. Trabajos relacionados.

Los trabajos que tratan de analizar en alguno de sus aspectos el Web son amplios y variados en cuanto a pretensiones y estudio, por ello indicaremos algunos de los que consideramos más interesantes.

En primer lugar estarían los trabajos que estudian o tratan de analizar aspectos cuantitativos del Web, como son todos los aspectos relacionados con tamaño del Web, tamaño de las páginas, tipos de etiquetas, y el cálculo de determinados índices como el factor de impacto Web (WIF), visibilidad, luminosidad, densidad de enlaces, endogamia, etc. diseminados en varios trabajos como los de (Aguillo, 2000), (Arellano, 1999), (Bray, 1996), (Ingwersen, 1998), (Larson, 1996), (Woodruff, 1996).

También nos encontramos con otras investigaciones que tratan de analizar la topología web, la estructura hipertexto, su diseño y sus características, creando una serie de índices que caractericen dichas estructuras. Nos encontramos con índices como

los de Randic, Compactación, Stratum, etc., que se analizan en trabajos como los de (Almind, 1997), (Botafogo, 1991), (Botafogo, 1992), (Ellis, 1998), (Smeaton, 1995).

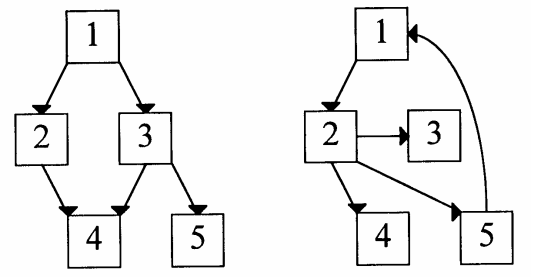
Finalmente, estarían todos aquellos trabajos que analizan el web desde el punto de vista de la recuperación de información, algoritmos de tratamiento de la información web, naturaleza y riqueza de los enlaces, etc., entre los que destacan trabajos como (Bernstein, 1992), (Bharat, 1998), (Chakrabarti, 1998), (Chakrabarti, 1999), (Figuerola, 1998), (Alonso, 1999).

Otra vía de trabajo novedosa es la de analizar la topología Web desde el punto de vista de unas nuevas leyes descubiertas por (Faloutsos, 1999), que puede ser otra vía alternativa de estudio. Precisamente sobre estas leyes vamos a realizar una introducción y comentaremos algunos de los resultados preliminares obtenidos en algunas de las investigaciones que tiene abiertas nuestro grupo de trabajo.

2. El Web visto como un grafo.

Para poder analizar la topología del Web es necesario realizar una adecuada recogida de información. Las teorías más interesantes trabajan sobre la base de considerar el Web como un grafo dirigido (Kleinberg, 1999), (Broder, 2000), (Ellis, 1994), y le aplican todas las teorías de grafos en su recogida y tratamiento posterior.

Debemos considerar cada página web como un nodo de dicho grafo y los enlaces de las páginas las aristas dirigidas de dicho grafo. Una representación de lo indicado anteriormente sería:



Para almacenar y tratar posteriormente esta información se generan matrices de adyacencia en las que se indica con un valor cero la no existencia de enlace y con un 1 la existencia del mismo, como se representa a continuación:

ij	1	2	3	4	5
1	–	1	1	0	0
2	0	–	0	1	0
3	0	0	–	1	1
4	0	0	0	–	0
5	0	0	0	0	–

ij	1	2	3	4	5
1	–	1	0	0	0
2	0	–	1	1	1
3	0	0	–	0	0
4	0	0	0	–	0
5	1	0	0	0	–

Una vez que tenemos dicha matriz, podemos realizar cualquier tipo de procesamiento de la misma, con infinidad de trabajos relacionados con ello como los de (Botafogo, 1992) (Ellis, 1994) que operan sobre dichas matrices y en la mayor parte de los casos obtienen índices que nos permiten caracterizar el grafo y por asociación el Web.

Las técnicas para realizar esta recogida de información son muy variadas, pero una metodología interesante y con desarrollos plenamente operativos se ofrecen en (Alonso, 1997).

En el caso que nos ocupa de las leyes de exponenciación, una vez realizada la recogida de datos y que tenemos las matrices de adyacencia correspondientes podemos operar sobre las mismas para obtener los datos de interés. Básicamente obtener el grado de apertura de cada nodo determinado por la siguiente fórmula $? \sum_{j?1}^{n} a_{ij} ? ?_i$, obteniendo la suma de los valores de la fila i.

También debemos obtener los valores propios de la matriz y con muy poco más podemos abordar el estudio de las leyes de exponenciación.

3. Las leyes de exponenciación.

Aparentemente el Web crece de forma aleatoria y sin mecanismos que de alguna forma regulen dicho crecimiento. Sin embargo, se han descubierto unas leyes muy sencillas que indican que la topología Web sigue algunas pautas de funcionamiento que son interesantes para analizar el Web y que pueden ser utilizadas para su análisis.

Los estudios de (Faloutsos, 1999) han determinado que las topologías Web siguen leyes del tipo $y ? x^{?}$, (similares a la de la ley de Zipf) dando lugar a cuatro leyes, que caracterizan dicha topología y que pasamos a comentar de forma breve.

Cada una de las leyes se caracteriza por tener un valor único para todos los datos analizados, y este valor es un exponente que nos va a permitir identificar diferentes grafos y por lo tanto realizar comparaciones.

Vamos a indicar en primer lugar algunos símbolos básicos, empleados en este tipo de trabajos, que van a ser utilizados:

Símbolo	Definición
G	Grafo
N	Número de nodos en el grafo
E	Número de aristas en el grafo
$?$	Diámetro del grafo
d_v	Grado de apertura del nodo v, definido como $? \sum_{j?1}^{n} a_{ij} ? ?_i$, es decir la suma de todos lo valores de una fila
$\overline{d}$	Media aritmética del grado de apertura de los nodos del grado, definida como $\overline{d} ? 2E/N$
f_d	Frecuencia de un grado de apertura d, que es el número de nodos con el grado de apertura d
r_v	Orden del nodo v, que es un índice en orden decreciente del grado de apertura
P(h)	Número de pares de nodos con menor o igual número de saltos
$?$	Valores propios de la matriz.
i	El orden de $?_i$

3.1. Ley 1. Exponente de Orden R.

El grado de apertura, d_v, de un nodo v, es proporcional al orden del nodo, r_v, elevado a un exponente, R,

$$d_v ? r_v^R$$

El exponente de orden R, es la pendiente que se obtiene con la representación del grado de apertura de los nodos frente al orden de los nodos en una escala logarítmica.

Una de las utilidades de este exponente, es que nos permite comparar diferentes topologías, diferenciando diferentes representaciones del grafo Web.

De esta primera ley se pueden sacar algunos lemas que completan dicha ley y nos ofrecen nuevos valores útiles en la caracterización de dicha topología.

En primer lugar, si consideramos que el mínimo grado de apertura de un nodo es 1 ($d_N = 1$), podemos decir que el grado de apertura, d_v, de un nodo v, es una función del orden del nodo, r_v, y del exponente R de la siguiente forma $d_v ? \frac{1}{N^R} r_v^R$. Aplicando este lema podemos relacionar el número de aristas con el número de nodos, N, y el exponente R

$$E ? \frac{1}{2 ? R ? 1 ? ? 1 ? \frac{1}{N^{R?1}} ? N}$$

Se ha estimado que el número de aristas obtenidas mediante la aplicación de este lema, difiere entre 9-20% de los datos reales.

3.2. Ley 2. Exponente de Grado de Apertura O.

La frecuencia, f_d, con un grado de apertura, d, es proporcional al grado de apertura elevado a un exponente, O.

$$f_d ? d^O$$

El exponente del grado de apertura O, es la pendiente que se obtiene con la representación de la frecuencia del grado de apertura frente a los grados de apertura en una escala logarítmica.

La presencia de esta ley indica que la distribución de los grados de apertura de los nodos Web no es arbitraria, y los nodos con un grado de apertura bajo son más frecuentes.

3.3. Ley 3. Exponente de representación de saltos H.

El número de pares de nodos, P(h), con h saltos, es proporcional al número de saltos elevado a un exponente H.

$$P(h) \propto h^{H}, \text{ si } h \gg 1$$

El exponente de salto H, es la pendiente que se obtiene con la representación de los pares de nodos P(h) con h saltos frente al número de saltos en una escala logarítmica.

Este exponente representa la conectividad de los grafos diferenciando eficientemente familias de grafos.

De la aplicación de esta ley, podemos sacar algunos datos muy útiles como la de calcular el diámetro efectivo del Web, diferenciándolo de los datos aportados por (Albert, 1999) para calcular dicho diámetro.

Dado un grafo de N nodos, E aristas y un exponente de salto H, podemos definir el diámetro efectivo, δ_{ef} como:

$$\delta_{ef} \approx \left(\frac{N^2}{N \cdot 2E} \right)^{1/H}$$

3.4. Ley 4. Exponente de valores propios λ.

Los valores propios, λ_i de un grafo son proporcionales al orden i, elevado a un exponente, λ.

$$\lambda_i \propto i^{\lambda}$$

El exponente de valores propios $?$ es la pendiente que se obtiene al representar los valores propios frente a su orden en una escala logarítmica.

Este exponente también nos permite caracterizar diferentes topologías Web, que además es independiente del crecimiento de dicho Web.

Los valores propios de una matriz están relacionados con algunas propiedades como pueden ser el diámetro, el número de aristas, el número de componentes conectados, el número de rutas existentes que poseen una determinada longitud, que son aspectos fundamentales dentro del análisis de cualquier topología.

3.5. Nuevos desarrollos de las leyes de exponenciación.

Estas leyes empiezan a ser ampliamente estudiadas y uno de los mejores trabajos es el de (Medina, 2000) que mediante generación de topologías de red en laboratorio ha extraído algunas conclusiones interesantes de las mismas.

Una de las primeras conclusiones es que las leyes 1 y 2 necesitan para aparecer de un crecimiento exponencial en la topología de red, permitiendo redes abiertas que aceptan nuevos nodos continuamente y de una conectividad preferencial, indicando una tendencia de los nuevos nodos a conectarse a nodos existentes con un alto grado de apertura.

En las leyes 3 y 4 no se precisan de estos dos parámetros y en las investigaciones los valores de su exponente no refleja ninguna variación en topologías que poseían estas características y en las que no se encontraban presentes.

4. Conclusiones.

Las leyes de exponenciación son un instrumento potente para analizar el Web y medir algunas de sus características más destacadas, en especial las relacionadas con la topología.

En comparación con otras técnicas de análisis de la topología, las leyes de exponenciación tienen la ventaja de un rápido y fácil procesamiento, que permiten en muy poco tiempo disponer de datos reales que posibiliten su estudio.

5. Bibliografía.

(Aguillo, 2000) AGUILLO, I. F. Indicadores hacia una evaluación no objetiva (cuantitativa) de sedes web . *Jornadas Espanolas de Documentación*, 2000, Vol. 7, p. 233-248.

(Albert, 1999) ALBERT, R., JEONG, H. y BARABÁSI, A.-L. The Diameter of the World-Wide Web. *Nature*, 1999, Vol. 401, p. 130-131.

(Almind, 1997) ALMIND, T. C. y INGWERSEN, P. Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, September 1997, Vol. 53, No. 4, p. 404-426.

(Alonso, 1997) ALONSO BERROCAL, J. L. *Herramienta software para el análisis de la documentación WEB: rastreo de dominios, estudio de etiquetas, tipología de ficheros, evolución de los enlaces*. Salamanca: Universidad de Salamanca, Facultad de Traducción y Documentación, 1997.

(Alonso, 1999) ALONSO BERROCAL, J. L., FIGUEROLA, C. G. y ZAZO RODRÍGUEZ. ÁNGEL FRANCISCO. Representación de páginas web a través de sus enlaces y su aplicación a la Recuperación de Información. *Scire. Representación y Organización del Conocimiento*, 1999, Vol. 5, No. 2, p. 91-98.

(Arellano, 1999) ARELLANO PARDO, C., RODRÍGUEZ MATEOS, D., NOGALES FLORES, J. T. y HERNÁNDEZ PÉREZ, T. Análisis de estructura de sitios web: el caso de las bibliotecas universitarias andaluzas. *2as. Jornadas Andaluzas de Documentación, JADOC'99*, (Granada, 1999), p. 39-50.

(Bernstein, 1992) BERNSTEIN, M. Contours of Constructive Hypertexts. *Proccedings of ACM ECHT CONFERENCE*, (Milano, 30 Noviembre-4 Diciembre de 1992), p. 161-170.

(Bharat, 1998) BHARAT, K. y HENZINGER, M. R. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information retrieval*, (1998), p. 104-111.

(Botafogo, 1992) BOTAFOGO, R. A., RIVLIN, E. y SHNEIDERMAN, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*, April 1992, Vol. 10, No. 2, p. 142-180.

(Botafogo, 1991) BOTAFOGO, R. A. y SHNEIDERMAN, B. Identifying aggregates in Hypertext structures. *Proceedings of Hypertext'91*, (Diciembre de 1991), p. 63-74.

(Bray, 1996) BRAY, T. Measuring the Web. *Fifth International World Wide Web Conference*, (Paris, France, 6-10 May 1996).

(Broder, 2000) BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. y WIENER, J. Graph structure in the web. *9th. International World Wide Web Conference*, (Amsterdam, May 15 - 19, 2000).

(Chakrabarti, 1999) CHAKRABARTI, S., DOM, B., GIBSON, D. y KLEINBERG, J. Mining the link structure of the World Wide Web. *IEEE Computer*, August 1999.

(Chakrabarti, 1998) CHAKRABARTI, S. y DOM, B. I. P. Enhanced hypertext categorization using hyperlinks. *Proceedings ACM SIGMOD*, (1998).

(Ellis, 1998) ELLIS, D., FORD, N. y FURNER, J. In search of the unknown user: indexing, hypertext and the world wide web. *Journal of Documentation*, January 1998, Vol. 54, No. 1, p. 28-47.

(Ellis, 1994) ELLIS, D., FURNER-HINES, J. y WILLETT, P. On the creation of hypertext links in full-text documents: measurement of inter-linker consistency. *Journal of Documentation*, June 1994, Vol. 50, No. 2, p. 67-98.

(Faloutsos, 1999) FALOUTSOS, M., FALOUTSOS, P. y FALOUTSOS, C. On power-law relationships of the internet topology. *ACM SIGCOMM*, (Cambridge, MA, September 1999), p. 251-262.

(Figuerola, 1998) FIGUEROLA, C. G., ALONSO BERROCAL, J. L. y ZAZO RODRÍGUEZ, Á. F. Nuevos puntos de vista en la Recuperación de Información en el Web. *Jornadas Espanolas de Documentación*, 1998, Vol. 6, p. 273-280.

(Ingwersen, 1998) INGWERSEN, P. The calculation of web impact factors. *Journal of Documentation*, March 1998, Vol. 54, No. 2, p. 236-243.

(Kleinberg, 1999) KLEINBERG, J. M., KUMAR, R. y RAGHAVAN, P. The web as a graph: measurements, models, and methods. *Proceedings of the Fifth Annual International Computing and Combinatorics Conference*, (1999).

(Koehler, 1999) KOEHLER, W. C. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 1999, Vol. 50, No. 2, p. 162-180.

(Larson, 1996) LARSON, R. R. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. *Annual meeting of the American Society for Information Science*, (Baltimore, October 19-24, 1996), p. 71-78.

(Medina, 2000) MEDINA, A., MATTA, I. y BYERS, J. On the origin of power laws in internet topologies. *Computer Communication review*, 2000, Vol. 30, No. 2.

(Smeaton, 1995) SMEATON, A. F. Building hypertext under the influence of topology metrics. *International Workshop on Hypermedia Design*, (Montpellier, June 1995).

(Turnbull, 1996) TURNBULL, D. (1996). Bibliometrics and the World-Wide Web. Url: http://donturn.fis.utoronto.ca/research/bibweb.pdf [Consulta: Diciembre 1998].

(Woodruff, 1996) WOODRUFF, A. An Investigation of Documents from the World Wide Web. *Fifth International World Wide Web Conference* , (París, May 6-10 1996).