ENFOQUE DE REDES DE MUNDO PEQUEÑO EN EL ANÁLISIS DE MEDIDAS TOPOLÓGICAS DEL WEB

Flores Rios Brenda L. [1], Ibarra Esquer Jorge E. [2], Cáceres González Abdiel Emilio [3] y Burtseva Larisa [4]
[1,4] Instituto de Ingeniería. [2] Facultad de Ingeniería.
Universidad Autónoma de Baja California
Calle de la Normal s/n y Blvd. Benito Juárez. Col. Insurgentes Este. C.P. 21280.
Mexicali, Baja California. México. Tel/Fax (686) 5664150.
[1,4] {bflores,lpb}@iing.mxl.uabc.mx [2]jorgeeie@uabc.mx

[3] Universidad Juárez Autónoma de Tabasco
Km. 1 Carretera Cunduacán Jalpa de Méndez. C.P. 86690 A.P. 24.
Conduacán, Tabasco. Tel/Fax (914) 3360300
abdiel.caceres@dacb.ujat.mx

RESUMEN
Se presenta un estudio cibermétrico realizado con los datos históricos de un sitio Web registrados durante 3 años. Este permitió detectar el fenómeno de red de mundo pequeño al modelar el comportamiento del sitio por medio de medidas topológicas y utilizar el modelo de Kohonen para descubrir las relaciones topológicas. La variabilidad en el grado de conexión entre los nodos del grafo, resultó contar con un diámetro pequeño, debido a las conexiones de largo alcance creadas para acceder a un mismo recurso.

1. INTRODUCCIÓN
A partir de 1998, la cibermetría se ha dedicado a analizar las pautas y mecanismos de crecimiento del Web [1]. Una de las primeras herramientas matemáticas que se usaron para desarrollar estudios cibermétricos fueron las series de tiempo, las cuales ayudaron a medir el comportamiento de los motores de búsqueda cuando son utilizados con búsquedas de una sola palabra común [2]. Una breve pero interesante descripción de lo que son las series de tiempo se encuentra en [3], donde dice que una serie de tiempo (o serie temporal) es una lista de números que representan el valor de la magnitud observada en intervalos regulares de tiempo. En varias investigaciones del Web se usan técnicas de estadística multivariadas [4, 5, 6] y modelos de mapas auto-organizados SOM (*Self-Organizing Maps*) [7, 8] para estudios cibermétricos. El modelo SOM ha sido empleado por su capacidad para descubrir las relaciones topológicas entre los espacios Web y sus variables específicas [8]. Por otro lado, el Web es considerado un sistema complejo [9, 10, 11], debido a que su estructura se forma de nodos que representan computadoras y las aristas las conexiones entre ellas [12]. Las redes sociales identifican la presencia de una organización jerárquica, ocasionando que grupos pequeños de nodos estén organizados jerárquicamente creando grupos más grandes, mientras se mantiene una topología de ausencia de escala característica [13]. Estas redes se manifiestan en formas diversas, tales como redes aleatorias de mundo pequeño, redes exponenciales de mundo pequeño y las redes fractales de mundo pequeño [14].

En el presente trabajo se expone la estrategia utilizada para conocer si los datos almacenados en el archivo histórico, el cual registra los ingresos al sitio Web bajo estudio, contenían características de redes de mundo pequeño. Se utilizaron medidas topológicas para reflejar las relaciones entre los nodos y las conexiones principales del Web, presentando las propiedades estructurales en forma de una red de pequeño mundo [15].

2. DIÁMETRO DEL WEB
Aparentemente, el Web crece de una forma aleatoria y sin mecanismos que de alguna forma regulen su crecimiento. Sin embargo, la topología del Web sigue algunas pautas de funcionamiento que han sido objeto de análisis. Los estudios de Faloutsos determinaron que la topología del Web tiene un comportamiento del tipo

$$y = x^{\alpha} \qquad (1)$$

donde α es una constante [16].

El diámetro se puede definir como la medida de la distancia más corta existente entre dos nodos de un grafo [1], lo que implica que cuando un grafo tiene determinado diámetro, hay una determinada probabilidad de que exista un enlace entre cualesquiera dos nodos del grafo. En 1995, Govindan analizó 900 dominios obteniendo el diámetro del subgrafo del Web entre 9 y 10 nodos [17]. Así mismo, Albert, Jeong y Barabási [18] construyeron un grafo dirigido del Web con N nodos, asignando a cada nodo k enlaces entrantes o salientes, donde k es un número aleatorio tomado de acuerdo a una de las leyes de distribución de probabilidad. El estudio consistió en seleccionar de manera aleatoria un nodo i e incrementar sus enlaces entrantes o salientes a k_i+1 si el número total de vértices con k_i+1 enlaces entrantes o salientes es menor que $NP_{\text{salida}}(k_i+1)$ o $NP_{\text{entrada}}(k_i+1)$. Siendo $P_{\text{entrada}}(k)$ y $P_{\text{salida}}(k)$ las probabilidades de que un documento tenga k enlaces entrantes y salientes, respectivamente [18].

El número de enlaces d en el URL (*Uniform Resource Locator*), utilizados para navegar de un documento a otro, representa la ruta más corta entre sus dos elementos en el proceso de búsqueda. El resultado es que el valor aproximado de d para todos los pares de nodos se define de acuerdo con la fórmula:

$$d = 0.35 + 2.06 \log(N) \qquad (2)$$

donde para el número de nodos en el Web, representado por $N = 8\text{x}10^8$ [18], se obtiene el d_{Web}=18.59; es decir, dos documentos seleccionados de manera aleatoria en el Web se encuentran en promedio a solo 19 vínculos de distancia entre ellos. La función logarítmica de N en d es importante para determinar la tendencia de desarrollo del Web en el futuro. Para un incremento del 1,000% en el tamaño del Web pronosticado en los próximos años, el valor de d crecerá de 19 apenas hasta 21 enlaces, lo que justifica, que el Web es una red de mundo pequeño [18].

3. DESARROLLO

Desde 2004, el Departamento de Computación e Informática del Instituto de Ingeniería de la Universidad Autónoma de Baja California, ha examinado la información del archivo histórico (bitácora) de su sitio Web [19].

El análisis consistió en aplicar técnicas cibermétricas para obtener patrones de comportamiento o determinar las similitudes que pudieran existir entre los usuarios que visitan dicho sitio Web. En el análisis cuantitativo de los datos almacenados en el archivo histórico, se observó que estos se incrementaban mensualmente formando una bitácora grande, en comparación con los datos registrados para el mismo periodo durante los primeros años. Esto creó la necesidad de establecer estrategias para determinar nuevas acciones en el manejo de grandes de volúmenes de información. Se utilizaron Redes Neuronales Artificiales (RNA), tipo Mapas de Kohonen, para observar cúmulos de información por medio de mapas tipo SOM comprobando la relación entre lo que demandan los usuarios y lo que se ofrecía en el sitio bajo estudio. Del mapa SOM, se analizó si la gráfica poseía características de redes de mundo pequeño. La última fase, consistió en aplicar métodos de la teoría de grafos para tratar el sitio Web como un grafo dirigido, donde sus páginas se representaron por medio de nodos (o vértices) y sus enlaces por arcos o aristas [15].

3.1. El modelo de Kohonen

Las redes neuronales artificiales tipo SOM son también conocidas como modelo de Kohonen o RNA WTA (Winner-Take-All) [20] por la capa competitiva que clasifica las entradas de entrenamiento (*clustering*). En este proceso de competencia, la neurona que posee más características similares con el resto de neuronas que la rodean (neurona ganadora), conseguirá inhibir a todas las demás, por lo que será la única que permanezca activada. De este modo, ante el mismo patrón de entrada, la neurona ganadora responderá con mayor intensidad. Si el espacio está dividido en grupos, cada neurona se especializará en uno de ellos, y la operación de la red se podrá interpretar como un análisis de cúmulos o clusters [21].

Los parámetros que se seleccionaron del archivo histórico fueron dirección IP (*Internet Protocol*) de origen, fecha y hora de acceso, horario GMT y recurso solicitado. En ese mismo archivo, se registraban los *links* (enlaces) contenidos en el sitio Web y los datos son acompañados de corchetes, guiones y otros símbolos innecesarios por lo que se realizó una depuración de toda aquella información irrelevante en el desarrollo de los clusters de información. Posteriormente, los

datos obtenidos se importaron hacia una base de datos para su manipulación.

En el modelo SOM, las unidades de la capa oculta cercanas físicamente, responden a vectores de entrada que se encuentran igualmente próximos; estas neuronas se organizan en un mapa bidimensional. En este trabajo, se utilizó una red de 10*x*20 neuronas y una topología hexagonal, de modo que cada neurona tenía 6 neuronas vecinas. El entrenamiento se realizó utilizando dos fases:

1) 100,000 épocas (cada época es la representación de un nuevo vector en la red), con una tasa de aprendizaje de 0.05 y 10 de vecindad;
2) 1,000,000 de épocas, con una tasa de aprendizaje de 0.02 y 2 de vecindad.

El resultado mostró las distancias entre los vectores de pesos de neuronas colindantes, es decir, entre los centroides de los clusters a los que da lugar cada neurona (Figura 1). Su interpretación se realizó por medio de mapas topográficos, donde las zonas con tonalidades claras indicaron los valles en los que las neuronas estaban mejor comunicadas y, por tanto, más relacionadas, mientras que los colores oscuros denotaron las neuronas mas aisladas [21].

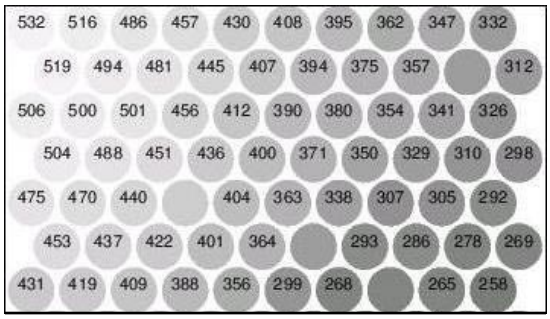

Figura 1. Mapa de cúmulos de los datos correspondientes a un año.

3.2. Redes de Mundo Pequeño

La teoría de sistemas complejos, desarrollada principalmente por científicos del campo de la física, es aplicable al estudio de estructuras sociales, tales como la comunicación electrónica, manejo de redes de energía eléctrica, transmisión de virus informáticos, entre otros [14]. En 1998, Watts y Strogatz expusieron que las redes de mundo pequeño pueden obtenerse a partir de una red regular, en la que cada nodo tiene el mismo número de enlaces [22].

Se toma un nodo y de manera aleatoria se decide borrar un enlace, sustituyéndose por otro enlace con otro nodo elegido también de manera aleatoria. Lo anterior, se repite para cada uno de los nodos, simulando conexiones entre computadoras que en el sitio Web estaban distantes pero que ahora son vecinas.

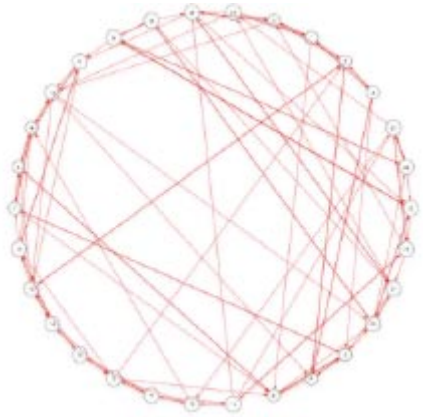

Figura 2. Ejemplo de red de mundo pequeño.

En 1999, Adamic concluye que las propiedades del mundo pequeño se aplican para el Web [23]. Sin embargo, desde el origen del Internet, las propiedades del mundo pequeño fueron remarcadas por Berners-Lee y Cailliau al especificar que es suficiente recorrer un número pequeño de enlaces para llegar de un lugar a cualquier otro [24]. La vinculación de científicos en redes de colaboración y redes semánticas de un lenguaje podrían presentar características de redes de mundo pequeño. Grafos correspondientes a este tipo de redes tiene la forma de cúmulos (clusters). La simulación computacional mostró que aquellas redes poseen una estructura intermedia entre un alto grado de orden y de carácter aleatorio, con un valor pequeño de longitud media de la ruta entre los nodos y un alto nivel de clusterización [22]; es decir, una pequeña probabilidad de ser vecinos para aquellos vértices, los cuales tienen a un vecino en común y habilitan una alta eficiencia en la propagación de la información, ideas, contactos, señas, energía, virus informáticos [25, 26].

Desde la década de 1990 se han aplicado métodos informétricos al estudio del ámbito del Internet. Varios aspectos de informetría, tales como la navegabilidad y la accesibilidad de la información, son aplicables para especificar la

estructura de enlace en redes de mundo pequeño [26].

El modelo matemático

Al estudiar la topología del Web, se podría suponer que los protocolos son independientes de la topología de la red, sin embargo se ha visto [27] que tienen un alto impacto en el desempeño global de la red. De manera que una alternativa para mantener un alto grado de buen desempeño es mantener un camino de conexiones del menos número de nodos.

Sobre el conjunto de vértices de un grafo G, puede haber una partición en subconjuntos no vacíos V_1, V_2, . . . , V_w tales que dos vértices u y v estan conectados sí y sólo si u y v pertencen al mismo subconjunto V_i. Los subgrafos $G[V_1]$,$G[V_1]$, ... ,$G[V_w]$ se llaman los *componentes* de G. Si G tiene exactamente un componente, entonces el grafo G es *conectado*.
Si de la figura 1, se toma un subconjunto con la zona de tonalidades claras, debido a que estas representan que las neuronas están mejor conectadas, y se representan por medio de un Grafo G. Entonces dos vértices *u* y *v* de *G* están *conectados*, sólo si existe un (*u*, *v*)−*camino* en *G*.

La relación *u-esta conectado con-v* es una relación de equivalencia sobre el conjunto de vértices *V*, porque la relación es:

• *Reflexiva:* $\forall\ v \in V$ se cumple que *v*R*v*, donde R es la relación *esta conectado con*. Vemos que la secuencia de nodos que forman el camino es P=*v*.
• *Simétrica:* $\forall\ u,v \in$ V , si *u*R*v* entonces *v*R*u*. Sea ahora $P = uv_1v_2 \ldots v_{k-1}v$ el camino desde *u* hasta *v*, vasta considerar $P^{-1} = vv_{k-1} \ldots v_2v_1v$.
• *Transitiva:* $\forall\ u,v,w \in$ V, si *u*R*v*, *v*R*w* entonces *u*R*w*. Lo que significa que si existe un camino $P = uv_1v_2 \ldots v_{k-1}v$ que une los vértices *u* y *v*, y otro camino $Q = vw_1w_2 \ldots w_{q-1}w$ que une los vértices *v* y *w*, entonces basta concatenar los caminos PQ para tener $PQ = uv_1v_2 \ldots v_{k-1}vw_1w_2 \ldots w_{q-1}w$, que conecta los puntos *u* y *w*.

Las redes de mundo pequeño tienen las siguientes características

1) La longitud característica de los caminos *L* es tan corta como se puede encontrar en las gráficas aleatorias.
2) El coeficiente de agregación C (*clustering coefficient*) es mucho mas grande del que se puede encontrar en una red aleatoria con el mismo número de nodos y el número promedio de aristas por nodo.

En el análisis realizado, un camino *L* entre dos IP distantes fue corto al tener el mismo interés sobre un recurso; es decir el diámetro definido es pequeño. Con respecto al coeficiente de agregación de la Figura 1, este se encuentra como sigue:

Sea *v* uno de los vértices del Grafo *G*, y que el vértice *v* tiene k_v vecinos, entonces pueden existir cuando mucho, $k_v\ x\ (k_v - 1)$ aristas dirigidas entre ellos. Entonces se denota por C_v la fracción que representa el número de aristas que efectivamente existen, y así *C* es el promedio de todas las C_{vi}, para cada vértice del grafo *vi*. De esta forma, los vértices representan las IP que accedieron a los recursos dentro del sitio Web y que comparten similitudes con sus k_v vecinos. Se detectaron 20 cuartetos de vértices donde cada cuarteto acceso a 20 recursos distintos.

4. CONCLUSIONES

En los últimos años, se ha incrementado fuertemente el interés por investigar la estructura topológica del Web. El aplicar medidas topológicas y RNA del tipo modelo SOM a estudios cibermétricos, permite visualizar las similitudes que comparten algunos datos bajo estudio.

El investigar el Web como un grafo no es un enfoque nuevo. Algunos trabajos dedicados al análisis del Web no se basan en este planteamiento [1]. El fenómeno de mundo pequeño fue comprobado en un experimento donde, para el total de nodos en el Web, evaluado por el número $8x10^8$ [28], se encontró que la distancia aproximada entre dos de ellos, seleccionados de manera aleatoria, es de 19 vínculos. Varios investigadores no comparten el punto de vista de autores referidos en este artículo. Así, según Berrocal [1], el diámetro del Web propuesto por Albert [28] no reflejó realmente el valor de la distancia debido a que su cálculo se realiza solamente a través del número de nodos y no se toma en cuenta la variedad del número de sus enlaces.

El enfoque de redes de mundo pequeño está orientado a examinar las distancias entre los nodos por medio de los enlaces. Las investigaciones muestran que el Web tiene propiedades estructurales de mundo pequeño. Así mismo, la importancia adicional de este enfoque es que permite contribuir en el análisis de propagación de virus informáticos; diseño y desarrollo de arquitectura de sistemas distribuidos; análisis de usabilidad de sitios Web; estructuras de servidores Web y Grid computacional, entre otras aplicaciones.

AGRADECIMIENTOS

Los autores desean expresar su agradecimiento a Rafael Villa Angulo, de la Universidad George Mason en EEUU, por la asesoría brindada para la aplicación del modelo SOM y a Jesús Alonso Rodríguez Godoy por el apoyo brindado en el proceso de experimentación.

5. REFERENCIAS

[1] J. L. Berrocal, C. G. Figuerola, A. F. Zazo y E. Rodríguez. "La Cibermetría en la recuperación de información en el Web". Universidad de Salamanca. España. 2004.

[2] R. Rousseau, dialy time series of common single word searches in Altavista and Nothern-Light, vol. 2/3. International Journal of Scientometrics, Informetrics and Bibliometrics, 1998-1999.

[3] I. Stewart,¿Juega Dios a los dados?: La nueva matemática del caos. Libro de Mano 48, Grijalbo Mondadori, S.A., en esta colección 1 Ed., 1986.

[4] R. R. Larson. "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace". Annual Meeting of the American Society for Information Science (ASIS). 1996. Disponible en http://sherlock.berkeley.edu/asis96/asis96.html. (Consultado el 31/03/2005)

[5] Chen y Cooper. "Using Clustering techniques to detect usage patterns in a Web-based information systems". Journal of the american society for information science& technology. Vol. 52. Num. 11 Pp. 888-904. 2001.

[6] J. Larsen, L. K. Hansen, A. S. Have, T. Christiansen y T. Kolenda. "Webmining learning from the World Wide Web". Computational Statistics & Data Analysis. Vol. 38. Num. 4. Pp. 517-532. 2002.

[7] T. Kohonen. "The Self Organizing Map". Proceedings of the IEEE. Pp. 1464-1480. 1995.

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero y A. Saarela. 1999. "Self organization of a massive text document collection". Editorial Oja, E y Kaski. Amsterdam. Pp. 171-182.

[9] P. Kihong. "The Internet as a Complex System". Department of Computers Sciences Purdue University.

[10] M. E. J. Newman. "The structure and function of structure networks". SIAM Review. Vol. 45. Num. 2. Pp. 167-256. USA. 2003.

[11] A. S. Balankin y J. Marquez González."Fractal behavior of complex systems". SIAM Review. Cientifica. Vol. 7 Num. 3. Pp. 109-128. USA. 2003.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins y J. Wiener. "Graph Structure in the Web". 9th International World Wide Web Conference. Amsterdam. 2000.

[13] Martínez, Balankin, O. Morales y E. García. "Modelación Fractal de la Red de Electores para diputados federales por el principio de mayoría relativa". 4to. Congreso Internacional de Ingeniería Electronmecánica y de Sistemas. México. 2005.

[14] M. San Miguel, R. Toral y V. M. Eguíluz. 2005. "Redes Complejas en la Dinámica Social". INGURUAK, Revista vasca de Sociología y Ciencia Política. Vol. 42. Pp. 127-146. Septiembre 2005.

[15] L. Björneborn. "Small-world link structures across an academic Web space: A library and information science approach". Doctoral thesis. Royal School of Library and Information Science. Copenhagen, Denmark. 2004.

[16] M. Faloutsos, P. Faloutsos y C. Faloutsos.. "On Power-Law relationships of the internet topology". ACM SIGCOMM. USA. Pp. 251-262. 1999

[17] R. Govindan y A. Reddy. "An analysis of Internet Interdomain topology and route stability". Proceedings IEEE INFOCOM. Kobe, Japan. 1997.

[18] R. Albert, H. Jeong y A. L. Barabási. "The Diameter of the World Wide Web. Nature". 1999. Vol. 401. Pp. 130-131. 1999.

[19] B. L. Flores Rios, J. A. Rodríguez Godoy y L. Burtseva. Estudio de cibermetría

aplicando medidas topológicas. ELECTRO 2005. Pág. 289-294. México.
[20] B. Martín del Brío y M. A. Sanz. "Redes Neuronales y Sistemas Difusos". Pp. 83-94. 1992.
[21] A. Rodríguez Godoy. "Estudio de Cibermetría para un sitio Web de la UABC utilizando Redes Neuronales Artificiales". Tesis de licenciatura. Universidad Autónoma de Baja California. México. 2005.
[22] D. J. Watts, y S. H. Strogatz. "Collective dynamics of small world networks. Nature". Vol. 393. June 1998. Macmillan Publishers. 1998.
[23] Adamic. "The small world Web". Proceedings of ECDL 1999. Pp. 443-452. 1999.
[24] T. Berners-Lee y R. Cailliau. "World Wide Web: proposal for a hypertext project". 1990. Disponible en: http://www.w3.org/Proposal.html (Consultado el 17/02/2006)
[25] M. Marchiori y V. Latora. "Harmony in the small world". Physica. Vol. 285. Pp. 539-546. 2000.
[26] L. Björneborn. 2004b. "Small World network exploration". Department of Information Studies. Royal School of Library and Information Science, Denmark.
[27] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W.Willinger, "Network topologies, power laws, and hierarchy," Tech. Rep. 01-746, Computer Science Department, University of Southern California. 2001.