Modelos de grafos para la web[1]

Carme Àlvarez[2] Josep Díaz[2] Maria Serna[2]

Uno de los grandes fenómenos sociales de finales de siglo es la popularización de internet y la facilidad de acceso a "información hipermedia" a través de la denominada *World Wide Web* o su abreviación la **web**. Los documentos o páginas electrónicas de la web pueden estar escritas en cualquier lenguaje y pueden contener información de cualquier tipo. Una característica relevante es la de contener enlaces a otras páginas. Cada persona o institución puede crear sus propias páginas cuando lo desee, se calcula que la web aumenta en un millón de páginas por día. Este crecimiento caótico implica una falta de organización y estructuración que repercute en la búsqueda eficiente de información. Un problema básico escomo extraer de la web una respuesta relevante a una petición concreta de información. En estos últimos cinco años se han realizado numerosos esfuerzos para encontrar propiedades topológicas de la web. Este conocimiento permitiría modelizarla y en consecuencia facilitaría el diseño de procedimientos eficientes para la búsqueda de información.

Desde un punto de vista estructural, podemos ver la web como un inmenso grafo dirigido. Un grafo dirigido se define como un conjunto de nodos o vértices y un conjunto de arcos o relaciones entre pares de nodos. En la web cada nodo es un documento o página, identificado por su URL[3], los arcos son los enlaces entre páginas. En uno de los recientes estudios sobre la web se estima que el grafo web contiene más de 8×10^8 nodos y se calcula que hay más de un billón de arcos. Experimentos realizados sobre una parte de la web correspondiente al dominio `.edu`, han revelado que en el grafo web la distancia esperada entre dos páginas es extremadamente corta. A pesar de los cientos de millones de páginas que forman la web, si escogemos como punto de partida una página web que sea *razonable*, podemos llegar a cualquier otra página, en un máximo de 19 clicks de ratón [2].

Evidentemente, hay que matizar el significado de *razonable* y una parte de los esfuerzos matemáticos para entender la web, se dirigen a precisar este concepto a fin de diseñar algoritmos eficientes para identificar páginas web razonables. Investigadores del grupo *clever*[4] en IBM Almadén, destacan dos tipos relevantes de nodos en la web: páginas **hub** y páginas **authority**. Las páginas hub son páginas que tienen muchos enlaces hacia otras páginas, es decir nodos con grado de salida muy alto. Las páginas authority se caracterizan por que tienen un alto grado de entrada, es decir muchas páginas

[1]Parcialemnte subvencionado por CICYT TIC1999-0754-C03

[2]Departament de Llenguatges i Sistemes, Universitat Politècnica Catalunya, Jordi Girona Salgado 1–3, 08034-Barcelona {alvarez,diaz,mjserna}@lsi.upc.es

[3]Uniforme Resource Location

[4]`http://www.almaden.ibm.com/cs/k53/clever.html`

enlazan con ellas. En el grafo web, los nodos hub tienen el papel de acortar distancias, como distribuidores de caminos, mientras que las páginas authority son páginas muy solicitadas por los usuarios de la web. El grupo *clever* ha diseñado un algoritmo para identificar los nodos hub y authority relevantes. También han diseñado un algoritmo para identificar comunidades en la web con temáticas comunes [6]. La característica más importante de ambos algoritmos es el uso de las propiedades de conectividad entre páginas. En el primero sólo se realiza una búsqueda por contenido para detectar un conjunto inicial de candidatos. El algoritmo refina este conjunto teniendo en cuenta las conexiones del subgrafo inducido por sus páginas. En el segundo algoritmo, se identifican comunidades virtuales de temática común con subgrafos del grafo web. Para obtenerlos, el algoritmo necesita conocer la topología del grafo web.

Aparte de tener distancia media pequeña, otra característica del grafo web es la de ser *esparso*. Un grafo dirigido con n nodos puede tener $n(n-1)$ arcos, habitualmente cuando un grafo tiene $O(n)$ arcos se dice que es esparso. También se puede observar que los nodos en la web tienden a apiñarse formando clusters [6], recordemos que un cluster es un conjunto de nodos altamente conectados entre sí, pero con conexiones esporádicas hacia el exterior. Estas tres propiedades del grafo web son compartidas por una familia de redes que simulan fenómenos sociales y biológicos, denominadas redes **small world** [8].

Vamos a presentar la evolución de algunas propuestas de modelos de grafo para las redes small world y en particular de aquellas que pueden tener utilidad para modelizar el grafo web. La primera idea fue utilizar el modelo clásico de grafo aleatorio. Dado un valor $p \in [0,1]$, un grafo aleatorio en el modelo $G_{n,p}$ se obtiene como un conjunto de n nodos, en el que cada uno de los posibles arcos se selecciona de manera independiente con probabilidad p. El estudio de grafos aleatorios iniciado por Erdös-Renyi, ha generado una teoría sólida, con multitud de resultados (ver por ej. [4]). En el caso de que p sea una constante y $p \geq 1/2$ estos grafos aleatorios tienen diámetro pequeño, de orden logarítmico en el número de nodos y son esparsos. Pero en contraposición con la web, sus nodos no tienden a agruparse en clusters. Después, de cara a incluir en el modelo la propiedad de formar clusters, se pensó en la utilización de mallas $n \times n$ que son esparsas y tienen la propiedad de agruparse formando clusters, cualquier subrectángulo de una malla es un cluster, pero su diámetro es demasiado grande, $\Theta(n)$ y hay demasiados pares de nodos que están a distancia $\Theta(n)$.

Watts y Strogatz desarrollaron un modelo intermedio [9]. Partiendo de un ciclo se conectan todos los nodos a distancia $\leq t$ y se añade una componente de aleatoridad en el diseño, de manera que, con probabilidad p, cada arista se redirige hacia otro nodo elegido uniformemente de entre el resto de nodos del grafo. Estas aristas servirán como atajos para reducir distancias. Notese que si $p = 0$ el grafo es el ciclo inicial, mientras que para

$p = 1$ el grafo se transforma en un grafo aleatorio en el modelo clásico (ver figura 1). El modelo presenta un problema estructural, ya que, para valores razonables de p, con alta probabilidad el grafo resultante no es conexo. La primera modificación consiste en añadir aristas en lugar de redirigirlas, la segunda en considerar otras estructuras de partida que sustituyen al ciclo, por ejemplo mallas bidimensionales.

Partiendo de esta idea, Jon Kleinberg [5], diseñó una modificación en donde a cada nodo de una malla $n \times n$ se le añaden primero conexiones a todos los nodos cercanos (a distancia $\leq t$) y conexiones aleatorias a nodos lejanos (a distancia $> t$), de manera que la distancia interviene en la probabilidad. En concreto se añade una conexión a un nodo a distancia d con probabilidad proporcional a d^{-r}, donde r es un parámetro del modelo. Los grafos resultantes tienen diámetro comparable con el de los grafos aleatorios. Pero no tienen la propiedad cluster, y ademas en el caso particular de la web, esta no tiene la regularidad de la malla. Una característica favorable de este modelo, es la existencia para el caso particular de $r = 2$ y $t = 1$, de un algoritmo eficiente para encontrar un camino corto entre dos nodos cualesquiera. Además, Kleinberg demostró que para ningún otro valor de r existe un algoritmo que calcule un camino corto.

Albert, Barbási y Jeong [3, 2] descubrieron que la probabilidad del número de enlaces de entrada y salida de una página en la web sigue una *ley de potencias* del tipo $k^{-\gamma}$ en particular, la probabilidad de que una página web tenga k enlaces (grado de salida k) es proporcional a $k^{-2.45}$ mientras que la probabilidad de que esté apuntada desde k páginas (grado de entrada k) es proporcional a $k^{-2.1}$. Esta distribución del grado de entrada y salida garantiza la existencia de nodos con grado elevado, hecho radicalmente diferente en los modelos $G_{n,p}$ (distribución tipo Poisson) y Watts-Strogatz. Los mismos autores también demostraron que la distancia entre dos nodos sigue una distribución Gaussiana con esperanza $0.35 + 2.06 \log n$ en dode n es el número total de nodos. Cuando substituimos n por 8×10^8 obtenemos nuestro comentario previo que 19 clicks de ratón son suficientes para navegar entre dos páginas razonables en la web. Además, la dependencia logarítmica en n, indica que un incremento en el tamaño de la web, repercutirá en un incremento pequeño de la distancia media entre páginas. Por ejemplo un incremento del 1000% en el tamaño de la web incrementará de 19 a 21 la distancia media.

Estos descubrimientos llevaron a los autores a proponer un nuevo modelo que toma en consideración una nueva característica, la creación dinámica de nuevas páginas web. El modelo dinámico propuesto comienza con n nodos y sin arcos. A partir de este momento se incorporan, uno a uno, nuevos nodos, cada nuevo nodo se conecta a $m < n$ nodos existentes. La probabilidad de tener un arco de u a v es directamente proporcional al grado actual del nodo v (ver figura 2). Curiosamente, cuando el grafo se estabiliza, la probabilidad de tener un nodo con k enlaces es k^{-3}. Este modelo parece capturar el crecimiento dinámico de la web, sin embargo los valores empíricos $k^{-2.45}$ y $k^{-2.1}$ están relativamente lejos de k^{-3} y de momento no se ha encontrado una explicación de la discrepancia.

Este último modelo expuesto, aunque no explica todas las características del grafo web, toma en consideración su estructura dinámica. En la actualidad se continua trabajado con modelos más realistas (y complicados) que incluyan destrucción de páginas, etc. Una línea de trabajo esperanzadora, son los nuevos modelos de grafos aleatorios basados en la distribución del grado de los nodos [7]. Algunos trabajos recientes han seguido esta línea. Por ejemplo en un artículo reciente [1], se ha propuesto la utilización de una ley de potencias como distribución del número de nodos de un grado dado en un grafo de llamadas telefónicas. Su modelo $P(\alpha, \beta)$ asigna probabilidad uniforme a todos los grafos que tienen e^{α}/x^{β} nodos con grado x. Es pronto para determinar su posible utilidad como modelo de la web.

Referencias

[1] W. Aiello, F. Chung, y L. Lu. A random graph model for massive graphs. 32nd Annual ACM Symposium on Theory of Computing, 2000. `http://math.ucsd.edu/~fan/pap.html`

[2] R. Albert, H. Jeong, and A.-L. Barabási. The diameter on the world wide web. *Nature*, 401:130–131, 1999.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] B. Bollobás. *Random Graphs.* Academic Press, London, 1985.

[5] J. Kleinberg. The small-world phenomenon: an algorithmic prespective. 32nd Annual ACM Symposium on Theory of Computing, 2000. `http://www.cs.cornell.edu/home/kleinber/`.

[6] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagoplan, and A. Tomkins. The web as a graph: Measurements, models and methods. *Computing and Combinatorics*, volume 1627 of *Lecture Notes in Computer Science*, pages 1–17. Springer-Verlag, Berlin, 1999.

[7] B. McKay and N. Wormald. The degree sequence of a random graph I. the models. *Random Structures and Algorithms*, 11:97–117, 1997.

[8] D. Watts. *The dynamics of Networks between order and randomness.* Princeton University Press, Princeton, 1999.

[9] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

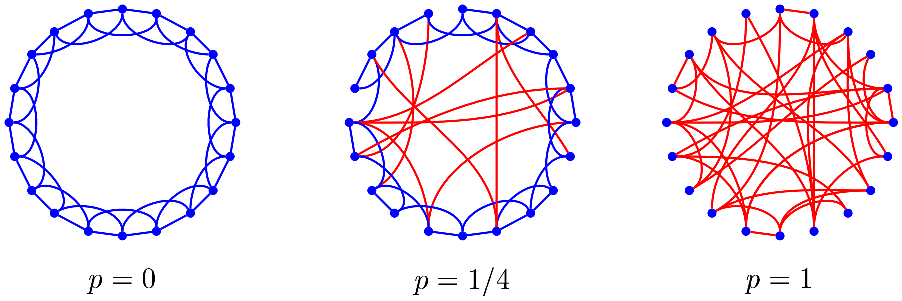

Figura 1: El modelo de Watts y Strogatz

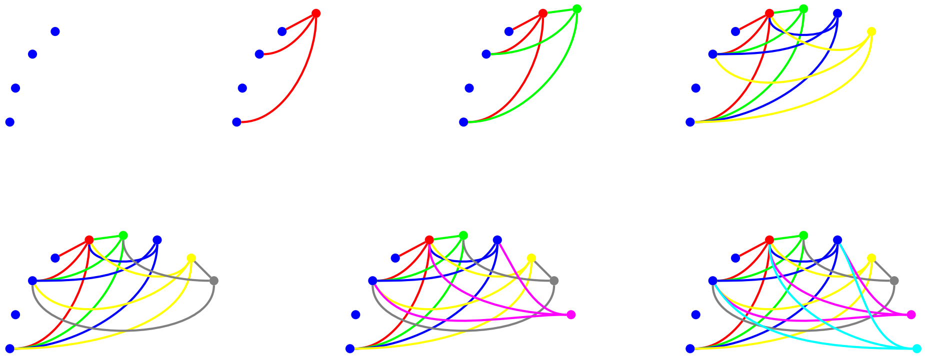

Figura 2: Un ejemplo de evolución del modelo de Albert, Barbási y Jeong con parámetros $n = 4$ y $m = 3$