MODELOS DE GRAFOS PARA LA WEB

Carme Álvarez, Josep Díaz, María Serna

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 (Uniforme Resource Location), 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 $8x10^8$ 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* 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 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 reali-