Resumen Tema 6: Minería de estructura

José Alberto Benítez Andrades

Marzo 2011

En este trabajo se resumen las conclusiones obtenidas después de haber realizado la lectura de los artículos propuestos Soumen Chakrabarti "Mining the Link Structure of the World Wide Web" , Ravi Kumar *"The Web as a Graph" y* Broder *"Graph Structure in the web"*.

1. Definición y objetivos de la minería de estructura de la web.

La World Wide Web contiene una cantidad enorme de información, pero puede ser extremadamente difícil para los usuarios el localizar recursos que sean de calidad y relevantes a las necesidades de información. Esto sucede porque la Web es un corpus de hipertexto de gran tamaño y continua creciendo exponencialmente. Pero la variación en las páginas es incluso peor que la escala de datos: el conjunto de páginas web no tienen una estructura web unificada, con variabilidad en el estilo de autores y contenido que es más grande que en las coleciones de documentos tradicionales. Este nivel de complejidad hace imposible aplicar técnicas de un gestor de base de datos y recuperadores de finromación.

Para mejorarlo se han desarrollado algoritmos que explotan la estructura de hipervínculos de la WWW para el descubrimiento de información y categorización, la construcción de listas de recursos de alta calidad y el análisis de las comunidades online enlazadas. Hay muchas maneras en que se puede utilizar la estructura de enlaces en la web para inferir cuáles son las páginas más importantes, y algunas son más efectivas que otras. La estructura de hipervínculos implica una estructura social subyacente en la manera de crear las páginas y los enlaces. El objetivo es desarrollar técnicas que se aprovechen de lo que observamos sobre la organización social intrínseca en la web mientras diseñamos algoritmos que minen información de los hiperenlaces.

La minería de estructura de la Web se definiría entonces como el proceso de usar la teoría de grafos para analizar los nodos y la estructura de conexión de un sitio web. De acuerdo con el tipo de dato estructural, la estructura de minería de la web puede ser dividida en dos tipos. El primer tipo consiste en extraer patrones de los hipervínculos en la web. Un hipervínculo es un componente estructural que conecta la página web a una localización diferente. El otro tipo es minar la estructura del documento. Consiste en usar la estructura en forma de árbol para analizar el XML o el HTML dentro de la página web.

2. Definición, modelado y uso de las nociones de:

2.1. Autoridad (authoritative page), prestigio

No se quieren sólo localizar un conjunto de páginas relevantes sino que se quieren las páginas relevantes de mayor calidad. Para limitar una búsqueda grande en Internet hasta un tamaño sensato

para un observador humano, se necesitan medios que identifiquen las páginas más "definitivas" o "Autoridad". Típicamente, la creación de un enlace por el autor de una página web representa un tipo implícito de "aprobación", de la página a la que se apunta. Recolectando el juicio colectivo en el conjunto de tales "aprobaciones", se puede obtener una comprensión más profunda tanto de la relevancia como de la calidad de los contenidos de la web. Un problema que surge es que las autoridades no suelen ser particularmente auto-descriptivas. –Por ejemplo no hay razón para encontrar "Fabricantes japoneses de automóviles" en la página de Toyota u Honda. Esta dificultad ilustra los problemas que hay en confiar sólo en el texto mientras buscamos Autoridades. Por eso es interesante utilizar la información de los enlaces. Pero también hay dificultad en usar la información de los hiperenelaces. Pese a que muchos enlaces representan el tipo de aprobación que se decribía anteriormente, otros se crean por razones que no tienen nada que ver con el otorgamiento de autoridad.

Estas consideraciones indican algunas de las dificultades con las que nos encontramos al buscar páginas autoridades. Hay dificultades en la información de los hipervínculos. Mientras muchos enlaces representan el tipo de respaldo que discutimos, otros son creados para razones que no tienen nada que ver con el contenido de la web. Algunos enlaces existen sólo para propósitos navegacionales ("Haga clic para regresar al menú raíz") o como anuncios pagados("Las vacaciones de sus sueños están a un solo clic de distancia") Lo que se espera es que en el sentido agregado, sobre un número lo bastante grande de vínculos, nuestra perspectiva de vínculos como "Dadores de autoridad" se mantendrá.

¿ Cómo podemos realizar el mejor modelado en que una autoridad es conferida en la web? Como expliqué anteriormente, las páginas web autorizadas no suelen ser muy auto-descriptivas; este caso se repite también en las autoridades en los temas generales que frecuentemente no enlazan directamente a otro. Está bastante claro el por qué esto debería ser cierto para cualquier tema con un aspecto comercial o de competencia; AltaVista, Excite e InfoSeek pueden todas ser autorizadas para el tópico "motores de búsqueda", pero ellos no tienen interés en enlazarse entre ellos porque son competencia.

¿Cómo determinar que una página es una Autoridad? Podríamos decirlo porque un número de páginas relativamente anónimas que son claramente relevantes a, por ejemplo, "Motores de Búsqueda" tienen enlaces a Google, AltaVista, Excite e Infoseek. Tales páginas son un componente recurrente de la web: "Hubs" o concentradores que enlazan a una colección de sitios prominentes en un tema común. Pueden aparecer en una variedad de formas, desde listas de recursos profesionales, hasta listas de vínculos recomendados en páginas web individuales. Los concentradores no necesitan ser prominentes, o siquiera tener enlaces apuntándoles. Su característica distintiva es que son potentes dadores de autoridad en un tema concreto. De esta manera, tienen un papel que es dual al de las autoridades: Una buena Autoridad es aquella que es apuntada por muchos buenos concentradores. . Esta relación mutuamente reforzada entre los hubs y las autoridades servirán como tema central en nuestra exploración de métodos basados en enlaces para la búsqueda, la compilación automatizada de recursos web de alta calidad, y el descubrimiento de comunidades web temáticamente cohesionadas.

2.2. Centralidad

Uno de los resultados importantes del análisis de hipervínculo de una red es la identificación de un nodo central, o en este caso, un sitio web central, generalmente definido como el sitio que proporciona la mayor parte de las conexiones y/o las conexiones más cortas a otros miembros del grupo(Scott, 1991; Wasserman & Faust, 1994). El sitio web central usualmente juega el papel de concentrador, Autoridad o sitio de prestigio. Existen varias medidas de centralidad.

El autovector de centralidad de Bonacci's se usa a menudo como un indicador global en el análisis de hipervínculo de red. Es apropiado en aquellos casos donde la red esté simétricamente interconectada

y las frecuencias de las conexiones de los vínculos entre sitios web no sean binarias. Y que sean relativamente densas (Bonacich & Lloyd, 2001). Sin embargo, esta métrica proporciona una descripción inadecuada de una red direccional(O asimétrica) Como resultado, los enlaces direccionales pueden ser analizados usando el grado de centralidad de Freeman. Mide el número de conexiones directas de hipervínculos de un sitio web con otros en el grupo (Freeman, 1979).

La métrica de Freeman consiste en grado de centralidad entrante y saliente. El grado de centralidad entrante se calcula basándose en el número de enlaces que un sitio web recibe de otro sitio web, mientras que el saliente se determina con el número de vínculos que se originan en un sitio. Además de estos valores están las métricas de proximidad (closeness) e interposición (betweeness).

La métrica de centralidad se utiliza para determinar qué sitio Web tiene el camino más corto a todos los otros en el grupo. La métrica de centralidad de interposición se refiere a la frecuencia con con la que un sitio web se encuentra entre pares de otros sitios en el grupo y representa el potencial para el control de comunicación, como un portero.(Freeman, 1979).

Finalmente, la centraldiad de Negopy de Richards es el número medio de vínculos requeridos para alcanzar a cada uno de los otros sitios web en el grupo, de tal forma que cuanto más bajo sea el valor. El sitio sera más central (Richards, 1995). La mayor parte de los WebSites agrupados, tales como los sitios web de un departamento de una Universidad, están conectados al sitio central de tal forma que los usuarios de Internet puedan navegar con pocos enlaces cuando están en uno de esos sitios.

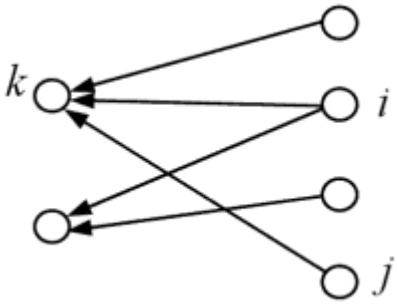
2.3. Co-cita

El análisis de co-cita ha sido usado para mapear la relación temática de conjuntos de autores, diarios o artículos. Ver [White & McCain1989] para una revisión de estas técnicas). Puede mostrar un agrupamiento significativo de autores relacionados por temas. [White & Griffith1981].

Grafo de co-cita:

Nodos = Páginas web·Aristas no dirigidas(Ponderadas)= co-cita(ponderada)

La co-citación es utilizada para medir la similitud de dos documentos. Si los papeles i y j son citados por un papel k, entonces ellos deben ser dichos para ser relacionados en algunos sentidos a otro, incluso ellos no se citan directamente el uno al otro. En la figura siguiente se muestra que los papeles i y j están co-citados por el papel k. Si el papel i y el papel j están citados juntos por otros papeles, significa que i y j tienen una relación o son similares.



Dejar L ser la matriz de citación. Cada celda de la matriz es definida como : $L_{ij} = 1$ si el papel i cita al papel j, y $L = 0$ si es de otra manera. La Co-citación (denotada por C_{ij}) es una medida similar definidia como el número de papeles que co-citan i y j, y es computado con:

$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$

donde n es el total de números de páginas. C_{ii} es naturalmente el número de papeles que citan i. Una matriz cuadrada C puede ser formada con C_{ij}, y es llamada matriz de citación. La co-citación es simétrica, $C_{ij} = C_{ji}$, y es comunmente usado como medidas de similitud de dos papeles en clustering a un grupo de papeles de tópicos similares juntos.

3. Ranking de páginas web basado en enlaces:

3.1. PageRank

El año 1998 fue un año importante para el análisis de enlace web y la búsqueda web. Se crearon los algoritmos de PageRank y HITS. HITS fue representado por Jon Kleingberg en Enero de 1998 en el *Noveno Simposium en algoritmos discretos*. PageRank fue presentado por Sergey Brin y Larry Page en la *Séptima conferencia de World Wide Web*. Basado en el algoritmo, ellos crearon un motor de búsqueda web llamado Google. Las ideas principales de PageRank y HITS son realmente similares. Sin embargo, existen algunas diferencias que veremos más adelante. PageRank ha emergido como el modelo de análisis de enlace dominante para los motores de búsqueda, particularmente debido a la evaluación de consultas independientes de las páginas web, para evitar el spamming.

PageRank se basa en la naturaleza democrática de la web utilizando su estructura de enlaces como indicador de la calidad individual de cada página. PageRank interperta hiperenlaces de una página x a una página y como un voto, de la página x a la página y. Sin embargo, PageRank busca un número de votos legal, o enlaces que reciba una página. También analiza las páginas que emiten el voto. Los votos emitidos por las páginas que son de importante peso ayudan a las otras páginas a tener mejor puntuación. Esta es la idea del ranking de prestigio en las redes sociales.

Algoritmo de PageRank

PageRank es un ranking estático de páginas web en el sentido de que un valor de PageRank es computado por cada página off-line y no depende de las consultas de búsqueda. Desde que PageRank se basa en medidas de prestigio en redes coaiels, el valor de PageRank de cada página puede ser recordado como ese prestigio. Los principales conceptos en los contextos web son los siguientes:

In-Links de la página i: Son los hiperenlaces que apuntan a una página i de otras páginas. Usualmente, los hiperenlaces de un mismo sitio no son considerados.

Out-Links de la página i: Son los hiperenlaces que apuntan a otras páginas desde una página i. Usualmente, los hiperenlaces de un mismo sitio no son considerados.

Desde otra perspectiva del prestigio, existe la siguiente derivación del algoritmo de PageRank:

1. Un hiperenlace de una página apuntando a otra página es un transporte de autoridad a la página de destino. Así, los enlaces de entrada que recibe una página i, dan más prestigio de lo que la página i tenía antes de ello.

2. Las páginas que apuntan a una página i también tienen sus puntuaciones de prestigio en el poder. Una página con un prestigio alto que apunta a una página i es más importante que na página con poco prestigio que apunta a i. En otras palabras, una página es imporatnte si es apuntada por otras páginas que son importantes.

Acorde con el ranking de prestigio en las redes sociales, la importancia de las páginas i, es determinado resumiendo las puntuaciones de PR de todas las páginas que apuntan a i. Una página que apunta a

muchas otras páginas, reparte el prestigio que tiene sobre todas las páginas a las que apunta. Notando la diferencia del ranking de prestigio, donde la puntuación no se comparte.

Para formular las ideas explicadas anteriormetne, tratamos la Web como un grafo directo *G = (V,E)* donde *V* es un conjunto de vértices o nodos, por ejemplo, el conunto de todas las páginas y *E* es el conjunto de enlaces directos en el grafo, por ejemplo, los hiperenlaces. Sabiendo que el número de páginas totales en la Web es de *n (n = |V|)*. El PageRank de la página *i (P(i))* es definido por:

$P(i) = \sum_{(j,i)\in R} \frac{P(j)}{O_j}$

donde O_j es el número de enlaces salientes de la página *j*. Matemáticamente, usamos un sistema de n ecuaciones lineales con n incógnitas. Podemos usar una matriz que represente todas las ecuaciones. Sabiendo que P es una columna de vectores n-dimensionales del valor del PageRank, por ejemplo :

$P = (P(1), P(2), ..., P(n))^r$

Dado A como la matriz de adyacencia de nuestro grafo con:

$$A_{ij} \begin{cases} \frac{1}{O_i} & if(i,j) \in E \\ 0 & enotrocaso \end{cases}$$

Nosotros podemos escribir el sistema de n ecuaciones con

$$P = A^T P$$

Esta es la ecucación característica del propio sistema, donde la solcuión para P es el *vector propio* con el correspondiente *valor propio* de 1. Se soluciona con un algoritmo iterativo. 1 es el mayor valor y el PageRank es el vector P que es el principal vector propio.

El problema que surge es que la ecucación no es suficiente porque el grafo web no cumple las condiciones. Para introducir estas condiciones nos vamoas en *la cadena de Markov*.

En esta cadena, cada página web o nodo en el grafo es recordado como un estado. Un hiperenlace es una transición que lleva de un estado a otro con una probabilidad. Esto modela una navegación web aleatoria, navegando por la red como un estado de transición. Cada probabilidad de transición es $1/O_i$, si nosotross asumismos al navegante cuando clicka en los hiperenlaces en la página i uniformemente de forma aleatoria. Dada una matriz de probabilidad de transición de estados A, una matriz cuadrada de la siguiente manera:

$$A = \begin{matrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ A_{n1} & A_{n2} & . & . & . & A_{nn} \end{matrix}$$

A_{ij}representa la probabilidad de transición de que el usuario que esta en internet en el estado i se mueva al estado j (vaya de la página i a la página j).

Dada un vector de distribución de probabilidad inicial $P_0 = (p_0(1), p_0(2), ..., p_0(n))^T$ y una matriz de probabilidad de transición A, nosotros tenemos:

$$\sum_{i=1}^{n} p_{0(i)=1}$$

$$\sum_{ij=1}^{n} A_{ij=1}$$

La última ecuación no es totalmente cierta para algunas páginas web porque no tienen enlaces de salida. Si la matriz A satisface esa ecuación, nosotros decimos que A es la matriz estocástica de la cadena de Markov.

Nosotros podemos determinar la prboablidad de que el sistema este ne el estado j después del primer paso usando el siguiente razonamiento:

$$\mathrm{p}_i(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$

donde $A_{ij}(1)$ es la probablidad de ir de i a j después de la primera transición.

Realizando una serie de matrices sucesivas que no voy a escribir en este documento porque me extendería demasiado en este trabajo con demasiadas fórmulas matemáticas, surge un problema con la mariz A que provoca que no sea estocástica porque hay una fila entera de 0 en esta matriz.

El problema se puede solventar de dos maneras distinas:

1. Borrando las páginas que no tienen enlaces desde el sistema durante la computación del PageRank ya que esas páginas no afectan en el ranking de otra página de forma directa. Los enlaces de salida de otras páginas apuntando a esas páginas también serán borrados. Después de que el PageRank se haya compuado, estas páginas y los hiperenlaces que apuntan a estas, pueden ser añadidos. Su PageRank es fácil de calcular utilizando estas matrices. Las probabilidades de transición de estas páginas con enlaces borrados puede afectar pero no significativamente.

2. Otro método es añadiendo un conjunto completo de enlaces de salida de cada página como la página i a todas las páginas en internet. Asi, la probabilidad de transición de ir de una página i a cada página es $1/n$ asumiendo una probabilidad uniforme. Reemplazaremos la fila de 0's por e/n donde e es el fector n-dimensional de todas las 1's.

Un grafo directo G=(V,E) esta directamente relacionado si y solo si, para cada par de nodos $u, v \in V$, hay una ruta de u a v.

Un estado i es periódico con $k > 1$ si k es el número más pequeño de todas las rutas que hay desde el estadio anterior a i hasta i teniendo la longitud que es k. Si el estado no es periódico, es aperiódico.

Así, el valor del PageRank de las páginas web puede ser creado usando un método de iteración, que produce el principal vector propio con el valor propio de 1. El algoritmo es simple. Uno puede comenzar con un valor de asignación de PageRank. La iteración termina cuando el valor de PageRank no cambia mucho. El algoritmo sería el siguiente:

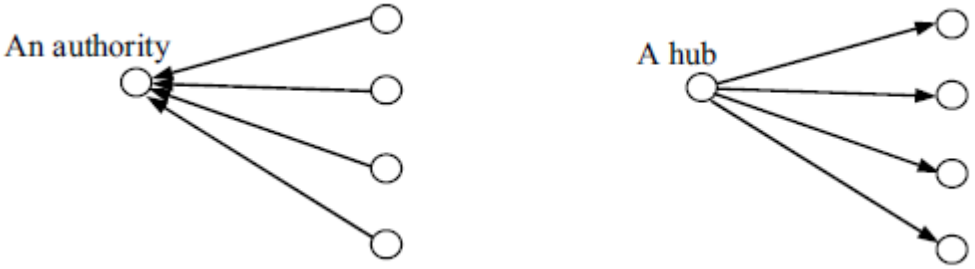
PageRank-Iterate(G)
....$P_0 \leftarrow e/n$
....$k \leftarrow 1$
....**repeat**
.....$P_k \leftarrow (1-d)e + dA^T P_{k-1}$
.....$k \leftarrow k+1$
....**until** $||P_k - P_{k-1}||_1 < \varepsilon$
....return P_k

Además de esto, hay una serie de variaciones en el algoritmo de PageRank que ha ido mutando debido al paso de los años y a los distintos avances en la tecnología web.

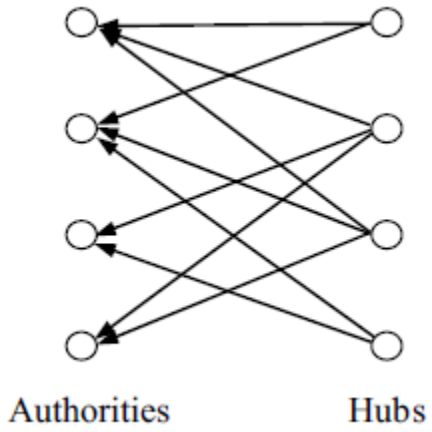
3.2. HITS

HITS es sinónimo de Hypertext Induced Topic Search. A diferencia de PageRank, que es un algoritmo de clasificación estática, HITS depende de la consulta de búsqueda. Cuando el usuario hace una consulta de búsqueda, la lista de webs exitosas amplía la lista de páginas relevantes devuelta por un motor de búsqueda y, a continuación produce dos clasificaciones de conjunto de páginas expandidas, el rango de autoridad y la clasificación del cubo (hub ranking).

Una autoridad es una página con muchos enlaces. La idea es que la página puede tener un buen contenido o autoridad sobre algún tema y por lo tanto muchas personas confían en él y enlazan con él. Un hub es una página con muchos de los enlaces. La página sirve como un organizador de la información sobre un tema en particular y apunta a muchas páginas que son una buena fuente sobre el tema. Cuando un usuario llega a esta página hub, él / ella se pueden encontrar muchos enlaces útiles que lo llevaría a las páginas de buen contenido sobre el tema. En la siguiente figura se puede observar bien este hecho.



La idea clave de HITS es que un buen centro de los puntos que apunte a muchas autoridades buenas y una buena autoridad es apuntada por muchas webs que son importantes. Por lo tanto, las autoridades y los centros tienen una relación de refuerzo mutuo.



A continuación, explicaré en primer lugar el algoritmo HITS, así como establecer una conexión entre HITS y la co-cita y el acoplamiento en la investigación bibliográfica bibliométricos. A continuación, analizaré las fortalezas y debilidades de HITS, y plantearé algunas líneas posibles para hacer frente a sus debilidades.

Algoritmo HITS

Antes de describir el algoritmo HITS, primero vamos a describir cómo HITS recoge las páginas para proceder a su clasificación. Dada una consulta amplia *q*, HITS recoge un conjunto de páginas de la siguiente manera:

1. Envía la q consulta a un sistema de motor de búsqueda. A continuación, recoge t (t = 200 se utiliza en el documento de HITS) más páginas clasificadas, que suponemos que es de gran importancia para la consulta de búsqueda. Este conjunto se conoce como la raíz del conjunto *W*.

2. A continuación, crece *W* mediante la inclusión de cualquier página a la que apunta una página de *W* y cualquier página que apunta a una página en *W*. Esto le da a un conjunto más amplio llamado *S*. Sin embargo, este conjunto puede ser muy grande. El algoritmo limita su tamaño permitiendo a cada página en *W* para llevar a las páginas de la mayoría de k (k = 50 se utiliza en el documento de HITS) que apunta a ella en *S*. El conjunto *S* se llama el conjunto base.

HITS trabaja en las páginas de *S*, y asigna a cada página una calificación *S* autoridad y una puntuación de cubo. El número de páginas a estudiar es n. Nosotros utilizaremos G = (V, E), que indica el gráfo de enlace (dirección) de S. V es el conjunto de páginas (o nodos) y E es el conjunto de aristas dirigidas (o enlaces). Utilizaremos L para denotar la matriz de adyacencia del grafo.

$$L_{ij} \begin{cases} 1 & if(i,j) \in E \\ 0 & enotrocaso \end{cases}$$

Definimos que la puntuación de la autoridad de la página i sea a(i), y la puntuación centro de la página i es h(i). La relación de refuerzo mutuo de las dos puntuaciones se representa como sigue:

$$a(i) = \sum_{(j,i)\in E} h(j)$$

$$h(i) = \sum_{(i,j)\in E} a(j)$$

El cálculo de las puntuaciones de la autoridad y las puntuaciones de centro es básicamente el mismo que el cálculo de las puntuaciones PageRank utilizando el método de iteración de energía. Si usamos a_k y h_k para denotar las puntuaciones autoridad y centro de operaciones en la iteración k-ésima, los procesos iterativos para la generación de las soluciones finales son:

$$a_k = L^T L a_{k-1}$$

$$h_k = L^T L h_{k-1}$$

Debemos tener en cuenta que las ecuaciones anteriores no utilizan el cubo (o autoridad) de vectores, debido a las sustituciones en las ecuaciones anteriores a estas.

Después de cada iteración, los valores también están normalizadas (para mantener las pequeñas) para que:

HITS-Iterate(G)

....$a_0 \leftarrow h_0 \leftarrow (1, 1, ..., 1)$;

....$k \leftarrow 1$

....**repeat**

.....$a_k \leftarrow L^T L a_{k-1}$

.....$h_k \leftarrow L^T L h_{k-1}$

.....$a_k \leftarrow a_k / ||a_k||_1$; // normalización

.....$h_k \leftarrow h_k / ||h_k||_1$; // normalización

....**until** $||a_k - a_{k-1}||_1 < \varepsilon_a$ and $||h_k - h_{k-1}||_1 < \varepsilon_h$

....return a_k y h_k

El algoritmo de iteración de para HITS es el mostrado justo encima de este texto. La iteración termina después de la 1-las normas de los vectores residuales que son menores de algunos umbrales ε_a y ε_h . Por lo tanto, el algoritmo encuentra los vectores propios de "equilibrio", como en el PageRank. Las páginas con autoridad y grandes puntuaciones hub son mejores las autoridades y los centros, respectivamente. HITS seleccionará a un puesto superior a pocas páginas como las autoridades y los cubos, y las devolverá al usuario.

Aunque HITS siempre convergen, hay un problema con la singularidad de limitar (convergentes) la autoridad y los vectores de cubo. Se muestra que para ciertos tipos de gráficos, inicializaciones diferentes al método de alimentación un mandato diferente final y vectores centro. Algunos resultados pueden ser incompatibles o malos. Farahat dio varios ejemplos. El corazón del problema es que no se repiten dominantes (principales) valores propios (autovalores varios son los mismos y son valores propios dominantes), las cuales son causadas por el problema que $L^T L$ (LL^T , respectivamente) se reduce. La primera solución de PageRank tiene el mismo problema. Sin embargo, los inventores PageRank encontrado una manera de evitar el problema. Una modificación similar a PageRank se puede aplicar a los HITS.

4. Análisis de comunidades en la web

Internet da cobijo a un gran número de comunidades-Grupos de creadores de contenidos que comparten un interés común y que se manifiesta como un conjunto de páginas web. A pesar de que muchas comunidades son definidas explícitamente (grupos de noticias, colecciones de recursos en portales, etc), muchos más son implícitos. Identificar estas comunidades no ayuda sólo a entender la evolución social e intelectual de la web, sino también en proporcionar información detallada a un conjunto de gente con ciertos intereses en el punto de mira. Debido a su número astronómico, naturaleza embrionaria, y flujo evolucionario, es difícil encontrar y analizar estas comunidades usando únicamente esfuerzo manual. Un enfoque puede consistir en tratar la web como un enorme grafo dirigido, usamos la estructura de grafo derivada del patrón de enlaces básico Autoridad-Concentrador como la "firma" de una comunidad y sistemáticamente escaneamos el grafo de la web para localizar tales estructuras.

Tal enfoque se basa en que las comunidades web temáticamente cohesivas contienen en su núcleo un denso patrón de enlaces de Concentradores a Autoridades. Esto enlaza las páginas juntas en la estructura de vínculos, pese al hecho de que los concentradores no necesariamente enlazan a concentradores ni las Autoridades enlazan necesariamente a Autoridades. Para poner esto en un lenguaje más orientado a la teoría de grafos: Se usa la noción de grafo bipartito dirigido –uno cuyos nodos puedan ser particionados en dos conjuntos A y B tales que cada enlace en el grafo se dirige del nodo

A al nodo B. Dado que las comunidades que se buscan contienen grafos bipartitos dirigidos con una gran densidad de aristas, se espera que muchos de ellos contengan subgrafos más pequeños que sean de hecho completos: Cada nodo en A tiene un enlace a cada nodo en B. Usando algoritmos de "poda"(pruning") se pueden enumerar todos los subgrafos de la Web en un PC Standard en unos tres días de ejecución. Este proceso, descrito en la referencia [1] dio como resultado en torno a 130.000 grafos bipartitos completos en la web en la cual 3 páginas web apuntaban al mismo conjunto de otras tres páginas web. Se realizó una inspección manual de unos una muestra aleatoria de unos 400 comunidades sugeridas y menos del 5 % de tales supuestas comunidades le faltaban un tema unificador o "Topic". En la fecha en que se llevó a cabo el experimento, el 25 % no estaban representadas en Yahoo. Y los que aparecían en Yahoo muchos aparecían en el sexto nivel del árbol de temas de Yahoo.

5. Otras aplicaciones de la minería de estructura

El objetivo principal de la minería de estructura es extraer previamente relaciones desconocidas entre las páginas Web. Esta estructura de minería de datos permite el uso de una empresa para vincular la información de su propio sitio Web para permitir la información de navegación y de grupo en los mapas de sitio. Esto permite a sus usuarios la posibilidad de acceder a la información deseada a través de la asociación de palabras clave y la minería de contenido. La jerarquía de hipervínculo se determina también mediante la ruta de la información dentro de los sitios relacionados a la relación de enlaces de la competencia y la conexión a través de motores de búsqueda y la tercera parte co-enlaces. Esto permite la agrupación de conectar las páginas web para establecer la relación de estas páginas. En la WWW, el uso de la minería de estructura permite la determinación de la estructura similar de páginas web de la agrupación a través de la identificación de la estructura subyacente. Esta información puede ser utilizada para proyectar las semejanzas de contenido web. Las similitudes se conoce, después, la capacidad para mantener o mejorar la información de un sitio para permitir el acceso de las arañas web, en una proporción superior. Cuanto mayor sea la cantidad de rastreadores Web, el más beneficioso para el sitio debido a su contenido relacionado con las búsquedas.

En el mundo de los negocios, la mineria de estructura puede ser muy útil para determinar la conexión entre dos o más sitios Web de negocios. La conexión determinada da a luz una herramienta útil para el mapeo de las empresas que compiten a través de enlaces de terceros, como distribuidores y clientes. Este mapa de agrupación de los contenidos de las páginas de negocios colocando sobre los resultados de búsqueda a través de la conexión de palabras clave y los vínculos de la relación de las páginas web. Esta información determinada proporcionará el camino adecuado a través de la minería de estructura para mejorar la navegación de estas páginas a través de sus relaciones y jerarquía de enlace de los sitios Web.

Con la mejora de la navegación de páginas Web en los sitios Web de negocios, que conecta la información solicitada para un motor de búsqueda sea más efectiva. Esta conexión más fuerte permite la generación de tráfico a un sitio de negocios para proporcionar resultados que sean más productivos. Los enlaces, siempre están dentro de la relación de las páginas web que permiten la navegación para obtener la jerarquía de enlace que permitan la navegación fácil. Esta mejora de la navegación atrae a los robots a las ubicaciones correctas proporcionar la información solicitada, demostrando ser más beneficiosa en los clics a un sitio en particular.

Por lo tanto, la minería web y el uso de la minería de estructura pueden proporcionar resultados estratégicos para la comercialización de un sitio Web para la producción de la venta. El tráfico más dirigido a las páginas web de un sitio en particular aumenta el nivel de visitas regresar al sitio y recordar los motores de búsqueda sobre la información o producto proporcionado por la empresa. Esto también

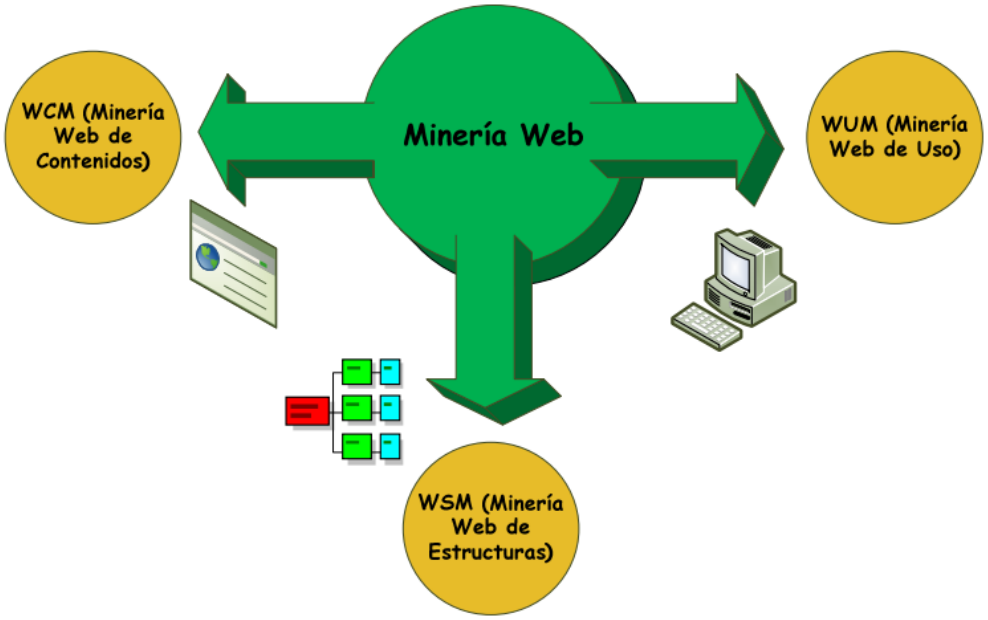
permite a las estrategias de marketing proporcionar resultados que sean más productivos a través de navegación de las páginas con enlaces a la página principal del sitio en sí. Para realmente aprovechar su sitio web como una herramienta de negocio en la Web la minería de estructura se convierte en una necesidad.

6. Áreas de investigación relacionadas

Como bien hemos ido viendo a lo largo de todo este curso, la Minería de Estructura se encuentra dentro del término global de Minería de la Web, que está compuesto por tres tipos de minería:

- Minería Web de Contenidos (WCM, Web Content Mining).
- Minería Web de Estructura (WSM, Web Structure Mining).
- Minería Web de Uso (WUM, Web Usage Mining).

WCM clasifica los documentos automáticamente o construye una base de información web multicapa. WSM extrae la estructura de una página web; WUM descubre patrones de acceso a las páginas en los usuarios. En la siguiente figura se ilustran las categorías de la Minería Web.



En resumen WSM extrae la estructura de los hiperenlaces, es decir, cómo están los documentos estructurados respecto a los otros (estructura interdocumental a diferencia de la estructura intradocumental de WCM). La estructura se representa como un grafo de los enlaces en un sitio web o entre sitios web. WSM revela más información que la información contenida en los documentos: por ejemplo, los enlaces que apuntan a un documento pueden indicar la popularidad o importancia de un documento, mientras que los enlaces salientes indican la riqueza o variedad de los temas que contiene. Esto nos lleva a una organización jerárquica por temas que puede ser inferida directamente de los patrones

de enlazado. Es posible incluso no especificar los documentos mediante palabras clave, sino mediante documentos ejemplares.

Un concepto íntimamente relacionado con la WSM por la importancia de la estructura es el de web semántica.

7. Conferencias internacionales

Algunas de las conferencias internacionales que abordan el tema de la minería de la web y que concretamente tratan el tema de la minería web de estructura son las siguientes:

- International World Wide Web Conference(IW3C2).
- Asia-Pacific Web Conference (APWeb 2008)
- International Conference on Internet and Web Engineering
- International Workshop on Web Mining for E-commerce and E-services (WMEE 2008) I
- International Conference on Web Intelligence, Mining and Semantics
- International Conference on Web-based Learning (ICWL 2010)
- Web Information Systems Engineering (WISE 2008)
- Call for book chapter: Web Mining Applications in E-commerce and E-services (Abril 2008)
- International Conference on Web Information Systems and Mining (WISM2010)

Referencias

[1] M. Abulaish, y L. Dey. "Biological Ontology Enhancement with Fuzzy Relations: A Text-Mining Framework". Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, France, pp. 379-385. 2005.

[2] A.O. Ajayi. G.A. Aderounmu y H.A. Soriyan. "An adaptive fuzzy information retrieval model to improve response time perceived by e-commerce clients". Expert Systems with Applications (ESWA) Vol. 37(1), pp. 82-91, 2010.

[3] S. Alag. "Collective Intelligence in Action". Manning Pubn, online ed., September 2008.

[4] A. An, J. Stefanowski, S. Ramanna y C. Butz. "Rough sets, fuzzy sets, data mining and granular computing". 11th International Conference, RSFDGrC 2007, Toronto, Canada, May 14-16, 2007.

[5] J. Barbancho. "Prescripciones técnicas para el diseño y construcción de apoyo al SOS de la Universidad de Sevilla". Departamento de Tecnología Electrónica. Universidad de Sevilla. Technical Report 0403-29. 2009.

[6] T. Berners-Lee y E. Miller. "The Semantic Web lifts off". ERCIM News No. 51, October 2002. Special Semantic Web.

[7] A. Bookstein, S.T. Klein y T. Raita, "Detecting content bearing words by serial clustering". SIGIR Forum (ACM Special Interest Group on Information Retrieval), p. 319-327, 1995.

[8] R. Burget, "Information Extraction from HTML Document Based on Logical Document Structure". Tesis Doctoral. Universidad de Brno, 2004.

[9] S. Chakrabarti, B. Dom, R. Agrawal y P. Raghavan, "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies". VLDB Journal Vol.7 3, p. 163-178, 1998.

[10] S. Chakrabarti, "Data Mining for hypertext: a tutorial survey". ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining, 2000.

[11] R. Cooley, B. Mobasher y J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9 th IEEE International Conference on Tools with Artificial Intelligence. ICTAI'97, 1997.

[12] D. de Kool y J. van Wamelen, "Web 2.0: A New Basis for E-Government" 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA 2008., vol. 1 pp.1-7, 7-11 April 2008.

[13] C. Ding y X. He. "K-means Clustering via Principal Component Analysis". Proceedings of the International Conference in Machine Learning, ICML04, pp 225-232. July 2004

[14] E. M. Eisman, V.Lopez y J.L. Castro, "Controlling the emotional state of an embodied conversationalagent with a dynamic probabilistic fuzzy rules based system". Expert Systems with Applications, Vol. 36, Issue 6, pp. 9698-9708, August 2009

[15] M. Friedman, M. Last, O. Zaafrany, M. Schneider y A. Kandel, "A new approach for fuzzy clustering of Web documents". Proceedings of the IEEE International Conference on Fuzzy Systems, vol.1, pp. 377-381, 25-29 July 2004.

[16] A. Gómez, J. Ropero y C. León. "A fuzzy logic system for classifying the contents of a database and searching consultations in natural language". Proceedings of the Mediterranean Electrotechnical Conference - MELECON 2006, pp. 721-724, 2006.

[17] N. Guarino y P. Giaretta. "Ontologies and Knowledge Bases: Towards a Terminological Clarification". Towards Very Large Knowledge Bases: Knowledge Building and Knowledge sharing, N. Mars (ed.) IOS Press, Amsterdam, 1995, pp. 25-32.

[18] J. Y. Hardeberg. "Acquisition and Reproduction of Color Images: Colorimetric and Multispectral Approaches". Universal-Publishers.com. 2001.

[19] Y. J. Horng, S. M. Chen, Y. C. Chang, C. H. Lee. "A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques". IEEE T. Fuzzy Systems 13, vol. 2, pp. 216-228. 2005.

[20] W. Klogsen y J. Zytkow, "Handbook of data mining and knowledge discovery"". New York: Oxford University Press. 2002