ARTÍCULO

Un modelo de minería de consultas para el diseño del contenido y la estructura de un sitio Web

Ricardo Baeza-Yates, Bárbara Poblete

Universitat Pompeu Fabra
Barcelona, España
& Centro de Investigación de la Web,
DCC, Universidad de Chile
Santiago, Chile
{ricardo.baeza,barbara.poblete}@upf.edu

Resumen

En este trabajo presentamos un modelo para hacer minería de consultas en sitios Web. Este modelo relaciona la información aportada por las consultas encontradas en un sitio, con los datos de uso, contenido y estructura de éste. El principal objetivo de nuestro modelo es descubrir en forma simple, información valiosa acerca de cómo mejorar la estructura y contenido del sitio, permitiendo que éste sea más intuitivo y adecuado a las necesidades de sus usuarios. Este modelo propone el análisis de los diferentes tipos de consultas registradas en las bitácoras o *logs* de uso de un sitio Web, tales como las consultas formuladas por los usuarios en el motor de búsqueda interno del sitio y las consultas realizadas en buscadores externos, que condujeron hacia documentos en el sitio. Estas consultas proveen información útil acerca de los temas que interesan a los usuarios que visitan un sitio Web, y los patrones de navegación asociados a estas consultas indican si los documentos encontrados en el sitio fueron satisfactorios para las necesidades de los usuarios en ese momento. Este modelo además propone una validación visual de la organización jerárquica del sitio, dada por los enlaces entre documentos y sus contenidos, además de su relación con las consultas.

Palabras clave: Minería de datos, Minería Web, Mejoramiento de sitios Web, Análisis de consultas.

1. Introducción

La Web se caracteriza por su acelerado crecimiento, su uso cada vez más masivo y ser un medio facilitador de nuevos negocios. Esto ha generado la necesidad por parte de los proveedores de sitios Web de hacer sitios más intuitivos y accesibles para los usuarios. Es de vital importancia que los interesados en los contenidos de un sitio, sean capaces de encontrarlo en la Web y a su vez recorrerlo de una forma que les resulte cómoda y fácil. El no tomar conciencia de este hecho puede significar la pérdida de muchos clientes.

Los motores de búsqueda surgen como la aplicación por excelencia en este nuevo medio, ya que se han convertido en la herramienta más utilizada para alcanzar sitios en la red y conducir a los usuarios hacia la información que ellos buscan en un momento determinado.

Al navegar a través de la Web los usuarios dejan registro de todas sus acciones, principalmente en las bitácoras o *logs* de acceso de los servidores de los sitios y buscadores que visitan. En los logs de un sitio Web, en particular, se pueden encontrar las consultas formuladas en el buscador interno del sitio (si es que existe uno) y las consultas realizadas en buscadores externos que produjeron visitas al sitio. Esta información es sumamente valiosa, ya que puede entregar la clave hacia los gustos e intereses de los usuarios en general y dentro de los diferentes sitios que estos recorren.

En este trabajo presentamos un modelo de minería de uso, contenido y estructura en un sitio Web, enfocado en consultas, con el objetivo de

descubrir información novedosa e interesante referente a nuevas formas en que el sitio puede ser mejorado. Nuestro modelo además genera una visualización de la distribución de los contenidos en relación a la organización de los enlaces entre documentos, así como también de las URLs seleccionadas como resultados de consultas. El resultado de este análisis consiste de diversos informes desde los cuales se pueden inferir mejoras al sitio.

A partir de los informes generados se puede: obtener nuevos contenidos para ampliar la cobertura de ciertos temas, cambiar o agregar palabras a las descripciones de los enlaces, agregar nuevos enlaces entre documentos similares y revisar los enlaces existentes entre documentos muy diferentes.

Hemos aplicado este modelo en diferentes tipos de sitios, desde sitios pequeños a sitios grandes, medido en número de documentos y cantidad de visitas. En sitios grandes resulta especialmente útil, dada la dificultad que presentan en el manejo de sus contenidos. En todos los casos planteados, nuestro modelo ayuda a visualizar posibles "puntos conflictivos" en el sitio y formas de mejorar su organización.

Este trabajo se organiza de la siguiente forma. La sección 2 presenta el trabajo relacionado. En la sección 3 presentamos los conceptos que se utilizarán en el resto de este trabajo. La sección 4 describe nuestro modelo, mientras que la sección 5 resume nuestro prototipo y algunos resultados obtenidos con éste. En la última sección entregamos nuestras conclusiones y discutimos el trabajo futuro.

2. Trabajo relacionado

La *minería Web* [9] es el proceso de descubrir diferentes patrones y relaciones en un conjunto de datos de la Web. La minería Web puede dividirse en tres áreas principales: *minería de contenido*, *minería de estructura* y *minería de uso*. En [26] se plantea que la clave para descubrir nuevos conocimientos en una base de datos es obtener un conjunto de datos apropiado para realizar minería de datos. En el caso de la minería Web los datos pueden ser obtenidos desde el lado del servidor, del cliente, de los servidores *proxy* o de la base de datos corporativa de la entidad a la cual pertenece el sitio. Desde este punto de vista, los datos encontrados en un sitio Web en particular, pueden ser clasificados en tres tipos:

Contenido: Son los datos reales que se entregan a los usuarios. Es decir, los datos que almacenan los sitios Web, los cuales consisten generalmente de texto e imágenes u otros medios. Este tipo de dato es el más importante y difícil de procesar, por ser multimedial, aunque consiste principalmente de texto.

Estructura: Son los datos que describen la organización del contenido de un sitio. Esto incluye, la organización dentro de una página, la distribución de los enlaces tanto internos al sitio como externos y la jerarquía de todo el sitio.

Uso: Son aquellos datos que describen el uso al cual se ve sometido un sitio, registrado en los logs de acceso de los servidores Web.

En el área de la minería de uso en sitios Web existe un alto interés comercial, lo cual es mencionado en [3, 11]. Al analizar los logs de acceso de servidores Web se puede determinar si una página es altamente visitada o no. A partir de esta información se podría concluir, por ejemplo, que un documento que nunca es visitado no tiene razón de ser, o por el contrario, si una página muy visitada no se encuentra en los primeros niveles de jerarquía de un sitio, esto sería un indicador para mejorar la organización y navegación del mismo.

Existe una extensa lista de trabajos previos que utilizan la minería Web para mejorar sitios, los que incluyen: el análisis de patrones frecuentes de navegación y reglas de asociación basadas en las páginas visitadas por los usuarios [23, 7, 6] y modelamiento de sesiones de usuarios, perfiles y análisis de *clusters* [13, 16, 17, 19, 18]. Además, son numerosas las herramientas de minería de uso y de sistemas de recomendación. WebSIFT [8] es una herramienta de minería Web creada para encontrar reglas y patrones interesantes en un sitio, definiendo como "interesantes" las reglas y patrones que son nuevos o inesperados. Otras herramientas dedicadas al mejoramiento de sitios [24, 15, 25] están enfocadas en la navegación (y a veces la estructura) de un sitio en forma dinámica e individual para cada visitante.

Las consultas formuladas en los motores de búsqueda son una herramienta valiosa para mejorar sitios Web. En [12] se propone un método para analizar consultas similares en buscadores. En [5, 4] se presenta un modelo vectorial para la recomendación de consultas. Otro tipo de análisis, presentado en [2] consiste en estudiar las consultas formuladas en el motor de búsqueda interno de un sitio Web e indica que puede descubrirse información valiosa a partir del estudio del comportamiento de los usuarios después de realizada una consulta. Esta propuesta es el punto de partida de nuestro trabajo.

3. Conceptos preliminares

Sesión: Una sesión corresponde a todos los accesos registrados para un usuario, en los logs de uso de un sitio, dentro de un intervalo de tiempo máximo. Este intervalo es definido por defecto en 30 minutos, pero puede modificarse a cualquier otro valor adecuado para el sitio [10]. Cada usuario es identificado en forma única por su dirección `IP` y software de acceso o `User-Agent` (esto es una heurística que puede ser mejorada).

Motor de búsqueda: Un motor de búsqueda puede ser tanto interno como externo a un sitio Web. La principal diferencia entre un motor de búsqueda interno y uno externo, es que el motor de búsqueda interno sólo se remite a las páginas encontradas en un sitio Web, mientras que el externo en general recopila los documentos de la Web en forma global.

Consulta: Una consulta corresponde a un conjunto de una o más palabras claves, formuladas por un usuario en un motor de búsqueda y representan una necesidad de información de ese usuario (en la mayoría de los casos).

Esencia de la información: La esencia de la información [20] es un término que indica qué tan buena es una palabra (o un conjunto de palabras) para describir cierto concepto en relación a otras palabras con la misma semántica. Por ejemplo, las palabras polisémicas (palabras con más de un significado) tienen menor esencia de la información debido a su ambigüedad.

4. Descripción del modelo

En esta sección presentamos una descripción de nuestro modelo de minería para el uso, contenido y estructura de un sitio Web, enfocado en consultas. Los datos de entrada del modelo son los logs de uso, la estructura del sitio y sus contenidos. La estructura del sitio es obtenida de los enlaces entre documentos y los contenidos corresponden al texto asociado a cada una de sus páginas. El objetivo de este modelo es generar información que permitirá mejorar la estructura y los contenidos de un sitio Web, además de evaluar la interconectividad entre documentos de contenidos similares. En un sitio Web se pueden encontrar dos tipos de consultas diferentes, las cuales quedan registradas en los logs de acceso. Estas consultas son:

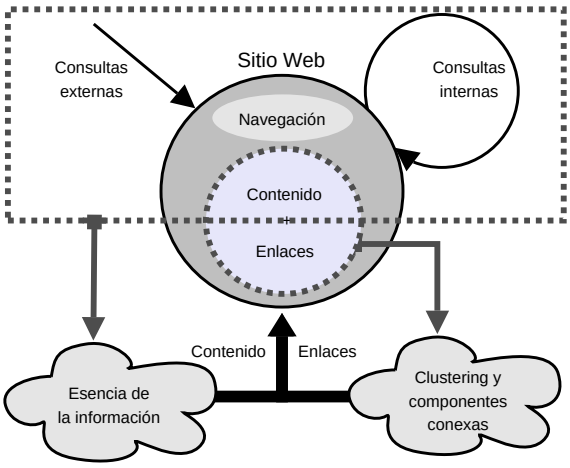

Figura 1: Descripción del modelo.

Consultas externas: Son las consultas realizadas a un buscador de la Web (externo al sitio), a partir de las cuales los usuarios seleccionaron y visitaron los documentos de un sitio en particular. Estas consultas pueden ser descubiertas utilizando el campo `referer` de los logs de acceso y siempre son consultas satisfactorias para el usuario (es decir que tienen respuesta).

Consultas internas: Son aquellas consultas formuladas en la caja de búsqueda interna de un sitio. En este tipo de consultas, se pueden dar las consultas sin resultados. Adicionalmente, consultas en buscadores externos restringidas a un sitio en particular, serán consideradas como consultas internas para ese sitio. Por ejemplo, consultas en Google.com que incluyen `site:example.org` son consultas internas para el sitio `example.org`.

Para cada consulta formulada en un motor de búsqueda, si una página de resultados es generada, esta página contiene enlaces a documentos considerados como respuestas apropiadas por el buscador. Al revisar el resumen que acompaña a cada documento (el cual permite al usuario establecer superficialmente si un documento es apropiado a su consulta) el usuario puede decidir visitar cero o más documentos de la lista.

En la figura 1 se muestra una descripción global del modelo, el cual recopila información de las consultas internas y externas del sitio, los patrones de navegación y los contenidos para descubrir esencia de la información que se puede utilizar para mejorar el contenido del sitio. Adicionalmente los datos de los enlaces y contenidos del sitio, son

analizados usando clustering de documentos similares y componentes conexas. Este procedimiento será explicado en mayor detalle más adelante.

4.1. Modelo de navegación

Al analizar el comportamiento de navegación de los usuarios que visitan un sitio Web, durante un periodo de tiempo determinado, el modelo puede establecer diferentes tipos de documentos en el sitio, estos son: *Documentos alcanzados Sin Buscador* ***(DSB)***, *Documentos alcanzados desde el Buscador Interno* ***(DBI)*** y *Documentos alcanzados desde un Buscador Externo* ***(DBE)***. A continuación se presentará la definición de estos tres tipos de documentos.

DSB: Son aquellos documentos que en el transcurso de una sesión, fueron alcanzados recorriendo enlaces sin la necesidad de utilizar un motor de búsqueda (interno al sitio o externo). Es decir, los documentos alcanzados desde la página de resultados de un buscador y los documentos recorridos desde estos resultados *no* son considerados en este conjunto. Sí son considerados en este conjunto todos los documentos navegados desde documentos visitados antes de utilizar un buscador. De esta manera se admite la posibilidad de que el usuario realice una o varias búsquedas para luego volver a alguna de las páginas previas a estas consultas y retomar la navegación a través de enlaces.

DBI: Son aquellos documentos que en el transcurso de una sesión, sólo son alcanzados por ser resultados directos de una consulta en un buscador interno del sitio Web.

DBE: Son aquellos documentos que en el transcurso de una sesión, sólo son alcanzados por ser resultados directos de una consulta en un buscador externo al sitio Web.

Es importante notar que DSB, DBI y DBE *no son conjuntos disjuntos* de documentos. Esto se debe a que en una sesión un documento puede ser alcanzado utilizando un motor de búsqueda (y por lo tanto pertenecer a DBI o DBE) y en otra sesión el mismo documento puede ser alcanzado navegando a través de enlaces (perteneciendo a DSB). Lo importante entonces es registrar *cuantas veces* cada uno de estos diferentes eventos ocurre para cada documento. Consideraremos la frecuencia de cada evento directamente proporcional a la relevancia de éste para mejorar el sitio Web. La clasificación de los documentos, en estas

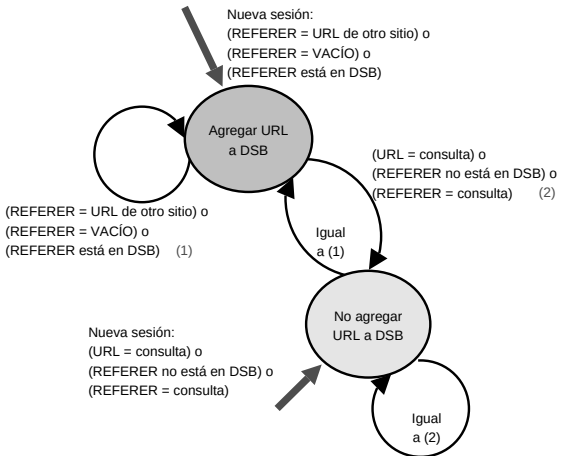

Figura 2: Heurística para DSB.

tres categorías, será esencial para que el modelo descubra información útil desde las consultas en un sitio.

Los documentos que pertenecen a los conjuntos DBI y DBE son fáciles de descubrir, y pueden ser identificados cuando la URL en un requerimiento HTTP es igual a la página de resultados de un buscador (interno o externo). En estos casos, *sólo la primera vez* que un documento es solicitado en una sesión se clasifica. Por otra parte, los documentos en el conjunto DSB son más difíciles de clasificar, debido a que la navegación hacia adelante y hacia atrás en el historial de documentos previamente visitados por el navegador no se registra en los logs de acceso del servidor. Para resolver este problema, hemos creado la heurística mostrada en la figura 2, la cual ha sido validada por nuestros resultados empíricos. La figura 2 muestra un diagrama de estados que comienza un nuevo ciclo de clasificación al inicio de cada sesión y luego procesa secuencialmente cada requerimiento producido por esa sesión al servidor Web del sitio. Al comienzo del ciclo de clasificación el conjunto DSB es inicializado con el valor de la página (o páginas) de inicio del sitio Web y cualquier documento solicitado desde un documento en el conjunto DSB, desde otro sitio o desde un referer vacío (es el caso de los documentos que son *bookmarks* o "favoritos") son agregados al conjunto DSB.

4.2. Consultas satisfactorias

El comportamiento de los usuarios en un sitio Web, reflejado en su navegación, permite establecer diferentes tipos de consultas dependiendo del comportamiento que se produce a partir de estas.

Si un usuario decide visitar documentos desde la página de resultados de una consulta, diremos que esa consulta es de clase A o B. Este tipo de análisis puede realizarse para consultas internas o externas. A continuación definiremos formalmente las consultas de clase A y B (ver la figura 3):

Consultas clase A: Son consultas para las cuales la sesión visitó uno o más documentos en DA, en donde DA contiene documentos encontrados en el conjunto DSB. En otras palabras, los documentos en DA son documentos que, en otra u otras sesiones, a su vez han sido alcanzados navegando sin utilizar un motor de búsqueda.

Consultas clase B: Son consultas para las cuales la sesión visitó uno o más documentos en DB, en donde DB contiene documentos que sólo han sido clasificados en DBI y/o en DBE y no en DSB. En otras palabras, los documentos en DB *sólo* han sido alcanzados utilizando búsquedas, para todas las sesiones analizadas.

El propósito de definir estas dos clases de consultas es que las consultas A y B *contienen palabras claves que pueden ayudar a describir los documentos que son alcanzados como resultado de estas consultas*. En el caso de consultas clase A, estas palabras pueden ser utilizadas en el texto que describe los enlaces a documentos en DA, contribuyendo con esencia de la información adicional para las descripciones de los enlaces existentes hacia estos documentos. El caso de consultas clase B es aún más interesante, debido a que las palabras utilizadas en en estas consultas describen los documentos en DB mejor que las palabras utilizadas actualmente en las descripciones de los enlaces de estos documentos, contribuyendo con nueva esencia de la información para los documentos en DB. Además, los documentos de DB más frecuentes deberán ser considerados por el administrador del sitio como buenas sugerencias para ser documentos alcanzables desde los primeros niveles del sitio (esto también se aplica, en menor grado a los documentos DA). De esta forma, es posible sugerir *hotlinks* basándose en las consultas y no en la navegación, como se suele hacer. En este punto, es importante notar, que la misma consulta puede ser clasificada como clase A y clase B (lo que no puede ocurrir es que un mismo documento pertenezca a DA y DB), así es que la relevancia que se asocia a cada consulta es proporcional a su frecuencia en cada una de las clases en relación a la frecuencia de los documentos en DA y DB.

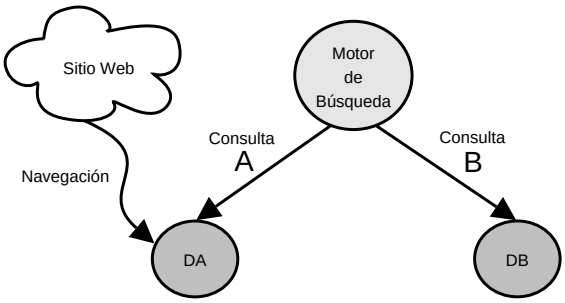

Figura 3: Consultas clase A y B.

4.3. Consultas no satisfactorias

En el caso de que el análisis de la navegación para una sesión, muestre que una consulta realizada al motor de búsqueda interno tuvo cero resultados visitados, esto normalmente puede significar una de dos cosas:

1. El buscador desplegó cero documentos en la página de resultados, debido a que no existen documentos apropiados para la consulta en el sitio.
2. El buscador desplegó uno o más resultados, pero ninguno de estos pareció apropiado a la consulta desde el punto de vista del usuario. Esto puede ocurrir cuando el contenido del sitio es reducido o con consultas que tienen más de un significado (palabras polisémicas).

Para clasificar estas consultas es necesario separar las diferentes causas. Con este objetivo definimos las clases C, C', D y E (ver figura 4), en donde las consultas C', D y E son obtenidas por medio de una clasificación manual realizada por el administrador del sitio:

Consultas clase C: Corresponde a consultas para las cuales el buscador desplegó uno o más resultados, pero sin embargo el usuario no eligió documentos para visitar. Esto puede ocurrir en el caso de consultas que tienen significados ambiguos y para las cuales el sitio tiene documentos que reflejan las palabras de la consulta, pero no el significado que el usuario estaba buscando. Las consultas clase C representan conceptos que deben ser desarrollados en los contenidos del sitio con el significado que los usuarios necesitan, enfocándose en las palabras claves de la consulta.

Consultas clase C': Corresponde a consultas para las cuales el buscador desplegó cero documentos como resultado, debido a que las

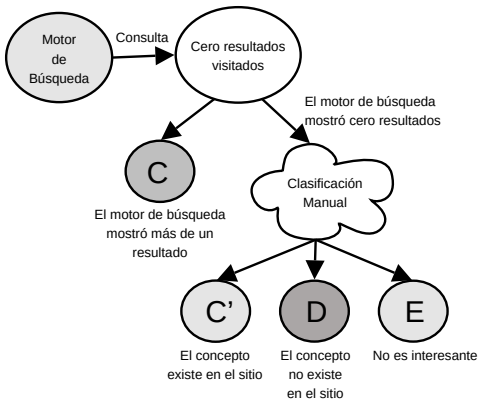

Figura 4: Consultas clase C, C', D y E.

palabras utilizadas en la consulta no existen dentro del sitio. En el caso de que la clasificación manual establezca que el significado de la consulta *existe* en el sitio, pero descrito con otras palabras, entonces la consulta es clase C'. Estas consultas representan palabras que deberían ser utilizadas en el texto que describe los enlaces y documentos que comparten el mismo significado que estas consultas.

Consultas clase D: Al igual que en las consultas clase C', para estas consultas el buscador desplegó cero documentos como resultado, debido a que las palabras utilizadas en la consulta no existen dentro del sitio. Sin embargo, después de hacer la clasificación manual se establece que el concepto expresado por la consulta *no existe* en el sitio (ni siquiera explicado con otras palabras) y que debería ser incorporado a los contenidos del sitio Web ya que representa nuevos temas de interés para los usuarios.

Consultas clase E: Son aquellas consultas para las cuales no existen resultados en el sitio y que no son de interés para éste, ya que no pertenece a las consultas clase C' ni D. Estas consultas deben ser omitidas de la clasificación e incluyen aquellas que poseen errores tipográficos. Sin embargo, si hay un error muy frecuente, puede ser importante incluir esta palabra en las páginas que poseen la palabra escrita en forma correcta. En este caso pasarían a ser consultas de una de las clases anteriores.

Cada clase de consulta es útil de forma diferente al momento de mejorar la estructura y el contenido de un sitio. La importancia de cada consulta se considerará proporcionalmente a su frecuencia

Clase	Existe semántica	Respuestas	Visitas	Tipo Documento	Interés
A	sí	sí	sí	DB ∩ DSB	bajo
B	sí	sí	sí	DB – DSB	medio
C	sí	sí	no	—	bajo
C'	sí	no	no	—	medio
D	debería	no	no	—	alto
E	no	no	no	—	ninguno

Cuadro 1: Clases de consultas.

en los logs de uso y cada consulta será contada sólo una vez por sesión. En el cuadro 1 se muestra una comparación entre las diferentes clases de consultas (en este cuadro DB = DBI ∪ DBE).

La clasificación de este tipo de consultas es con memoria, es decir, una consulta previamente clasificada manualmente no necesita ser clasificada en ejecuciones posteriores de la herramienta. Además se utiliza un tesauro simple para asociar consultas con sus sinónimos. De hecho, con el tiempo, esta herramienta construye un tesauro a la medida para cada sitio.

4.4. Patrones frecuentes de consulta

Todas las consultas formuladas en el motor de búsqueda interno del sitio Web se analizan para descubrir patrones frecuentes de consulta. Estos patrones son conjuntos de palabras que se repiten frecuentemente en las consultas. Una vez descubiertos estos conjuntos, estos son verificados en el motor de búsqueda interno para ver si tienen respuestas en el sitio o no. Si patrones altamente frecuentes no tienen respuestas en el sitio, entonces es necesario revisar los contenidos relacionados a estas palabras para profundizar estos temas.

4.5. Clustering de texto

Nuestro modelo de minería de sitios Web realiza *clustering* de los documentos en el sitio de acuerdo a la similaridad de su texto (el número de clusters utilizado es un parámetro para el modelo). En esta etapa de clustering los documentos son representados como vectores, en los cuales cada coordenada representa una palabra del vocabulario del sitio Web y el valor de esa coordenada es la frecuencia con que ocurre esta palabra en el sitio. Esto se hace para obtener una representación simple y global de la distribución del contenido entre los documentos y para comparar esto con la organización de los enlaces en el sitio. Esta característica es utilizada para evaluar si es que documentos con texto similar, que no tienen enlaces entre ellos, deberían ser enlazados para mejorar la

estructura del sitio Web. Es importante destacar que no estamos diciendo que todos los documentos de texto similar deberían tener enlaces entre sí, ni que este sea el único criterio existente para asociar documentos, pero consideramos que este es un método útil y directo para evaluar la interconectividad de sitios Web (en especial sitios muy grandes).

El modelo además relaciona los resultados del clustering con la información de la clasificación de consultas. Esto permite conocer cuales documentos en cada cluster pertenecen a los conjuntos DA y DB, y la frecuencia con que estos eventos ocurren. Esto apoya la idea de añadir nuevos grupos de documentos (o temas) de interés en la distribución de contenidos de los primeros niveles del sitio. Enfocando, tal vez, los contenidos del sitio hacia los clusters más visitados. Además este análisis entrega información de los *clusters* más visitados y de si fueron alcanzados utilizando un buscador o por medio de navegación a través de enlaces en el sitio.

5. Nuestro prototipo y un caso de uso

La implementación del modelo involucró diferentes tareas de minería Web, tales como: limpieza de datos, identificación de sesiones y usuarios, descubrimiento de patrones, clasificación de consultas, separación de contenido y estructura, y análisis de clusters de texto. Los datos de los contenidos y la estructura se extrae del índice generado por el recolector de páginas del motor de búsqueda del sitio [1], la herramienta utilizada para el clustering de texto es CLUTO [14] por ser de gran eficiencia y rapidez. Este proceso utiliza el método de bisecciones secuenciales, el cual genera buenos resultados para clustering de texto [14], aunque nuestro modelo puede incorporar cualquier otra técnica de clustering. Para el análisis de patrones de consulta se utilizó LPMINER [22]. Las consultas internas provienen del buscador interno y consideramos sólo consultas externas provenientes de Google.com (aunque pueden incorporarse más buscadores externos). Más detalles de la implementación se encuentran en [21].

El cuadro 2 contiene una muestra de las diferentes clases de consultas obtenidas de un log de dos meses perteneciente a un portal chileno dirigido a estudiantes universitarios. El log contiene más de 595 mil sesiones, 3.8 millones de documentos visitados (contando sólo visitas únicas por sesión), 57 mil consultas internas y 162 mil consultas externas. En este cuadro se muestran las consultas más frecuentes para cada clase. Algunas de las sugerencias que se infieren son el añadir información sobre becas en España y proveer un test vocacional. Además, los usuarios no gustan de los resultados entregados para la consulta *leyes* y no encuentran información acerca de clases nocturnas ya que en el sitio no se utiliza la palabra *vespertinas*. Las consultas de clase E no se muestran ya que no son relevantes para el portal (ej. *msn* o *inteligencia emocional*).

Clase A	**Clase B**
examen admisión	exámenes admisión
universidades	currículum vitae
chat	librerías
inserción laboral	universidades
becas universidades	carta presentación
ensayos admisión	examen inglés
tesis	becas universidades

Clase C	**Clase C'**
becas	carreras vespertinas
admisión	diplomas
carreras	España
ensayo	**Clase D**
universidad chile	test vocacional
leyes	becas España
resultados admisión	calcular puntaje

Cuadro 2: Ejemplos de diferentes clases.

La figura 5 muestra parte de la visualización del análisis del clustering, mencionado en la sección anterior. El número de enlaces entre documentos de diferentes nodos y clusters se presenta para cada nivel del árbol jerárquico de clustering. Las "áreas problemáticas", tales como los clusters que no tienen enlaces entre sus documentos o clusters muy similares que tienen pocos enlaces, son representadas utilizando diferentes colores, y pueden ser exploradas inmediatamente. Esto permite a un humano experto decidir cuales clusters necesitan mejorar su interconectividad.

Al aplicar las principales sugerencias obtenidas desde los reportes de la herramienta, se pudo observar un incremento de aproximadamente un 20 % en las visitas desde buscadores externos. Esto se tradujo en un aumento en el número de visitas diarias al sitio. Este aumento se debió principalmente al mejoramiento del contenido del sitio y las descripciones de los enlaces, lo cual fue validado por el análisis de las palabras claves utilizadas en las consultas externas. Adicionalmente a esto, se pudo observar una leve disminución en el número de consultas internas, lo cual coincide

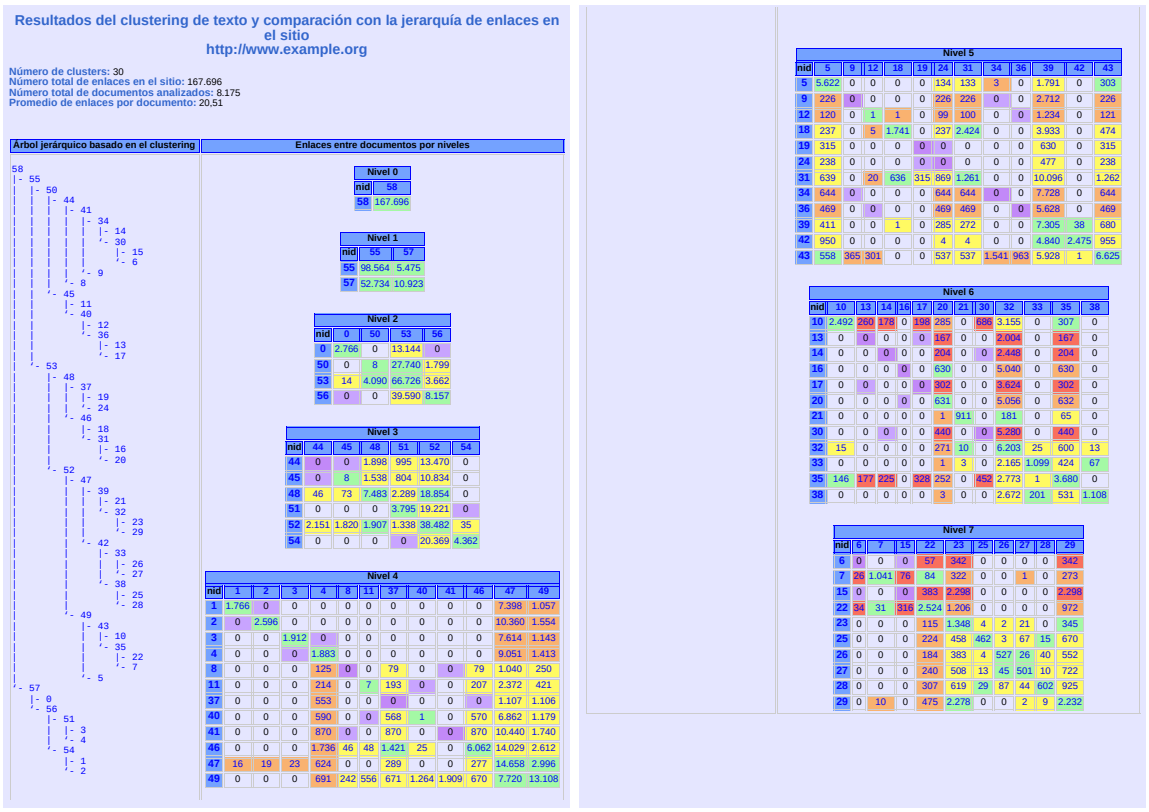

Figura 5: Visualización de clustering.

con nuestra teoría de que, gracias a las sugerencias, los contenidos están siendo encontrados más fácilmente por los usuarios. Todas estas mejoras se han mantenido en los informes generados por la herramienta para los meses posteriores.

6. Conclusiones y trabajo futuro

En este artículo hemos presentado el primer modelo de minería de sitios Web enfocado en consultas. El objetivo de este modelo es encontrar mejor esencia de la información, contenidos y estructura de enlaces para un sitio Web. Nuestra herramienta descubre, en forma muy simple y directa, información interesante. Por ejemplo, las consultas de clase D pueden representar carencia de contenidos, productos o servicios importantes en el sitio. Aunque la clasificación manual pueda parecer algo lenta en un principio, en nuestra experiencia, a largo plazo es casi insignificante, ya que son pocas las consultas nuevas importantes que se presentan.

El trabajo futuro incluye mejorar nuestras visualizaciones del análisis de clustering y extender nuestro modelo para incluir el origen de consultas internas (páginas desde las cuales se formula la consulta). También, añadiremos información de la clasificación y/o el tesauro, además del texto de los enlaces, para mejorar la etapa de clustering. Asimismo nos gustaría modificar el algoritmo de clustering para que determine automáticamente el número de clusters adecuado para el sitio y para que realice un análisis más en profundidad de los clusters más visitados. La etapa de clustering de texto posiblemente será extendido para que incluya lematización en al menos dos idiomas, inglés y español.

Otra característica que nuestro modelo debería incluir es la cuantificación incremental de la evolución del sitio, midiendo los cambios en las consultas y los comportamientos de los usuarios en la medida que pasa el tiempo. En este punto nos interesaría evaluar y validar la calidad de los cambios realizados al sitio gracias a las recomendaciones generadas a partir del modelo.

Finalmente, tenemos la certeza de que existe un gran potencial para descubrir nuevos usos de las consultas para mejorar sitios Web.

Referencias

[1] Akwan Information Technologies. Myweb search. `http://www.akwan.com.br`.

[2] Ricardo Baeza-Yates. Excavando la web (mining the web, original in spanish). *El profesional de la información (The Information Professional)*, 13(1):4–10, Jan-Feb 2004.

[3] Ricardo Baeza-Yates. Web usage mining in search engines. In *Web Mining: Applications and Techniques, Anthony Scime, editor.* Idea Group, 2004.

[4] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In *Web Clustering Workshop at EDBT 2004*, Crete, Greece, 2004. LNCS Springer.

[5] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Ranking boosting based in query clustering. In *Atlantic Web Intelligence Conference*, Cancun, Mexico, 2004. LNCS Springer.

[6] Paulo Batista and Mario J. Silva. Mining on-line newspaper web access logs, 2002.

[7] Bettina Berendt and Myra Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. In *VLDB Journal, Vol. 9, No. 1 (special issue on "Databases and the Web")*, pages 56–75, 2000.

[8] R. Cooley, P. Tan, and J. Srivastava. Websift: the web site information filter system. In *KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag, in press*, 1999.

[9] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: information and pattern discovery on the world wide web. In *9th International Conference on Tools with Artificial Intelligence (ICTAI '97)*, pages 558–567, 1997.

[10] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[11] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Discovery of interesting usage patterns from web data. In *WEBKDD*, pages 163–182, 1999.

[12] Brian D. Davison, David G. Deschenes, and David B. Lewanda. Finding relevant website queries. In *Poster Proceedings of the Twelfth International World Wide Web Conference*, Budapest, Hungary, May 2003.

[13] Z. Huang, J.Ñg, D. Cheung, M.Ñg, and W. Ching. A cube model for web access sessions and cluster analysis. In *Proc. of WEBKDD 2001 (San Francisco CA, August 2001), 47-57.*, 2001.

[14] George Karypis. CLUTO a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. Available at http://www.cs.umn.edu/~cluto.

[15] F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. *ACM SigWeb Letters vol. 8, num. 3*, pages 1–19, 1999.

[16] O. Nasraoui and R. Krishnapuram. An evolutionary approach to mining robust multiresolution web profiles and context sensitive url associations. *Intl' Journal of Computational Intelligence and Applications, Vol. 2, No. 3*, pages 339–348, 2002.

[17] O. Nasraoui, R. Krishnapuram, and A. Joshi. Relational clustering based on a new robust estimator with application to web mining. In *Proceedings of NAFIPS 99, (New York)*, pages 705– 709, 1999.

[18] O. Nasraoui and C. Petenes. Combining web usage mining and fuzzy inference for website personalization. In *Proceedings of the WebKDD workshop*, pages 37–46, 2003.

[19] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu. Mining access patterns efficiently from web logs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 396–407, 2000.

[20] Peter Pirolli. Computational models of information scent-following in a very large browsable text collection. In *CHI*, pages 3–10, 1997.

[21] Barbara Poblete. A web mining model and tool centered in queries. M.sc. in Computer Science, CS Dept., Univ. of Chile, 2004.

[22] Masakazu Seno and George Karypis. Lpminer: An algorithm for finding frequent itemsets using length-decreasing support constraint. In *Proceedings of the 2001 IEEE In-*

ternational Conference on Data Mining, pages 505–512. IEEE Computer Society, 2001.

[23] Myra Spiliopoulou. Web usage mining for web site evaluation. *Commun. ACM*, 43(8):127–134, 2000.

[24] Myra Spiliopoulou and Lukas C. Faulstich. WUM: a Web Utilization Miner. In *Workshop on the Web and Data Bases (WebDB98)*, pages 109–115, 1998.

[25] Myra Spiliopoulou, Carsten Pohle, and Lukas Faulstich. Improving the effectiveness of a web site with web usage mining. In *WEBKDD*, pages 142–162, 1999.

[26] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.