

Creación y posicionamiento de un Wiki

Minería de textos Web



Universidad Carlos III de Madrid
Curso 2006-2007

Herrero Núñez, Julio Alberto 100033200

Contenido

1. INTRODUCCIÓN	3
1.1. SELECCIÓN Y RECOPIACIÓN DE DATOS	4
1.2. TRATAMIENTO PREVIO DE LOS DATOS	4
1.3. TRANSFORMACIÓN DE LOS DATOS	4
1.4. ANÁLISIS DE LAS INFERENCIAS SOBRE LOS DATOS.....	4
1.5. TIPOS DE MINERÍA DE TEXTOS WEB (WEB MINING)	5
2. EL <i>WEB MINING</i> DE CONTENIDO	6
3. EL <i>WEB MINING</i> DE ESTRUCTURA.....	7
4. EL <i>WEB MINING</i> DE USO	8
5. HERRAMIENTAS PARA EL <i>WEB MINING</i>	9
5.1. LOGS	9

1. INTRODUCCIÓN

Una de las extensiones del data mining consiste en aplicar sus técnicas a documentos y servicios del Web, lo que se llama *Web Mining* (minería de web). Se usa para el estudio de varios aspectos esenciales de un sitio y ayuda a descubrir tendencias y relaciones en el comportamiento de los usuarios que sirven como pistas para, por ejemplo, mejorar la usabilidad de un sitio. Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, galletas, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (log).

Las herramientas de Web Mining analizan y procesan estos logs para producir información significativa, por ejemplo, cómo es la navegación de un cliente antes de hacer una compra en línea. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hiperligas, investigaciones recientes usan el término multimedia data mining (minería de datos multimedia) como una instancia del web mining para tratar ese tipo de datos.

Los accesos totales por dominio, horarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del web mining. En definitiva podemos decir que todo el proceso consiste en la **integración de información obtenida mediante los métodos tradicionales de la minería de datos con información recogida sobre la web, es decir, la minería de datos aplicada a las especificidades de la web.**

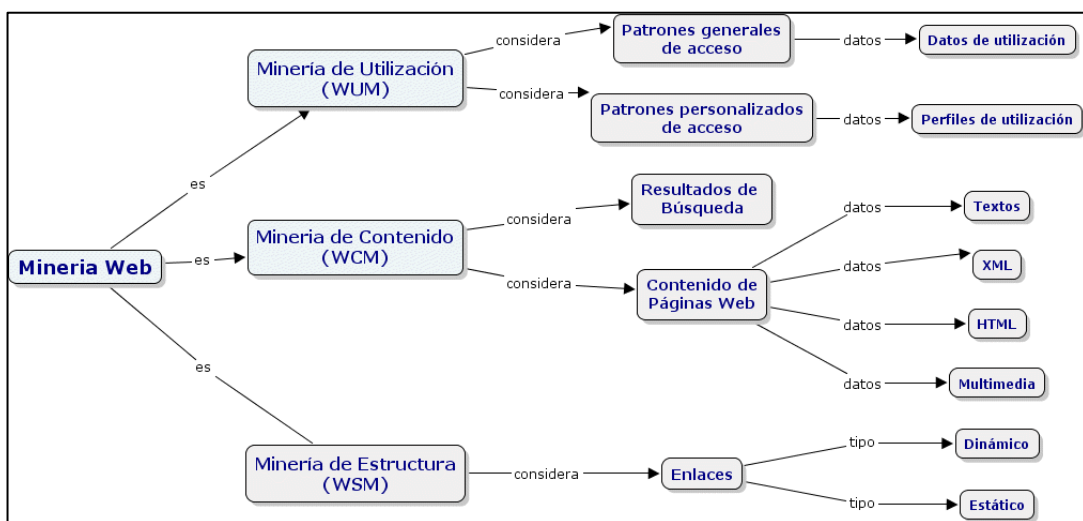


Figura 1. Minería Web

1.1. Selección y recopilación de datos

En primer lugar decidir qué se quiere estudiar y cuáles son los datos que nos facilitarán esa información. Posteriormente se localizan los documentos o archivos a adquirir. Estos se capturarán y se almacenarán los datos pertinentes.

1.2. Tratamiento previo de los datos

Se trata de filtrar y limpiar los datos recogidos. Una vez extraída una determinada información a partir de un documento, ya sea HTML, XML, texto, ps, PDF, LaTeX, FAQs,, se realizan tareas de criba y normalización, eliminando los datos erróneos o incompletos, presentando los restantes de manera ordenada y con los mismos criterios formales hasta conseguir una homogeneidad formal, etc. y demás labores enfocadas a la obtención de unos datos originales listos para su transformación por medios automáticos.

1.3. Transformación de los datos

En esta fase se utilizan algoritmos inteligentes de búsqueda de patrones de comportamiento y detectar asociaciones. Estos algoritmos se elaboran previamente utilizando recursos estadísticos, técnicas procedentes del data mining, etc, se procede a transformar los datos para obtener como resultado, información sobre ellos.

Los principales algoritmos se basan en la reunión de grupos homogéneos (ej. Usuarios que visitan más de un número determinado de páginas), reglas de asociación de páginas, seguimiento de rutas o historial de navegación de una persona, etc.

Esta metamorfosis suministra información que englobe a la mayor parte de los datos estudiados. En esta fase se consiguen generalizaciones que se perciben en el establecimiento de enlaces, en muchas ocasiones en forma gráfica. Esta fase, junto con la próxima, son las más cercanas al campo de la visualización, especialmente en métodos de visualización.

1.4. Análisis de las inferencias sobre los datos

La simple inferencia no tendría un sentido completo si no se razonan los resultados, si no se logra encontrar una justificación a dichos resultados. Es aquí donde, dependiendo del tipo web mining, utilizaremos recursos de las ciencias sociales y económicas. Ya que, como bien se ha comentado, la W3 es una comunidad, un territorio donde los comportamientos

automatizados de relaciones y contenidos vienen decididos por personas que se encuentran tras cada ordenador conectado a la red.

1.5. Tipos de minería de textos web (Web Mining)

El Web Mining nos ayuda a descubrir información, encontrar documentos relacionados, mostrar temáticas, averiguar el grado de satisfacción de recursos web, etc. Según el fin deseado, la actividad de excavar en la web se desglosa en tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos:

- *Web content mining* (minería de contenido web)
- *Web structure mining* (minería de estructura web)
- *Web usage mining* (minería de uso web)

2. EL *WEB MINING* DE CONTENIDO

Busca la regularidad y dinámica de los contenidos en la W3. Los documentos Web pueden ser datos sin estructurar, archivos html parcialmente estructurados, o información procedente de bases de datos generadas en páginas con formato html. Estos documentos hipertexto incluyen texto y también a imágenes, audio, vídeo, metadatos e hiperenlaces.

La metodología utilizada en este apartado, va desde las tradicionales relaciones entre términos hasta la tecnología que se utiliza en la minería textual (text mining). Esta última consiste en analizar elementos textuales con el fin de identificar, deducir y ampliar conocimiento a partir de cualquier organización de documentos (por ejemplo, bases de datos, web...).

La extracción (mining) de información, intenta inferir la estructura del sitio web (web site) para transformarla y convertirla en una base de datos a nivel lógico.

Además, la localización de patrones en el texto de los documentos, el descubrimiento del recurso basado en conceptos de indexación o la tecnología basada en agentes también pueden formar parte de esta categoría.

En definitiva, podemos obtener datos acerca de la forma de escribir que es más atractiva para el usuario, de si la catalogación que usamos sirve para mejorar un ranking, si los temas que se tratan interesan o no.

3. EL *WEB MINING* DE ESTRUCTURA

Web Mining de estructura, intenta descubrir la organización de los enlaces del conjunto de hiperenlaces dentro del documento para generar un informe estructural sobre la página y el sitio web. Obtenemos información acerca de si los usuarios encuentran la información, si la estructura de sitio es demasiado ancha o demasiado profunda, si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación se entiende, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central.

Según el objetivo a estudiar, se pueden dar tres tipos de informes:

- Basándose en los hiperenlaces, clasifica las páginas Web y genera el informe.
- Revelando la estructura del documento Web en sí.
- Descubriendo la naturaleza de la jerarquía o de la red de hiperenlaces del sitio Web de un dominio particular.

Suele dar como resultado representaciones gráficas para una mejor visión del conocimiento obtenido y pueden utilizarse como guía para el usuario en busca de información.

4. EL *WEB MINING* DE USO

El Web Mining de uso es la aplicación de las técnicas de data mining para descubrir pautas de conducta a la hora de utilizar la web por parte de los usuarios.

Esta extracción se refiere a patrones de navegación que podemos descubrir en nuestros usuarios y nos pueden servir para mejorar la misma, por ejemplo si el 80 % de nuestros usuarios recurren al campo de búsqueda cuando entran a nuestro sitio es que deberemos poner énfasis en la mejora de esa interfaz y que el motor que se encuentre detrás devuelva la información deseada. Este proceso se basa en el uso de *logs* de los accesos al web.

En definitiva, se tratan seguir una serie de pautas sobre:

- el acceso que utilizan los clientes cuando consultan el sitio web de una empresa
- los usuarios que interrogan a una aplicación que precede a una base de datos
- los individuos que navegan por páginas determinadas, ...

A partir de datos secundarios derivados de interacciones automáticas de los usuarios mientras navegan por la web se pueden cubrir mejor las necesidades que se solicitan a través de aplicaciones basadas en protocolos W3.

5. HERRAMIENTAS PARA EL WEB MINING

En los tres tipos de extracción de información web se utilizan técnicas que se venían utilizando con la minería de datos y otras que se han planteado y perfeccionado en ambos casos. Se trata de campos extremadamente ligados, el primero centrado en datos hipertextuales en red (W3) y el segundo aplicado a información estructurada o semi-estructurada que se encuentra en bases de datos.

Según pues la rama en la que se esté trabajando dentro de la extracción de información web, se utilizan más los elementos formales o los elementos de contenido. En especial destacar el uso de ficheros *logs*

5.1. LOGS

Los ficheros logs son una grabación de la actividad de un servidor o de un sitio web a lo largo de un período de tiempo determinado. La información se genera automáticamente y suelen incluir la dirección IP de los visitantes, la página solicitada junto con la fecha y hora de la consulta, tiempo de lectura, si han accedido desde buscadores, ...

Suelen ser ficheros voluminosos y registran visitas automáticas de robots, no efectuadas por usuarios de manera voluntaria y con una intención.