EL WEB MINING: UNA TECNOLOGÍA PARA LA INDAGACIÓN EN LA WORLD WIDE WEB

AUTORA: LOLA GARCÍA-SANTIAGO

Resumen:

Este trabajo presenta a nivel divulgativo, las características generales de la extracción de información en la World Wide Web (W3) mediante nuevas técnicas. Esta rama nueva, denominada Web Mining en el mundo anglosajón, trata de profundizar en todos los aspectos de la W3 y que no se encuentran fácilmente al alcance del usuario. Además, se diferencia este concepto de otros similares como la extracción textual (el Text Mining) o la minería de datos (el Data Mining). Se presentan las líneas de investigación en las que se trabaja y las dificultades con las que aún se enfrenta. Finalmente, se indican los potenciales usos del Web Mining.

Palabras clave: Web Mining; Extracción de información; World Wide Web; Hipertexto.

1 Introducción

1.1 La World Wide Web (W3)

La W3 establece conexiones, enlaces entre el universo propio de cada uno y los universos de los otros. Existen lazos entre usuarios, activos y pasivos, a través de los productos informativos que establecen el universo de cada autor. En Internet existen recursos que ofrecen una nueva forma de comunicación interpersonal a tiempo real (ej. Los Internet Relay Chat, IRC) o no (ej. Listas de distribución). En la W3 es diferente, la comunicación se establece a través de los enlaces que forman un entramado, una textura (origen etimológico del vocablo texto). Nos movemos en la red digital y nos proyectamos de un continente a otro sin que exista una verdadera separación. Saltamos de un enlace a otro sin que notemos barreras geográficas ni tecnológicas. La posibilidad técnica de estos enredos deriva de la característica principal de la W3, la hipertextualidad. Concepto, a su vez, muy allegado al de transversalidad.

Nuestra personalidad de red, la imagen que presentamos según la información que mostremos en la W3, se compone de la combinación de varios papeles, identidades y funciones que nos permiten aislarnos o conectarnos con otros. Y es que el ciberespacio ya es la casa de miles de grupos de personas que se reúnen para compartir información, discutir intereses comunes, jugar y llevar a cabo negocios. El concepto de comunidad aparece como el conjunto de interacciones entre personas en un espacio determinado.

1.2 Información en la W3

Existe mucha información en la W3 o derivada de ella. Por un lado la que percibimos a simple vista, los documentos hipermedia que conforman la W3 de manera explícita y que abunda cada vez más. Y, por otro lado, la información subyacente y que se encuentra en capas más profundas.

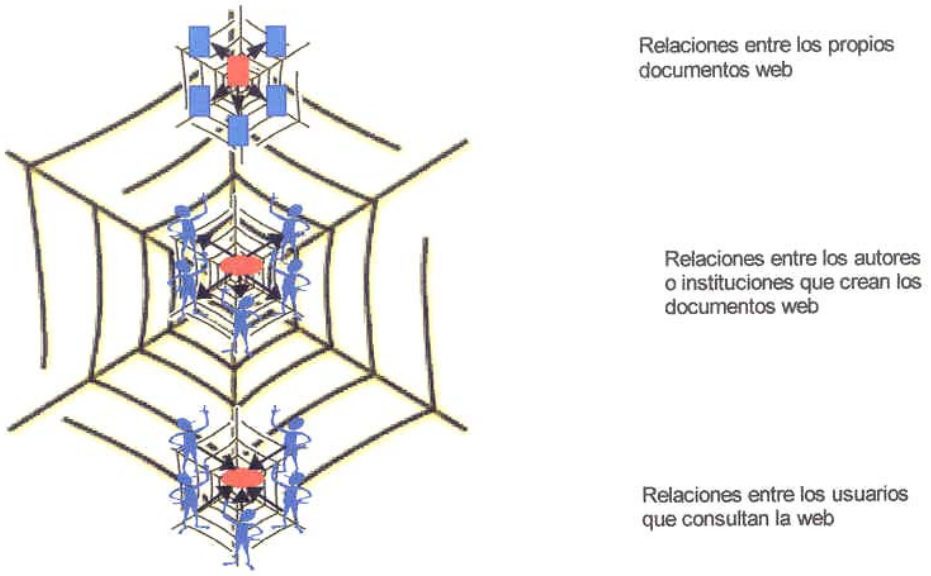
Ya no nos conformamos con lo primero que nos llega a las manos, cada vez exigimos mayor precisión. Nos sumergirnos en niveles más profundos para conocer qué otros datos podemos extraer y que nos pueden servir de utilidad. Para conseguir esto, nos basamos en otras características de la W3.

1.3 Las relaciones en la W3.

Las relaciones son representadas en forma de enlaces. Estos, se incluyen en los documentos hipermedia que alberga la W3. Pueden ser objetivos o subjetivos

según su función y reflejan las relaciones y las redes de relaciones que dichos documentos establecen.

- Relaciones de información: cuando nos referimos a las establecidas entre los textos y contextos.
- Relaciones entre los autores o las instituciones que generan o albergan dichos documentos, hasta llegar a establecer verdaderas redes dentro de la comunidad.
- Relaciones entre los usuarios que utilizan los recursos que tiene la W3.



[Fig.1: Tipos de relaciones en la W3]

1.4 Problemas a la hora de encontrar tanta información

El browsing, el serendipity,… son maneras de buscar y encontrar información saltando de una página web a otra a través de los enlaces. A partir de una ubicación concreta se decide navegar y amplíar la búsqueda a otras páginas remitidas por la página de partida y así sucesivamente. Este sistema provoca un fenómeno de "desorientación" dentro de la W3, que no siempre nos conduce a la información deseada. Algunas razones para que se produzca este fenómeno son:

- La ingente cantidad de documentos web en este territorio del ciberespacio.
- La cobertura limitada de la W3. Con recursos ocultos o poco accesibles (licencias, suscripciones y acceso previo pago), generalmente datos procedentes de bases de datos.
- Programa de consulta basado en búsquedas por palabra clave.
- Personalización para usuarios individuales.

Con el uso de **buscadores** tenemos solventado el problema de la desorientación al obtener listados de direcciones. Se trata de una recuperación automatizada, previa búsqueda en las bases de datos de estos recursos. La pertinencia de los resultados vendrá en función de la calidad de las técnicas utilizadas para la búsqueda,

almacenamiento y elaboración de las consultas. Estos robots, han ido evolucionando con el paso del tiempo.

> En una primera fase, los buscadores, se basaban únicamente en la comparación de cadenas de caracteres. En la segunda generación, se tienen en cuenta las direcciones que más han sido enlazadas. Y en la tercera, se establecen ponderaciones sobre los enlaces relacionados y que además que contengan dichas cadenas de caracteres.

Pero un inconveniente que todavía persiste en los buscadores es la barrera lingüística. La búsqueda de información y la forma de interrogar al motor de búsqueda queda aún limitado por el idioma.

2 El Web Mining

2.1 ¿Qué es?

La minería de datos o web mining se refiere al proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web (Etzioni 1996).

Es un campo multidisciplinar donde convergen áreas como la recuperación de información, el data mining, la estadística, la visualización de datos, lenguajes de etiquetas, tecnología web, etc, con el objetivo de descubrir redes de relaciones existentes en la W3, utilizando su información desestructurada o semi-estructurada.

Es decir, una vez transformados los datos y planteado el algoritmo a seguir, es el sistema el que muestra representaciones y sugiere modelos. Esta visión es diferente al tradicional planteamiento de leyes (modelos preestablecidos por investigadores) o cualquier otra hipótesis que, una vez reconvertidos los datos, es ese analista el que comprueba si los resultados se ajustan al patrón previamente planteado. Y este campo se diferencia de la minería de datos o data mining en que éste pretende descubrir modelos existentes dentro de bases de datos estructurados.

2.2 Origen del término

La primera aparición del término ***Web Mining*** es en 1996 en un artículo de Oren Etzioni [Etzioni 1996]. Y los define como "el uso de las técnicas de data mining con el fin de descubrir y extraer información de los servicios y documentos de la World Wide Web de manera automática".

2.3 ¿Cuáles son sus objetivos?

- Mejorar la navegación del usuario en un espacio tan vasto y cambiante como es la W3. Tener representaciones gráficas que reflejen los cambios sufridos y/o representar la estructura general de la red.
- Descubrir recursos, extraer información, analizar datos e inferir generalidades.
- Encontrar información relevante
- Obtener nuevos conocimientos provenientes de la información disponible en la W3
- Personalizar la información
- Saber más sobre usuarios o clientes

2.4 Cómo se trabaja en el Web Mining

2.4.1 Selección y recopilación de los 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.

2.4.2 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.

2.4.3 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.

2.4.4 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.

3 Tipos de 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 líneas.

3.1 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 imagenes, 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.

3.2 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. 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.

3.3 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. 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.

4 Herramientas para el Web Mining

Como ya he comentado al principio, 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. Apuntamos algunos de ellos.

4.1 Metadatos

Los ***metadatos***, entendidos como normas de representación de la estructura autoidentificativa del documento. El análisis de estos proporciona un mecanismo formal para la categorización y clasificación automática de documentos. Aplicando a los metadatos unas determinadas escalas conceptuales, se pueden construir espacios conceptuales facetados según la perspectiva que le interese a cada usuario. Esta modalidad de identificación permite el uso de programas y servicios informáticos.

Utilizado principalmente en el Web Mining de Contenido. Con estos elementos y con la ayuda de la inteligencia artifical, se intenta conseguir deducciones terminológicas, predicciones en respuestas a consultas complejas,… todo ello cuando las relaciones entre términos y los conceptos que representan no mantienen una relación lineal directa.

4.2 Hiperenlaces

En la tecnología hipertextual, cada bloque de texto contiene una multitud de palabras clave, pictogramas y/o dibujos que son susceptibles de ser marcados con el ratón. Estos puntos de intersección, denominados "enlaces".
Estos enlaces se pueden desglosar en:
Externos: entre documentos diferentes
Internos: que a su vez pueden ser estructurales (incluyendo elementos multimedia en el documento, o de referencia a otros puntos del mismo documento.

4.3 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.

4.4 Métodos estadísticos

Como el clustering o proceso de encontrar grupos tras un procesamiento de los datos. Es decir, a priori se desconoce el número de grupos o las características de los mismos. Otro método es el escalamiento multidimensional (MDS),…

4.5 Reglas de Asociación

Las relaciones planteadas entre elementos web (contenidos, documentos, instituciones, usuarios,…) se materializan con la inclusión de hiperenlaces. El poder de decisión a la hora de incluir o no un nuevo enlace muestra el grado de interés hacia ese enlace establecido.
Una de las herencias procedentes del campo de la recuperación de información son los análisis de citas. Bajo este planteamiento, se establecen relaciones entre elementos u actores sociales. Las asociaciones entre usuarios que consultan una misma página, los entes que son enlazados por otros entes, los textos más utilizados a lo largo del tiempo y su conexión con otros textos,…son claros ejemplos de relaciones sincrónicas y diacrónicas.
Estas reglas son una técnica alternativa para detenerse en modelos que se repiten entre usuarios que comparten caminos transversales similares. En algunos motores de búsqueda se ha implantado ya esta filosofía de relaciones para una mayor precisión en los resultados obtenidos.

5 Futuro del Web Mining

El potencial que tiene el Web Mining o extracción de información web para detectar colegios invisibles es muy alto y además de práctico, necesario ante el crecimiento de la información en todo tipo de formatos, más aún en la W3. Estos colegios invisibles se establecen como redes de relaciones existentes, directas o indirectas, entre autores de documentos web que versan sobre una misma temática o línea de investigación concreta.

El reconocimiento y representación de las comunidades científicas latentes, permitirán a las personas a navegar, a buscar y ver los contenidos que alberga la W3.

Por una parte permite descubrir y describir redes de relaciones y pautas de comportamiento en la W3, lo que proporciona guías para el usuario y la navegación por ámbitos concretos. Por otro lado, facilita el poder de predicción y el grado de exactitud a la hora de recuperar información tras una consulta compleja y sin la ayuda de lenguajes controlados que analicen el contenido de los documentos.

Con esta introducción se ha pretendido hacer un esbozo de lo que es el Web Mining los usos para la mejora en la recuperación de información web y para proporcionar más información sobre la red hipertextual y mostrar las dificultades con las que trabaja dadas las características de la red. Se trata de una red poco estructurada, pero menos aleatoria de lo que se puede percibir en un primer momento.

El Web Mining nos da la oportunidad de encontrar nuevos recursos, extraer la información más interesante y, tras un proceso de análisis, finalmente mostrar modelos de información de carácter general en la W3.

6 Bibliografía

- COOLEY, R. (2000) http://www.cyberartsweb.org/cpace/ht/lanman/bibli.htm [Cooley 2000]
- ETZIONI, O. (1996). "The World-Wide Web: Quagmire or Gold Mine?". Communications of the ACM, november 1996, Vol. 39, No. 11
- JIAWEI, H. y MICHELINE, K. "Data Mining: Concepts and Techniques" http://www.cs.uiuc.edu/~hanj
- WANG, Y. "Web mining and knowledge discovery of usage patterns - A survey"