Web Mining: Fundamentos Básicos

Francisco Manuel De Gyves Camacho
Doctorado en informática y automática
Universidad de Salamanca

fdegyves@usal.es

Resumen. La web es uno de las aplicaciones o fenómenos más importantes que han surgido en los últimos tiempos, ¿por que?, no hay más que ver la rápida aceptación que tuvo en la sociedad, ya que aquí descubrieron un medio significativo para exponer riquezas de información. Los motores de búsqueda son actualmente los mejores repositorios de información de la web, esto es debido al volumen de datos que contienen. Los usuarios acuden a este medio con el fin de localizar información más sin embargo si no se utilizan adecuadamente o no se busca bien, pueden ser no fructíferos. La web mining juega un rol importante hacia lograr la efectividad en las relaciones de patrones interesantes.

1. Introducción

El crecimiento desmedido de la información que se encuentra en la web ha sido exponencial debido a la necesidad de los usuarios (personas físicas, empresas, universidades, gobierno, etc.) de contar con datos para la interrelación en el mundo globalizado. De acuerdo con Baeza–Yates, la información de la web es finita pero el número de páginas web es infinita0 [1]. Actualmente existen alrededor de 4 mil millones de páginas estáticas, es decir la información que poseen los buscadores web, más sin embargo es importante mencionar que la mayoría de las páginas web y que no son indexables que existen en la web son dinámicas, es decir son aquellas que se generan automáticamente con datos extraídos de bases de datos [2].

Existen diferentes problemas a los que se enfrentan los usuarios debido al crecimiento exponencial. Uno de esos problemas es el que representa encontrar información relevante, esto es por dos aspectos muy relevantes, la baja precisión y la escasa cobertura. La escasa cobertura es debido a que no todos los motores de búsqueda tienen la suficiente capacidad de indexar la web, debido a varios factores; el ancho de banda, el espacio de disco duro, el costo económico, etc.

La web mining actualmente es un área de investigación extensa dentro de varios grupos de investigadores, especialmente interesados debido al alto crecimiento de la información que existe en la web y por el movimiento económico que ha generado el e-commerce y sobre todo para intentar resolver los problemas que se han mencionado anteriormente, ya sea de manera directa o indirectamente. Actualmente

lo que se ha pretendido realizar es aprender de acuerdo a los comportamientos de los usuarios en su andar por la web y así proporcionarles información realmente relevante, útil y personalizada en muchos casos.

El presente documento pretende adentrarnos un poco en el mundo de la web mining, permitiéndonos conocer los aspectos básicos tales como, el proceso general de la web mining así como unas breves aproximaciones del web content mining, web structure mining y de la web usage mining

2. Particularidades de la Web

Es común encontrarse en la web con ciertas características preponderantes que parecieran problemas, sin embargo podrían catalogarse como oportunidades sin precedentes para la obtención de información y mejoramiento de la web.

En los tiempos actuales, es común encontrar casi todo tipo de información en la web en cantidades descomunales y fácilmente accesibles. La información en la Web por lo regular es heterogénea, es decir muchas páginas presentan la misma o similar información usando formatos diferentes. Podemos decir que la información es redundante. La Web normalmente esta compuesta por una mezcla de tipos de información, por ejemplo, contenido principal, anuncios, paneles de navegación, noticias de copyright, etc. Para una aplicación en particular solo parte de la información es útil y el resto es basura ó no útil [3]. La información de la web cambia constantemente, es decir es dinámica.

3. Web mining y su proceso

Algunos autores definen a la web mining como el uso de técnicas para descubrir y extraer de forma automática información de los documentos y servicios de la web[0]. Según M. Scotto [4] la web mining es el proceso de descubrir y analizar información "útil" de los documentos de la Web. Sin embargo y tomando en cuenta lo expuesto en la introducción la minería web se puede definir como el descubrimiento y análisis de información relevante que involucra el uso de técnicas y acercamientos basados en la minería de datos (*Data Mining*) orientados al descubrimiento y extracción automática de información de documentos y servicios de la Web, teniendo en consideración el comportamiento y preferencias del usuario.

En la web mining, los datos pueden ser coleccionados en diferente niveles; en el área del servidor, en el lado del cliente (cookies), en los servidores proxys (log files), etc.

De acuerdo con Etzioni 0] el proceso general de la web mining es el siguiente:

Recuperación de Información (IR)

No referimos básicamente al proceso del descubrimiento automático de documentos relevantes de acuerdo a una cierta búsqueda.
Documentos relevantes disponibles en la web tales como noticias electrónicas, newsgroups, newswires, contenido de las html, etc.

Extracción de Información(IE)

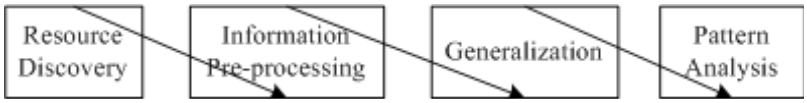
Tiene como objetivo transformar los documentos extraídos en el proceso de recuperación de información, en documentos que sean más digeribles, fáciles de leer y de analizar.

Generalization

Reconocimiento de Patrones generales de una página en particular o bien también patrones de diferentes páginas.

Analysis

Una vez que los patrones han sido identificados, la parte humana juega un papel importante haciendo uso de herramientas adecuadas para entender, visualizar e interpretar los patrones



3.1 .Motores de búsqueda

Los motores de busqueda han tenido un éxito considerado gracias a la enorme necesidad que tienen los usuarios de indagar información no disponibles en sitios físicos y si en la Internet. Un motor de búsqueda tiene como objetivo indexar archivos almacenados en los servidores web, un ejemplo son los *buscadores de Internet.* El resultado de la búsqueda es un listado de direcciones Web en los que se mencionan temas relacionados con las palabras clave buscadas [6].
Basándonos en las descripciones hechas por Henzinger [7] podemos decir que los motores de búsqueda están integrados básicamente por tres grandes componentes; el crawler, el indexador y el query processor.

Crawler: Recorren las páginas recopilando información sobre los contenidos de las páginas. Recolectan páginas de manera recursiva a partir de un conjunto de links de páginas iniciales. Cuando buscamos una información en los motores, ellos consultan

su base de datos, y nos la presentan clasificados por su relevancia. De las webs, los buscadores pueden almacenar desde la página de entrada, a todas las páginas de la web.

Indexador: el objetivo principal del indexador es procesar las páginas coleccionadas por el crawler. La indexación proporciona agilidad en las búsquedas, lo que se traduce en mayor rapidez a la hora de mostrar resultados.

Query processor: procesa las consultas de los usuarios y regresa resultados de acuerdo a esas consultas y de acuerdo a un algoritmo de posicionamiento.

3.2 Taxonomía de la Web mining

3.2.1 Minería de contenido de la web

Su objetivo es la recogida de datos e identificación de patrones relativos a los contenidos de la web y a las búsquedas que se realizan sobre los mismos. Es decir son los datos reales que se entregan a los usuarios, los datos que almacenan los sitios web[8].

La minería de contenidos consiste de datos desestructurados tales como texto libre, semi-estructurado como documentos HTML, y mas estructurados como datos en tablas o páginas generadas con datos de BD.

Existen dos grupos de estrategias sobre minería de contenidos: aquellas que minan directamente el contenido de los documentos y aquellas que mejoran en la búsqueda de contenidos.

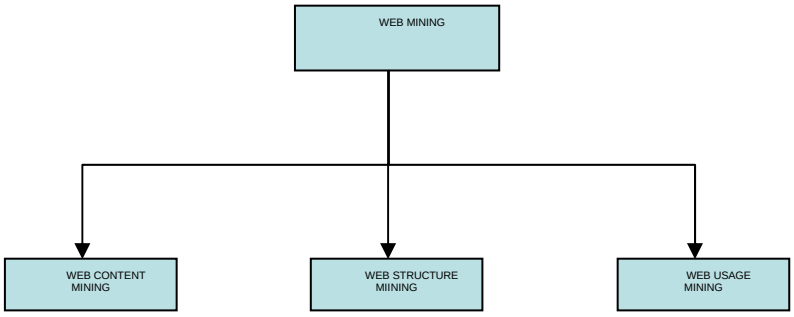
3.2.2 Minería de estructura de la web

La minería de estructura intenta descubrir el modelo subyacente de las estructuras de los enlaces del web. El modelo se basa en la topología de los hiperenlaces con o sin la descripción de los enlaces. Este modelo puede ser usado para categorizar las páginas web y es útil para generar información tal como la similitud y relación entre diferentes páginas web [9]. Es decir pretende revelar la estructura real de un sitio web a traves de la recogida de datos referentes a su estructura y, principalmente a su conectividad. Típicamente tiene en cuenta dos tipos de enlaces: estáticos y dinámicos.

3.2.3 Minería de uso de la web

La minería de uso intenta dar sentido a los datos y comportamientos generados en las sesiones de navegación del web. Es decir son aquellos datos que describen el uso al

cual se ve sometido un sitio, registrado en los logs de acceso a de los servidores web. A partir de esta información se podría concluir, por ejemplo, que documento visitado no tiene razón de ser, o si una página no se encuentra en los primeros niveles de jerarquía de un sitio [8]. Analizar los logs de diferentes servidores web, puede ayudar a entender el compartamiento del usuario, la estructura de la web, permitiendo de este modo mejorar el diseño de esta colección de recursos [10].



4. Web Content Mining

La minería de contenido, tiene como principal objetivo otorgar datos reales o finales a los usuarios que interactúan con la Web. Es decir, extraer información "útil" de los contenidos de las páginas web.

Generalmente la información disponible, se encuentra de forma no estructurada (minería de Texto), semi-estructurada y un poco más estructurada como es el caso de tablas html generadas automáticamente con información de bases de datos.

De acuerdo con Raymon Kosala y Hendrick Blockeel0 [9], la minería de contenido puede ser diferenciada desde dos puntos de vista; desde el punto de vista de la **Recuperación de Información (IR) y desde la vista de Base de Datos (DB).** Es decir asistir en el proceso de recogida de información o mejorar la información encontrada por los usuarios, usualmente basada en las solicitudes hechas por ellos mismos (IR). Desde el punto de vista de DB principalmente trata de modelar los datos e integrarlos en la Web a través de queries sofisticadas

4.1. Minería de Contenido desde el punto de vista de Recuperación de Información y Extracción de Información.

La recuperación de información es el proceso de encontrar el número apropiado de documentos relevantes de acuerdo a una búsqueda hecha en una colección de documentos. La IR y la web mining tienen diferentes objetivos, es decir la web mining no busca remplazar este proceso. La web mining pretende ser utilizada para

incrementar la precisión en la recuperación de información y mejorar la organización de los resultados extraidos0 [5]. La recuperación de información es altamente popular en grandes empresas del mundo web, las cuales hacen uso de este tipo de sistemas, la máquinas de búsqueda (google y altavista), directorios jerárquicos (yahoo) y otros tipos de agentes y de sistemas de filtrado colaborativos.

La diferencia principal, independientemente de las técnicas que usan, que existe entre la Recuperación de la información y la Extracción de la Información recae principalmente en que uno recupera documentos relevantes de una colección y la otra recupera información relevante de dichos documentos. La IE se centra principalmente en la estructura o la representación de un documento mientras que la IR mira al texto en un documento como una bolsa de palabras en desorden [11].

Podemos decir que dichas técnicas son complementarias una de otra y usadas en combinación pueden generar valor agregado.

Datos no estructurados, semi-estructurados y estructurados, son los objetivos de la Extracción de Información, generalmente para los datos no estructurados se hacen uso de técnicas de Lenguaje Natural. Dichas reglas son generalmente basadas en el uso de relaciones sintácticas entre palabras y clases semánticas. Reconocimiento de objetos de dominios tales como, nombres de personas y compañias, análisis sintáctico y etiquetado semántico, son algunos de los pasos para la extracción de información en documentos no estructurados.

Recientemente se ha hecho uso de una tecnología llamada Text mining , que hace referencia principalmente al proceso de extracción de información y conocimiento interesante, no trivial desde documentos no estructurados [6].

Las principales categorías de la Web Text mining son Text Categorization, Text Clustering, association analysis, trend prediction.

Text Categorization: dada una predeterminada taxonomía, cada documento de una categoría es clasificada dentro de una clase adecuada o más de una. Es más conveniente ó fácil realizar búsquedas especificando clases que buscando en documentos. Actualmente existen varios algoritmos de text categorization, dentro de los cuales encontramos, K-nearest, neighbor-algorithm y naive bayes algorithm.

Text Clustering: el objetivo de esta categoría es el de dividir una colección de documentos en un conjunto de clusteres tal que la similitud intra-cluster es minimizada y la similitud extra-cluster es maximizada. Podemos hacer uso de text clustering a los documentos que fueron extraidos por medio de una máquina de búsqueda. Las busquedas de los usuarios referencian directamente a los clusters que son relevantes para su busqueda. Existen dos tipos de text clustering, clustering jerarquico y clustering particional (G-HAC y k-means).[13]

4.2. Minería de Contenido desde el punto de vista de BD

La Web es una fuente enorme de documentos en línea que regularmente contienen datos semi-estructurados. La Extracción de información en la web se afronta de diferente manera a lo antes hecho, ahora hay que enfrentarse a un volumen extenso de documentos web, a los documentos nuevos que aparecen con periocidad y al cambio en el contenido de los documentos web. Una gran parte de los documentos o páginas web contienen datos semi-estructurados y estructurados y generalmente o siempre contienen información a través de links[14].

El objetivo principal que tiene la web content mining desde el punto de vista de BD es que busca representar los datos a través de grafos etiquetados.

La publicación de datos semi-estructurados y estructurados en la web ha crecido fuertemente en los últimos años y existe la tendencia a seguir creciendo, más sin embargo el crecimiento ha sido preponderante en las "hidden Web" [13 14] páginas ocultas, las cuales son generadas automáticamente con datos de bases de datos a través de consultas hechas por usuarios. Dichas páginas no son accesibles para los crawlers y para las máquinas de búsqueda no están a su alcance. Es así pues que existe la necesidad de crear ciertas aplicaciones o herramientas para la extracción de información de tales páginas. Para la obtención de dicha información en las web se hacen uso actualmente de los llamados "wrappers".

Los wrappers pueden ser vistos como procedimientos para extracción de contenido de una fuente particular de información

La extracción de estos datos permite otorgar valor agregado a los servicios, por ejemplo, en los comparativos de compras, meta busquedas, etc. Existen varios enfoques para la extracción de información estructurada; manual wrapper, wrapper induction y el enfoque automático [3]. El primero consiste en escribir un programa para extracción de información de acuerdo con los patrones observados en un Web site en específico. Los segundos consisten en identificar un grupo de páginas de entrenamiento y un sistema de aprendizaje generará reglas a partir de ellas, finalmente dichas reglas serán aplicadas para obtener objetos identificados dentro de páginas Web. Finalmente el método automático tiene como objetivo principal identificar patrones de las páginas web y luego usarlas para extraer información. Seguramente éste último es el método más utilizado en la actualidad para extraer información de la Web.

5. Web Structure Mining

De acuerdo con WangBin [**¡Error! No se encuentra el origen de la referencia.**] las estructuras de links permiten otorgar mayor información que otro documento normal. La Web Structure Mining se centra principalmente en la estructura de los hiperlinks de la web, es decir interesada en la entrada y salida de links de las páginas. Los links que

apuntan a una página puede sugerir la popularidad de la misma, mientras que los links que salen de la página demuestran los tópicos o la riqueza de contenido.

Algoritmos como el PageRank y los HITS son usados con frecuencia para modelar la topología de la web.

En PageRank, cada página Web tiene una medida de prestigio que es independiente de cualquier necesidad de información o pregunta. En línea general, el prestigio de una página es proporcional a la suma de las páginas que se ligan a él. PageRank es un valor numérico que representa lo importante que es una página en la web. Para Google, cuando una página(A) enlaza a otra(B), es como si la página(A) que tiene el enlace, votara a la página enlazada(B). Mientras más votos tenga una página, más importante será la página. También, la importancia de la página que vota determina lo importante que es el voto. Google calcula la importancia de una página a partir de los votos que obtiene. En el cálculo del PageRank de una página se tiene en cuenta lo importante que es cada voto.

HITS (Hyperlink.induced topic research) es un algoritmo que interactivo que tiene como finalidad excavar el grafo de la Web para identificar “hubs” y “authorities”. Entendemos como authorities a las páginas que de acuerdo a un topico son las que mejor posicionadas están. Los hubs son aquellas páginas que hacen liga hacia las authorities. El número y el peso de hubs apuntando a una página determina el nivel de posicionamiento.

6. Web Usage Mining

Los logs que se generan constantemente en los servidores debido a los requerimientos de los usuarios, generan un gran volumen de datos provenientes de dichas acciones.
Recientemente este gran volumen de información relevante empezó a usarse para obtener datos estadísticos, analizar accesos inválidos y para analizar problemas que se produjeran en el servidor.
Los datos almacenados en los logs siguen un formato standard. Una entrada en el log siguiendo este formato contiene entre otras cosas, lo siguiente: dirección IP del cliente, identificación del usuario, fecha y hora de acceso, requerimiento, URL de la página accedida, el protocolo utilizado para la transmisión de los datos, un código de error, agente que realizó el requerimiento, y el número de bytes transmitidos. Esto es almacenado en un archivo de texto separando cada campo por comas (",") y cada acceso es un renglón distinto

Association Rules, Sequential Patterns y Clustering ó Clasificación son algunas de las técnicas de data mining que se aplican en los servidores web.

6.1 Association Rules

La Association Rules juega un papel muy importante en el contexto de la nueva visión de la web, es decir con el auge de las técnicas de comercio que se manejan de forma electrónica permiten el desarrollo de estrategias voraces de marketing.
Normalmente esta técnica está relacionada con el uso de Bases de Datos transaccionales, donde cada transacción consiste en un conjunto de ítems. En este modelo, el problema consiste en descubrir todas las asociaciones y correlaciones de ítems de datos donde la presencia de un conjunto de ítems en una transacción implica la presencia de otros ítems.
Está técnica generalmente está asociada con en el número de ocurrencias de los ítems dentro del log de transacciones[15], por lo tanto, podemos identificar la cantidad de usuarios que acceden a determinadas páginas (60% de los clientes que acceden a la página con URL /company/products/, también acceden a la página /company/products/product1.html). Por otro lado nos permite mejorar considerablemente la estructura de nuestro site, por ejemplo, si descubrimos que el 80% de los clientes que acceden a /company/products y /company/products/file1.html también acceden a /company/products/file2.html, parece indicar que alguna información de file1.html lleva a los clientes a acceder a file2.html. Esta correlación podría sugerir que ésta información debería ser movida a /company/products para aumentar el acceso a file2.html.

6.2 Sequential Patterns

En general en las Bases de Datos transaccionales se tienen disponibles los datos en un período de tiempo y se cuenta con la fecha en que se realizó la transacción; la técnica de sequential patterns se basa en descubrir patrones en los cuales la presencia de un conjunto de ítems es seguido por otro ítem en orden temporal.
En el log de transacciones de los servidores de Web, se guarda la fecha y hora en la que un determinado usuario realizó los requerimientos. Analizando estos datos, se puede determinar el comportamiento de los usuarios con respecto al tiempo.
Con esto, se puede determinar por ejemplo:
<60% de los clientes que emitieron una orden on-line en /company/products/product1.html, también emitieron una orden on-line en /company/products/product4.html dentro de los siguientes 15 días.
El descubrimiento de sequential patterns en el log puede ser utilizado para predecir las futuras visitas y así poder organizar mejor los accesos y publicidades para determinados períodos. Por ejemplo, utilizando está técnica se podría descubrir que los días laborables entre las 9 y las 12 horas muchas de las personas que accedieron al servidor lo hicieron para ver las ofertas y en los siguientes días la mayoría compró productos. Entonces por la mañana debería facilitarse el acceso a las ofertas y brindar la publicidad más llamativa posible.
También puede ser utilizado para descubrir tendencias, comportamiento de usuarios, secuencias de eventos, etc. Esta información puede ser aprovechada tanto en el aspecto comercial (pensar una campaña de marketing) como en el aspecto técnico (mejorar los tiempos de acceso).

En general todas las herramientas que realizan mining sobre el log enfocan el análisis sobre secuencias de tiempo ya que los eventos que son almacenados están muy relacionados con el tiempo en que se producen.

6.3 Clustering

Las técnicas de clasificación permiten desarrollar un perfil para los ítems pertenecientes a un grupo particular de acuerdo con sus atributos comunes. Este perfil luego puede ser utilizado para clasificar nuevos ítems que se agreguen en la base de datos.
En el contexto de Web Mining, las técnicas de clasificación permiten desarrollar un perfil para clientes que acceden a páginas o archivos particulares, basado en información demográfica disponible de los mismos. Esta información puede ser obtenida analizando los requerimientos de los clientes y la información transmitida de los browsers incluyendo el URL.
Utilizando técnicas de clasificación, se puede obtener, por ejemplo, lo siguiente:
50% de los clientes que emiten una orden on-line en /company/products/product2.html, están entre 20 y 25 años y viven en la costa oeste.
La información acerca de los clientes puede ser obtenida del browser del cliente automáticamente por el servidor; esto incluye los accesos históricos a páginas, el archivo de cookies, etc. Otra manera de obtener información es por medio de las registraciones y los formularios on-line.
La agrupación automática de clientes o datos con características similares sin tener una clasificación predefinida es llamada clustering.

7. Conclusiones

La Web Mining ha despertado gran interés en la actualidad, particularmente debido a los avances de la comunidad científica en distintas líneas de investigación relacionadas con Data Mining orientado a la www. La web mining en los últimos años esto se ha potenciado fuertemente en virtud del gran aumento en volumen del tráfico, tamaño y complejidad de las fuentes de información disponibles en la Web y el reciente interés en el desarrollo aplicaciones para el comercio electrónico. La iniciativa privada actualmente es el principal precursor de que la información sea particular para cada individuo creando sistemas que incorporan personalización construyen modelos de los objetivos, características, preferencias y conocimientos de cada usuario. La web mining es un área con expectro amplio de investigación. En este documento se hace una simple aproximación para tener ciertas bases en proyectos profundos de investigación.

8. Referencias

1. Baeza-Yates, R. Castillo, C. Marin, M. and Rodríguez, A. “Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering”, WWW Conference / Industrial Track, ACM, pp. 864-872. Chiba, Japan, 2005.

2. Baeza-Yates, R. Excavando la web. El profesional de la información. v13, n1, 2004

3. Liu, B. and Chen-Chuan Chang, Kevin. Editorial: “Special Issue on Web Content Mining”. WWW 2005 Tutorial, Page 1-4, 2005.

4. Scotto, M. Sillitti, A. Succi, G. Vernazza, T. “Managing Web-Based Information”, *International Conference on Enterprise Information Systems (ICEIS 2004)*, Porto, Portugal, April 2004. Page 1-3

5. Etzioni O. “The World Wide Web: quagmire or gold mine?” , In Communications of the ACM 39(11). 1996

6. www.wikipedia.org

7. Henzinger M. “**Web Information Retrieval an Algorithmic Perspective”,** European Symposium on Algorithms, p 1-3, 2000

8. Baeza-Yates, R. Pobrete, B. “Una herramienta de minería de consultas para el diseño del contenido y la estructura de un sitioWeb” Actas del III Taller Nacional de Minería de Datos y Aprendizaje TAMIDA2005”, pp.39-48, 2005

9. Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. Page 1-9, 2000.

10. Galeas, P. Web Mining by Patricio Galeas. http://www.galeas.de/webmining.html

11. Wilks, Y. **Information Extraction as a Core Language Technology Source** Lecture Notes In Computer Science; Vol. 1299 Pages: 1 – 9, 1997

12. J. Wang, Y. Huang, G. Wu, and F. Zhang, Web Mining: Knowledge Discovery on the Web *Proc. Int'l Conf. Systems, Man and Cybernetics (SMC '99),* vol. 2, pp. 137-141

13. Wang, J. Huang, Y. Wu, G. and F. Zhang, "Web Mining: Knowledge Discovery on the Web" *Proc. Int'l Conf. Systems, Man and Cybernetics (SMC '99),* vol. 2, pp. 137-141, 1999.

14. Eikvil, L. "Information Extraction from World Wide Web - A Survey", *Rapport* Nr. 945, July, 1999. ISBN 82-539-0429-0

15. http://www2.ing.puc.cl/gescopp/Sergio_Maturana/SAG/Webmining.html

16. Wang, B. Liu, Z. "Web Mining Research," Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03), *iccima*, p. 84, 2003.