CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO

Minería de Datos

SERIE GRIS

RNPS No. 0552
ISSN Solicitado

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Minería de Datos

SERIE GRIS

Estado del Arte del Web

José E. Medina Pagola

RT _ 001 **Agosto _ 2007**

RNPS No. 0552
ISSN Solicitado

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Estado del Arte del Web Mining

José E. Medina Pagola

Centro de Aplicaciones de Tecnología de Avanzada (CENATAV), 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
jmedina@cenatav.co.cu

RT_001 CENATAV
Fecha del camera ready: 11 de febrero de 2005

Resumen. En este trabajo se presenta una estructuración y descripción de los tipos de Minería de Web, junto a algunos de sus principales tipos de aplicaciones. Se abordan, además, algunas de las técnicas y métodos propuestos, y se analizan diversas tecnologías asociadas con este tipo de minería. Como resultado de este análisis se incluye un listado de los especialistas consultados, junto a sus instituciones, direcciones de correo y páginas personales. Por último, se incluye una relación de eventos científicos que abordarán la Minería de Web en los años 2005 y 2006.

Palabras clave: Minería de web; Sistemas multiagentes; Minería de texto, Ontologías, Bibliotecas digitales, Filtraje cooperativo, Recuperación de informaciones, Extracción de informaciones.

Abstract. In this work, a structuring and description of Web Mining types are presented, with some of their main types of application. The proposed techniques and methods are also presented, and many technologies associated with this kind of mining are analyzed. As a result of this analysis, a list of the consulted specialists, with their institutions, e-mail and home pages, is included. Finally, several scientific events in 2005 and 2006, which will include Web Mining, are included.

Keyword: Web mining; Multiagent systems; Text mining; Ontology, Digital libraries, Collaborative filtering, Information retrieval, Information extraction.

1 Introducción

El crecimiento explosivo de la WWW (*Word Wide Web*) ha generado una enorme cantidad de información almacenada en muy diversas fuentes y formatos. Este crecimiento ha estado aparejado con una popularidad tal que este medio resulta un recurso imprescindible de la sociedad humana, observándose usuarios de muy disímiles profesiones, y con muy diversos fines, accediendo a capacidades nunca antes vistas con la esperanza de lograr informaciones y servicios sin precedentes en la historia de la humanidad. Sin embargo, no siempre tales esperanzas han sido alcanzadas a plenitud.

Con el propósito de aprovechar ese cúmulo informativo y tratar de satisfacer esas expectativas, se han estado desarrollando y perfeccionando diferentes herramientas de cómputo que permiten encontrar, extraer, filtrar, y evaluar las informaciones requeridas. Sin embargo, muchas de esas informaciones representan patrones y conocimientos, no pocas veces ocultos en toda esa telaraña de objetos informáticos dispersos e interconectados por todo el mundo.

Prácticamente, desde los inicios de la Internet se observaron estas posibilidades. Por ejemplo, ya en 1995 Bray encontró patrones interesantes en millón y medio de documentos, mostrando las complejidades que la Web ofrecía [Bray, 1996].

Todos estos factores han propiciado la creación de sistemas sobre la Web que permiten la conformación y manipulación de todos esos patrones y conocimientos disponibles en las redes globales y locales para su mejor comprensión y aprovechamiento, y así lograr mejores servicios.

Para la conformación de tales sistemas se han desarrollado diferentes técnicas y tecnologías englobadas bajo la denominación de Minería de Web.

En este trabajo se abordan tales técnicas y tecnologías, tratando de analizar algunas de las principales líneas de investigación actuales, presentando algunos de los especialistas, instituciones y trabajos vinculados con esta temática, así como los principales eventos relacionados con ella de los años 2005 y 2006.

2 Minería de Web

La Minería en la Web o Minería de Web[1] (*Web Mining*) es un término relativamente nuevo en el argot informático; por ello, no es de extrañar que no exista una definición precisa y de aceptación generalizada por los especialistas del tema.

Una definición de Minería de Web dada por Tingshao Zhu en su tesis de doctorado es la siguiente [Zhu, 2003]:

> *"La Minería de Web es la aplicación de la minería de datos o de otras técnicas de procesamiento de informaciones a la WWW, para encontrar patrones útiles que se espera ayuden a las personas a acceder a la WWW más eficientemente."*[2]

Como puede observarse, aunque esta definición relaciona los términos de "minería de datos" y la "WWW" como elementos distintivos, presenta imprecisiones como la expresión, un tanto ambigua, de "otras técnicas de procesamiento de informaciones". Algo similar ocurre con la referencia a que "ayuden a las personas a acceder a la WWW más eficientemente", lo cual depende objetivamente del tipo de aplicación y los intereses de los que la apliquen, aunque esta referencia fue suprimida por el autor en la definición que se observa en su página Web [Zhu, 2004].

Otra definición, referida como fuente por el propio Zhu, es la dada por Bamshad Mobasher en 1996, cuando planteó que [Mobasher, 1996]:

> *"La Minería de Web es la aplicación de las técnicas de la minería de datos a grandes repositorios de datos de la Web"*[3]

En donde ofrece una definición algo más precisa, aunque posteriormente en el mismo artículo restringe el "objetivo primario de la Minería de Web" al acceso de las páginas Web y a un servidor particular.

Quizás, una de las definiciones mejor formulada, a pesar del tiempo transcurrido, es la dada por Oren Etzioni [Etzioni, 1996], y referenciada por Kosala y Blockeel [Kosala, 2000], cuando planteó que:

> *"La Minería de Web es el uso de las técnicas de minería de datos para el descubrimiento y extracción automática de informaciones de documentos y servicios de la Word Wide Web."*[4]

Como se observa, bajo esta definición se tienen una serie de aplicaciones y problemáticas como son las siguientes [Kosala, 2000], [Han, 2000], [Rojo, 2002], [Yao, 2003], [Kawamae, 2004], [Park, 2004], [Liu, 2004], [Gleich, 2004], [Velásquez, 2004]:

- Identificar las páginas Web de mayor interés.
- Clasificar los documentos de la Web.
- Clasificar los resultados de los motores de búsqueda de la Web.
- Identificar y recuperar eficientemente los documentos en la Web.
- Encontrar regularidades en los *Logs* de Web.
- Responder inteligentemente a solicitudes de búsqueda en la Web.
- Personalizar las informaciones que se muestran en la Web.
- Identificar los intereses, preferencias e intenciones de los usuarios de una empresa o Intranet.
- Sugerir términos de búsqueda, como ocurre en el mercado de búsqueda "pay-for-performance" tipo "Overture".
- Recomendar visitas a páginas Web según perfiles de usuarios.
- Asistir a los investigadores en las búsquedas en la Web.
- Encontrar informaciones interesantes en bases de datos públicas.

Algunas de estas problemáticas serán presentadas en mayores detalles en los siguientes epígrafes.

---

[1] También se ha encontrado la expresión "Minería Web", aunque se considera una incorrecta traducción de su equivalente en inglés.

[2] En inglés: *"Web mining is the application of data mining or other information process techniques to WWW, to find useful patterns which are expected to help people access WWW more efficiently."*

[3] En inglés: *"Web mining is the application of data mining techniques to large Web data repositories."*

[4] En inglés: *"Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services."*

Para la tipificación de las aplicaciones y problemáticas de la Minería de Web se han dado diversas clasificaciones.

Algunos autores han considerado dos tipos de Minería de Web, siendo estas: Minería de Contenido de Web y Minería de Uso de Web [Cooley, 1997], [Xu, 2003]. Sin embargo, la mayoría de los especialistas consideran tres los tipos de Minería de Web, atendiendo al tipo de información a minar [Kosala, 2000], [Zhu, 2004], los cuales son:

- Minería de Estructura de Web.
- Minería de Uso de Web.
- Minería de Contenido de Web.

A continuación se expondrán en mayores detalles cada uno de estos tipos.

2.1 Minería de Estructura de Web

La Minería de Estructura de Web (*Web Structure Mining*) trata de descubrir los patrones que subyacen en la estructura y topología de los enlaces de la Web, con el propósito de identificar preferencias y clasificaciones de los objetos relacionados [Kosala, 2000], [Zhu, 2004].

Generalmente, la Web es modelada como un grafo orientado, cuyos vértices representan las páginas y sus aristas los enlaces entre ellas. La relevancia de las páginas es evaluada en conjunto con las páginas adyacentes. Para ello, se toman diferentes medidas, como son: el grado de entrada (la cantidad de enlaces externos hacia la página), el grado de salida (cantidad de enlaces internos hacia otras páginas), etc.

Evidentemente, una forma simple de evaluar la relevancia de una página sería considerando su "popularidad" a partir de su grado de entrada [Davison, 2003]. Otra forma más elaborada es mediante el método PageRank utilizado en Google [Brin, 1998].

En el método PageRank, desarrollado por Larry Page y Sergey Brin, las páginas con mayores grados de entrada tienen valores de relevancia más altos. Sin embargo, esto es matizado con las relevancias de las páginas. De esta forma, para el cálculo de la relevancia (PR - PageRank) de una página T (PR_T) se consideran los valores asociados con las n páginas ($t1$, …, tn) desde las que se acceden a ella [Craven, 2004].

$$PR_T = (1-d) + d * (PR_{t1}/C_{t1} + \cdots + PR_{tn}/C_{tn}) \qquad (\mathbf{1})$$

Cada parámetro C_i representa la cantidad de enlaces de la i-ésima página y d representa la probabilidad de salir y saltar hacia otra página. Esta expresión presupone un proceso convergente de los PageRank de las páginas adyacentes.

Otra forma de evaluar la relevancia de una página es mediante métodos tipo HITS [Kleinberg, 1998], [Chakrabarti, 2004]. En los métodos tipo HITS (*Hyperlink Induced Topic Search*) se genera un subgrafo con las páginas que satisfacen un criterio indicado en algún motor de búsqueda, junto a las páginas citadas por ellas o que estas citan. Estas páginas pueden caracterizarse como Autoridades y Concentradoras. Las páginas Autoridades (*Authorities*), presentan las mejores informaciones de un tema y son las más requeridas por los usuarios. Las páginas Concentradoras (*Hubs*), son las que contienen múltiples enlaces a páginas Autoridades [Chakrabarti, 2000].

En la implementación del método, con cada página T del subgrafo se tiene asociado un valor de Concentrador (H_T) y uno de Autoridad (A_T). Estos valores son inicializados con valores no negativos, iterándose hasta que converjan las expresiones [Davison, 2003].

$$\begin{aligned} H_T &= \sum A_{ti} \quad \text{donde } T \rightarrow ti \\ A_T &= \sum H_{ti} \quad \text{donde } ti \rightarrow T \end{aligned} \qquad (\mathbf{2})$$

Diversas son las variaciones realizadas sobre estos dos métodos básicos, resaltándose en algunos de los trabajos consultados el alto costo computacional del método HITS. Ejemplos de esas propuestas son las variaciones de HITS de Cohn y Hofmann [Cohn, 2001], los métodos derivados de PageRank propuestos por Rafiei y Mendelzon [Rafiei, 2000], y por Richardson y Domingos [Richardson, 2002].

Otros tipos de métodos y aplicaciones pueden observarse asociados con la minería de uso y de contenido de Web, como se expondrá en los siguientes epígrafes.

2.2 Minería de Uso de Web

La Minería de Uso de Web (*Web Usage Mining*) utiliza las técnicas de minería de datos para descubrir patrones de uso a partir de los *Logs* de acceso a la Web, para comprender y mejorar los requerimientos de las aplicaciones en este medio [Zhu, 2003]. O sea, a diferencia de los otros tipos de Minería de Web, en este tipo se analizan los datos secundarios de las interacciones de los usuarios, tales como los *Logs* de acceso en servidores, *Browsers* (Buscadores), datos de sesiones, de *Cookies*, de registros y transacciones de los usuarios, etc.

Existe una tendencia generalizada en dividir este tipo de minería en dos grupos, siendo estos [Kosala, 2000], [Han, 2000], [Zhu, 2003]:

- Descubrimiento de patrones de acceso general.
- Descubrimiento de patrones de uso personalizado.

2.2.1 Descubrimiento de patrones de acceso general

En esta clasificación se incluyen diversos tipos de aplicaciones y técnicas que tratan de comprender los patrones y tendencias de los ficheros *Logs*. Estos ficheros contienen, al menos, la dirección IP del usuario, la fecha y hora de la solicitud, y la URL de la página requerida. Con esta información es posible reconstruir las sesiones de navegación de los usuarios; o sea, las páginas visitadas por un usuario en un tiempo dado [Berent, 2001].

Las sesiones de navegación pueden ser procesadas y modeladas de diversas formas. Un método usual es generando tablas relacionales, tratándolas por técnicas estándares [Mobasher, 2002]. Otra forma de evaluar la información disponible es procesando directamente los ficheros *Logs* [Spiliopoulou, 2001], [Borges, 2000].

En la generalidad de los casos, las sesiones de navegación son tratadas como cadenas o secuencias de páginas, o como caminos de un grafo de navegación, o trazas, rememorando el término empleado por Vannevar Bush, el precursor de los hipertextos, cuando hablaba de la "telaraña de trazas" (*Web of Trails*) [Bush, 1945].

Con el análisis de las informaciones de los *Logs* y las secuencias o trazas pueden lograrse resultados tales como:

- Encontrar las trazas más probables. Por ejemplo, si se descubre que el ochenta y cinco por ciento de los clientes que acceden a /productos/noticias.html acceden también a /productos/historias_suceso.html, puede deberse a determinadas noticias interesantes de la empresa, lo que permitiría el rediseño de las páginas [Molina, 2002].
- Encontrar, entre las trazas más probables, las más largas, evitando lo conocido como "pérdidas de hiperespacio" entre los usuarios [Wheeldon, 2003], [Oyanagi, 2003].
- Descubrir grupos de usuarios, páginas o sesiones con intereses comunes [Cadez, 2000], [Zhu, 2003].

Diferentes algoritmos han sido propuestos para el procesamiento de estas informaciones. Muchos de ellos aplican técnicas estándar de minería de datos, tales como los métodos tipo Apriori [Agrawal, 1994], DHP [Park, 1997] o PatriciaMine [Pietracaprina, 2003] para la inducción de reglas de asociación, los asociados a la búsqueda de agrupamientos, etc. Otros representan desarrollos específicos para estas tareas, tales como las observadas en el sistema WebMiner que descubre automáticamente reglas de asociación y patrones secuenciales en los *Logs* de acceso [Cooley, 1997]; en los métodos basados en clasificaciones, agrupamientos y análisis de secuencias [Baglioni, 2003], como el MDSAM (Multidimensional Sequence Alignment Method) [Hay, 2003]; en la búsqueda de asociaciones indirectas negativas vinculadas por subtrazas mediadoras [Tan, 2002]; el *Best Trail Algorithm* [Wheeldon, 2003] o el Clustering Matricial de Shigeru Oyanagi *et al*. [Oyanagi, 2003], para encontrar las trazas más probables y largas; aplicaciones de las Cadenas de Markov a las trazas en los *Logs* [Cadez, 2000], [Davison, 2004]; o modelando las trazas mediante Gramáticas Probabilísticas de Hipertextos o HPG (*Hypertext Probabilistic Grammar*), las cuales representan un tipo especial de gramática regular donde los nodos del autómata asociado son los URL visitados [Borges, 2004].

2.2.2 Descubrimiento de patrones de uso personalizado

El descubrimiento de patrones de uso personalizado abarca aplicaciones que modelan los perfiles de los usuarios para mejorar los servicios, personalizar las interfaces e informaciones ofrecidas en la Web, o perfeccionar las posibilidades inherentes al comercio electrónico. Estas posibilidades se logran identificando los objetivos intrínsecos en las sesiones de navegación y descubriendo las relaciones semánticas ocultas entre los usuarios, y entre estos y los objetos de la Web [Zhou, 2004].

La personalización (*Personalization* o *Profiling*) de la Web, según Bamshad Mobasher, se puede analizar a partir de tres tipos de sistemas: los Sistemas de Reglas de Decisión Manual, los Sistemas de Filtraje Basados en el Contenido, y los Sistemas de Filtraje Cooperativo [Mobasher, 2004 a].

Los Sistemas de Reglas de Decisión Manual y los de Filtraje Basados en el Contenido consideran las informaciones de los perfiles obtenidas durante las actividades de registro manual de los usuarios, junto a conocimientos de apoyo manipulados por los administradores de los sitios Web y, en el segundo caso, de diversas informaciones obtenidas de múltiples fuentes. Estos tipos de estrategias son utilizadas por la mayoría de los portales y sitios, como los de comercio electrónico, que incluyen registros de usuarios. Ejemplo de estos, citados por Mobasher, es la suite Broadvision[5] y el WebWatcher de Thorsten Joachims *et al.*[6].

A diferencia de los sistemas anteriores que actúan sobre informaciones estáticas ofrecidas por los usuarios durante su registro, los Sistemas de Filtraje Cooperativo tratan de captar "al vuelo" los intereses y preferencia de los usuarios para perfeccionar las predicciones o recomendaciones que se realicen.

El Filtraje Cooperativo (*Collaborative Filtering*), una forma de filtraje de información, permite recomendar objetos preferidos por usuarios similares o predecir la utilidad de ciertos objetos para un usuario particular. Las preferencias utilizadas, generalmente, son representadas por una evaluación numérica, las cuales pueden obtenerse de forma explícita o derivarse de los registros de compra, de los tiempos de acceso a las páginas Web, etc. [Nakamura, 2003], [Ziqiang, 2004].

Muchos sitios Web han incorporado mecanismos de recomendación basados en el Filtraje Cooperativo, como sucede con Amazon, el mayor portal de ventas de bienes en línea, Barnes and Noble, una librería virtual, o Netflix, que ofrece servicios de alquiler de DVD. En particular, en el comercio electrónico se ha observado un polémico movimiento del *Direct Marketing* al *Permission Marketing*; o sea, a la mercadotecnia en la que se le ofrece a los usuarios informaciones que les gusten, como noticias, facilidades de descarga de materiales de interés, participación en sorteos o eventos que le motiven, etc. [Campuzano, 2002]. Lógicamente, este movimiento no es concebible sin una adecuada personalización de los intereses de los usuarios. En general, la aplicabilidad de estas técnicas son tan diversas que se observan trabajos desde el comercio electrónico hasta propuestas asociadas a la enseñanza a distancia [Hösch, 2005].

En los sistemas de personalización de la Web se han desarrollado y aplicado una gran diversidad de algoritmos. De hecho, la generalidad de los métodos aplicados en el descubrimiento de patrones generales puede utilizarse en el descubrimiento de patrones de uso personalizado. Ejemplo de ello es el empleo de las Cadenas de Markov en la predicción de accesos, como ocurre con el método Hybrid-order Tree-like Markov Model (HTMM) propuesto por Dongshan y Junyi [Dongshan, 2002].

Otros métodos aplicados en esta clase de aplicación son los de Aprendizaje Basado en Instancias (IBL - *Instance Based Learning*). En estos métodos se buscan los usuarios con similares patrones de acceso, preferencias o valoraciones de los objetos de la Web mediante métricas de semejanza y se evalúan (probablemente de forma pesada) esas preferencias y valoraciones para recomendarle posibles opciones al usuario [Heylighen, 2001] [Zhu, 2005].

También se han propuesto técnicas de clasificación, como son las conocidas como SVM (*Support Vector Machines*), considerando las votaciones emitidas por los usuarios sobre los objetos de la Web [Calderón, 2004]. Otra propuesta para generar las recomendaciones de las páginas a visitar consideran Modelos de Variables Latentes, tales como el Análisis de Componentes Principales (*Principal Component Analysis*) y el Análisis de Factores Principales

[5] URL: http://www.broadvision.com
[6] URL: http://www-2.cs.cmu.edu/afs/cs/project/theo-6/web-agent/www/project-home.html

(*Principal Factor Analysis*), como la realizada por Yanzan Zhou *et al*. mediante el método IPF (*Iterative Principal Factor*) [Zhou, 2004]. En otras propuestas se han sugerido técnicas de inducción de reglas de producción, tipo C4.5 o CART, para generar automáticamente reglas de recomendación, aunque se observa la dificultad en la evaluación de las medidas de interés [Zhu, 2003].

Otras técnicas empleadas en el Filtraje Cooperativo son los métodos de agrupamiento (*Clustering*). Usualmente, mediante los algoritmos de agrupamiento se segmentan los ítems pertenecientes a los usuarios, utilizándose cada partición encontrada en las recomendaciones y predicciones [O'Connor, 1999], [Rodríguez, 2003]. Un ejemplo de la aplicación de estas técnicas es el trabajo de Juan Velásquez *et al*. [Velasquez, 2004], donde se propone un SOM (*Self Organizing Map*) para la generación de los *Clusters* temáticos, definiéndose con los administradores las reglas del negocio a ser utilizadas en las recomendaciones a los usuarios.

Incluso, se han evaluado las técnicas denominadas *Stigmergic*; o sea, las que consideran estrategias auto-organizativas encontradas en las sociedades de insectos, interrelacionadas o no con otras técnicas como es la Computación Evolutiva [Ramos, 2004].

Varias de las técnicas anteriores han sido incorporadas en muchos de los sistemas existentes. Ejemplo de estos son los sistemas Analog[7], Urchin[8], Webalizer[9] y Weblog Expert[10]. Para ver un amplio listado de aplicaciones de Minería de Uso de Web puede consultarse la página "KDnuggets: Software: Web Mining and Web Usage Mining"[11].

2.3 Minería de Contenido de Web

La Minería de Contenido de Web (*Web Content Mining*) centra su atención en la búsqueda y extracción de patrones de los objetos disponibles en la Web, tales como ficheros HTML, imágenes, correos electrónicos, bases de datos en línea, etc., permitiendo el descubrimiento de palabras claves, frases, ontologías, e informaciones interesantes sobre esos objetos, facilitando la evaluación, identificación, recuperación de datos y documentos, la elaboración de sumarios, etc. [Kosala, 2000], [Han, 2000], [Zhu, 2004].

Aunque los tres tipos de Minería de Web (de Estructura, de Uso y de Contenido) son tres categorías de amplia aceptación, no existe una clara división entre ellas [Baglioni, 2003]. Esto se evidencia no solo en los múltiples trabajos consultados sino, incluso, en la estructuración de las temáticas de uno de los eventos científicos más importantes en esta área: WEBKDD, realizado en conjunción con *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* donde, en su edición del 2004, combinó estas categorías como:

- Web Usage Mining and Web Analytics.
- Web Content and Structure Mining.

Sin embargo, se incluyó en la segunda temática el subtópico: *Integration of Web content, usage, and structure data for Web mining*, como el reconocimiento explícito de tales interrelaciones [WEBKDD, 2004].

Como ha podido observarse, debido a la diversidad de tipos de datos, las problemáticas asociadas y a su interrelación con los otros tipos de minería, la clasificación de la Minería de Contenido de Web resulta una tarea difícil. No obstante, algunos especialistas han propuesto posibles subdivisiones de esta categoría. Ejemplo de ello es la estructuración dada por Kosala y Blockeel donde, bajo esta temática, aborda la Recuperación de Información (*Information Retrieval*) y las Bases de Datos [Kosala, 2000]. Por su parte, Jiawei Han en una conferencia tutorial estructuró este tipo en: Minería del contenido de las páginas Web y Minería de los resultados de búsquedas [Han, 2000]. También se puede citar a Baglioni *et al*., cuando clasifican esta minería según el tipo de documento; o sea como: Textos plano, Documentos semi-estructurados (HTML, XML, etc.), Documentos estructurados (*Digital Libraries*), Documentos dinámicos, y Documentos de multimedia [Baglioni, 2003].

---

[7] URL: http://www.analog.cx/
[8] URL: www.urchin.com/
[9] URL: http://www.mrunix.net/webalizer/
[10] URL: http://www.weblogexpert.com/
[11] URL: http://www.kdnuggets.com/software/web.html

Como puede observarse, no existe un consenso en la clasificación de la Minería de Contenido de Web. Por ello, no se asumirá una en particular. No obstante, en los siguientes epígrafes se analizarán diferentes tipos de aplicaciones y problemáticas afines.

2.3.1 Recuperación de información

La Recuperación de Información (IR - *Information Retrieval*) es un término aplicable a la búsqueda y recuperación de informaciones, sobre un asunto determinado, en colecciones de documentos y textos; o sea, en informaciones no estructuradas o semi-estructurada. Estas técnicas fueron aplicadas en los últimos 20 años en la indexación de documentos, pasando por sus aplicaciones a la Multimedia e Hipertextos hasta, en la actualidad, las que se realizan sobre la WWW [Baeza-Yates, 2001]. Las principales actividades de la IR son: la modelación del conocimiento a indexar, la clasificación y categorización de los documentos, el filtraje de las informaciones, y el perfeccionamiento de las interfaces usuarias. De estas actividades, la clasificación y categorización presentan una evidente relación con las técnicas de Minería de Web de Contenido, permitiendo con ello mejores indexaciones [Kosala, 2000].

Es comprensible que para recuperar información en la Web se precisan de sistemas de búsqueda eficientes. Según Xu, Huang y Madey, existen dos tipos principales de instrumentos de búsqueda en la Web: los Directorios, como Yahoo, Netscape, etc., y los Motores de Búsqueda, como Lycos, Google, etc. [Xu, 2003]. Los sistemas tipo directorio están orientados a las búsquedas interactivas, por lo que, en general, precisan de una eficiente y rápida organización y clasificación de las informaciones que muestran y recomiendan.

En la literatura, los Motores de Búsqueda (*Search Engine*), o simplemente Buscadores, suelen diferenciarse de los llamados Rastreadores de la Web, aunque estos se consideren un elemento esencial de los primeros. Los Rastreadores de la Web (*Web Crawlers*), conocidos también como Arañas (*Spiders*), Robots, Gusanos (*Worms*), etc., son programas que viajan automáticamente a través de los sitios Web, bajando documentos, navegando por los enlaces, manteniendo copias de las páginas visitadas y facilitando la indexación de esas páginas. Un tipo particular de Buscador son los denominados como Meta-buscadores, los cuales toman las salidas de otros Motores de Búsqueda para filtrar y generar una única salida. Una de las problemáticas que deben resolver los Meta-buscadores es la fusión de diferentes tipos de ranking [Díaz, 2004].

Un tipo de sistema vinculado estrechamente con las técnicas de IR son las denominadas Bibliotecas Digitales. Las Bibliotecas Digitales (*Digital Libraries*) es un término que se relaciona con las tecnologías que permiten el acceso a recursos académicos (e investigativos) heterogéneos en bases de datos bibliográficas, catálogos, revistas electrónicas, servidores de documentos y páginas Web [Friedrich, 2004]. Un ejemplo de este tipo de biblioteca es el proyecto de Biblioteca Digital de Nueva Zelandia desarrollado por la Universidad de Waikato utilizando el software *Open-source* conocido como *Greenstone* [12].

La Recuperación de Informaciones en las Bibliotecas Digitales, así como en la mayoría de los sistemas de este tipo, se encuentra estrechamente vinculado con el conocimiento asociado e incluso, como se indicará en el siguiente epígrafe, con el descubrimiento automático de tales conocimientos. Ejemplo de esto es el trabajo de Fanguy y Raghavan, en donde se propone la creación de sistemas de recuperación de información basados en conceptos, permitiéndoles a los usuarios la selección de conceptos y definiciones. Estos conceptos y definiciones, además de que puedan ser especificados por expertos, se sugiere su generación de forma automática a partir de árboles de decisión [Fanguy, 2003].

Como ya fue planteado, los diferentes tipos de aplicaciones asociadas con la Minería de Contenido de Web pueden realizarse en conjunción con otras técnicas y tipos de Minería de Web. La Recuperación de Informaciones, como pudo observarse en la propuesta de Fanguy y Raghavan, no está exenta de estas posibilidades. Otro ejemplo de ello es el trabajo de Kawamae y Takahashi, en donde se trata el problema de la recuperación de documentos de interés basado en el Filtraje Cooperativo a partir del concepto de Clases Latentes (*Latent Classes*) para construir las interrelaciones entre los usuarios y objetos [Kawamae, 2004].

[12] URL: http://www.sadl.uleth.ca/nz/cgi-bin/library

2.3.2 Extracción de información

La Extracción de Información (IE - *Information Extraction*) es una forma de aprendizaje o descubrimiento de patrones textuales a partir de informaciones no estructuradas o semi-estructuradas, como se observan en la Web [Soderland, 2004].

Las técnicas y aplicaciones de Extracción de Informaciones presentan una clara interrelación con las observadas en la Minería de Texto. El concepto de Minería de Texto (*Text Mining*) se aplica a los procesos de descubrimiento de patrones interesantes y nuevos conocimientos en colecciones de textos; o sea, a la extensión de la Minería de Datos aplicada a los documentos para descubrir informaciones no contenidas en ninguno en específico [Montes y Gómez, 2002].

Las técnicas de Minería de Texto suelen actuar sobre secuencias de términos obtenidos a partir de un pre-procesamiento de los documentos. Los términos obtenidos pueden representarse de diferentes formas, generalmente en forma de grupos o "bultos" (*bags*) de términos almacenados casi siempre según el modelo vectorial [Raghavan, 1986]. Como puede observarse, en esta representación no se tiene en cuenta la secuencia en que aparecen los términos, ni sus relaciones sintácticas; o sea, se consideran como unigramas con una supuesta independencia de sus ocurrencias. Los valores de tales vectores suelen representar pesos, los cuales pudieran asumir, entre otras, las siguientes interpretaciones [Pons, 2004]:

- Booleana – Cada términos se asocia con un valor booleano representando su presencia o no en los documentos.
- Frecuencia de Términos (TF - *Term Frequency*) – Cada término se hace corresponder con la frecuencia (absoluta o normalizada) con que aparece en los documentos.
- TF-IDF – Cada término se corresponde con un valor denominado TF-IDF (*Term Frequency – Inverse Document Frequency*), el cual no sólo considera la frecuencia de los términos en un documento sino que, además, se ajusta por el inverso de la cantidad de documentos que lo contienen. Dentro de este tipo de medida existen diferentes variaciones, tales como: el pesado *ltc*, el pesado de *Okapi tf*, etc.

Estos vectores de términos son utilizados, entre otras tareas, para analizar semejanzas entre documentos o grupos de ellos, usando diferentes medidas. Aunque se han propuesto diferentes medidas como la de Jaccard o Dice [Zhong, 2003], una de las más utilizadas es la del Coseno, definida como [Pons, 2004]:

$$sim(d_i, d_j) = \cos(d_i, d_j) = \frac{(d_i \bullet d_j)}{\|d_i\| * \|d_j\|} = \frac{\sum w_{ir} * w_{jr}}{\sqrt{\sum {w_{ir}}^2 * {w_{jr}}^2}}, \qquad (3)$$

donde d_i, d_j son los vectores de los documentos i, j; $||d_i||$, $||d_j||$ las normas de esos vectores; y w_{ir}, w_{jr} los pesos de los términos de los vectores d_i, d_j, respectivamente.

Aunque el tratamiento vectorial, propuesto inicialmente por Salton [Salton, 1971], ha sido el dominante en la literatura consultada, algunos autores han considerando otras representaciones y dimensiones semánticas u ontológicas de los documentos.

Una alternativa al tratamiento semántico del contenido de los documentos es el empleo de Mapas Conceptuales para su representación, agrupamiento y tratamiento en general [Montes y Gómez, 2002], [Simón, 2004]. Otra forma de incluir una dimensión ontológica son los métodos basados en corpus, junto a taxonomías léxicas, para calcular semejanzas semánticas entre palabras/conceptos.

Ejemplos de estos métodos son aquellos desarrollados sobre la taxonomía de amplio cubrimiento conocida como Wordnet [Budanitsky, 2001].

Alternativas interesantes al modelo del espacio vectorial son los modelos lingüísticos. Estos modelos, en general, consideran las probabilidades de ocurrencia de las frases S en un lenguaje M, indicado por $P(S/M)$. Un ejemplo de este tipo de modelo es la propuesta de Kou y Gardarin [Kou, 2002]. En esta propuesta se considera la semejanza entre dos documentos según la siguiente expresión:

$$sim(d_i, d_j) = d_i \bullet d_j = \sum_r w_{ir} w_{jr} + \sum_r \sum_{s \neq r} w_{ir} w_{js} (t_r \bullet t_s), \qquad (4)$$

donde w_{ir} y w_{js}, usando la terminología de Kou-Gardarin, son los pesos de los términos en los vectores d_i, d_j, respectivamente, y $(t_r \bullet t_s)$ es la correlación a priori entre los términos t_r y t_s. Los autores proponen estimar esas correlaciones mediante un proceso de entrenamiento. Puede notarse que esa expresión puede reducirse a la medida del Coseno (normalizado por el largo de los vectores) si se considera la independencia de los términos.

Todas estas medidas analizadas anteriormente son variantes del modelo del Espacio Vectorial Generalizado (*Generalized Vector Space Model*) propuesto por Wong *et al.* [Wong, 1985].

Entre los objetivos de la Minería de Texto se tienen la descripción semántica del contenido de uno o varios documentos, la realización de sumarios, descubrir regularidades, tendencias, desviaciones, asociaciones, patrones estructurales o de estilo, etc. [Berry, 2003], [López, 2002], [Molina, 2002]. Es comprensible que para lograr tal diversidad de propósitos, casi todos los algoritmos que existen en la Minería de Datos pueden ser aplicados a la Minería de Texto. En particular, la clasificación o categorización [Yu, 2004], [Bloehdorn, 2004], [Liu, 2004], el agrupamiento [Zhong, 2003], [Sprague, 2003], [Walls, 1999], y el descubrimiento de itemsets frecuentes y de dependencias (o asociaciones) [Feldman, 1996], son tres de las tareas más aplicadas en este tipo de minería.

Uno de las principales dificultades de estas técnicas, para el caso de la Minería de Texto, es el problema de la dimensionalidad y su reducción [Soucy, 2003]. Entre los métodos aplicados a esta problemática se tienen: los Testores Típicos, los métodos Bayesianos, el K-Nearest Neighbors, el SVM, el LSI, etc.

Como ejemplo de lo anterior se tiene el trabajo de Chakrabarti, Roy, y Soundalgekar, en el que se propone el algoritmo SIMP para disminuir la dimensionalidad de los rasgos de los documentos aplicando el método SVM, clasificando posteriormente los documentos por Árboles de Decisión [Chakrabarti, 2003]. Otro ejemplo es la tesis doctoral de Aurora Pons, en la que se aplica el algoritmo LEX, basado en el método de los Testores Típicos, para discriminar entre grupos de noticias previamente generadas [Pons, 2004].

Uno de los métodos con mayores trabajos en la literatura reciente es el conocido como Indexado Semántico Latente (LSI - *Latent Semantic Indexing*). En este método se aplica la técnica SVD (*Singular Value Decomposition*) a la matriz formada por los vectores de un conjunto de documentos.

La reducción de la dimensión se logra considerando las Variables Latentes, representando estas Espacios Latentes de dimensión reducida [Berry, 2003].

Una aplicación de este última estrategia es el trabajo de Park y Ramamohanarao, en el que se propone un método de recuperación de documentos mediante un mapeo de consultas híbridas sobre Espacios Vectoriales Semánticos Latentes [Park, 2004]. Otro ejemplo de la aplicación del SVD es la propuesta de David Gleich y Leonid Zhukov, en el que se aplica este método en sistemas de sugerencia de términos de búsqueda, como ocurre en el mercado de búsqueda "p*ay-for-performance*" tipo *Overture*. En esta propuesta se considera la proyección de subespacios ortogonales de refinamientos positivos y negativos para la sugerencia y ordenamiento de los términos [Gleich, 2004].

Todo lo anteriormente expuesto sobre la Minería de Texto es aplicable sin grandes restricciones al procesamiento y minería de los documentos en la Web. No obstante, las peculiaridades de la Web pudieran introducir otras posibilidades y necesidades determinadas por el medio.

Ejemplo de lo anterior es el trabajo de Velásquez *et al*., indicado en el epígrafe de la Minería de Uso de Web. En este trabajo se recomiendan las páginas a visitar combinando la secuencia de páginas y sus contenidos evaluados vectorialmente [Velasquez, 2004].

Otro ejemplo es la propuesta de Raymond Kosala *et al*., en la que se analizan las características de los documentos HTML y XML. En ese trabajo se considera la estructura arbórea que explícitamente poseen estos documentos. Para ello, los árboles de los documentos se convierten en árboles binarios pesados, induciendo autómatas arbóreos "*k-testable*" para extraer la información contenida [Kosala, 2003].

También puede darse como ejemplo el trabajo de Gui-Rong Xue *et al*., en donde se propone el algoritmo IRC (*Iterative Reinforcement Categorization*) el cual, a partir de una pre-clasificación de las páginas con los datos disponibles tales como: plain text, título, metadata, etc., converge a una estructura representada por un grafo bipartito, con los *Queries* de un lado y las páginas del otro, enlazadas con las frecuencias de "cliqueo" de cada página dada un *Query*, y las probabilidades de pertenencia a las categorías previamente encontradas [Xue, 2004].

Un tipo de sistema en el que se observa la aplicación de la Extracción (y Recuperación) de Informaciones en la Web, y que va cobrando auge en los últimos años, es lo denominado por Yi Yu Yao como WRSS (*Web-based Research Support Systems*); o sea, como tipos de sistemas de soporte a la Investigación (RSS - *Research Support Systems*), los cuales mejoran los sistemas existentes de búsqueda de artículos, indización, y de análisis de cita, tales como el *Current Content*, DBLP, *Science Citation Index*, y el *CiteSeer* [Yao, 2003], [Xu, 2003], [Yao, J.T., 2003], [Tang, 2003], [Xiang, 2003].

Otras formas de este tipo de Minería de Web se observan en lo conocido como Minería de Enlaces (*Link Mining*); o sea, la intersección de las investigaciones sobre las redes sociales, el análisis de estructuras de hiperenlaces en la Web y la minería de grafos [Chakrabarti, 2004].

En esta misma línea, Ben-Dov, Wu, Feldman y Cairns proponen usar técnicas de análisis de enlaces (*Link-analysis*) sobre los rasgos extraídos para encontrar nuevo conocimiento. Esos enlaces son creados mediante un proceso de Extracción de Información. Para ello, se consideran como estrategias la búsqueda de co-ocurrencia y las técnicas semánticas, verificando que estas últimas son mejores para encontrar informaciones específicas, mientras que la primera permite cubrir mayores informaciones [Ben-Dov, 2004].

También puede mostrarse en esta línea el trabajo de Qing Lu y Lise Getoor, los cuales proponen un ambiente para la modelación de la distribución de enlaces, que soporta modelos discriminativos que describen tanto las distribuciones de enlaces como los atributos de objetos enlazados. En este trabajo se aplica un método de regresión logística estructurada, capturando tanto el contenido de los documentos como sus interrelaciones. Este trabajo fue aplicado en la Web y en colecciones de citas [Lu, 2003].

Las estrategias relativas a las redes sociales han sido aplicadas incluso a las Bibliotecas Digitales, como se observa en el trabajo de Peter Mutschke, el cual considera una red de co-autores para la búsqueda de expertos en dominios científicos [Mutschke, 2003].

3 Otras áreas relacionadas

Además de estas temáticas, pudieran abordarse otras por su relación con el Web Mining, destacándose entre estas las siguientes.

3.1 Sistemas de agentes

Los Sistemas Multiagentes son sistemas en los que varios agentes, usualmente inteligentes, interactúan entre si para lograr determinados objetivos o realizar ciertas tareas. Estos agentes realizan estas tareas a partir de determinadas habilidades, tales como: la reactividad, o la capacidad de percibir y responder a cambios del entorno; la pro-actividad, o la capacidad de poseer iniciativas; y la sociabilidad, o la capacidad de interactuar con otros agentes y personas [Weiss, 2000], [Rojo, 2002], [Olivares, 2002], [Pazienza, 2003].

La relación con la Minería de Web se observa en que estos sistemas actúan en un entorno distribuido, como son las redes locales y la Internet. Un tipo especial de los Sistemas Multiagentes es el conocido como Agentcities. Los Agentcities son un conjunto de sistemas de software (plataformas) conectados a redes de Internet pública. Cada una de esas plataformas posee sistemas de agentes capaces de comunicarse con el mundo exterior usando mecanismos estándares de comunicación[13].

Las plataformas de agentes suelen estructurarse con diferentes tipos de agentes. Entre los tipos de agentes relevantes en la Minería de Web se tienen los siguientes [Kosala, 2000]:

- Agentes de interfaces usuarios – Aquellos que incluyen técnicas de aprendizaje de los intereses de los usuarios. Ejemplos de estos son los Agentes de Recuperación de Información, los Agentes Recomendadores, los Agentes Personales, etc.

[13] URL: http://www.agentcities.org/

- Agentes distribuidos – Aquellos que conforman grupos de agentes que actúan de forma cooperada para descubrir determinados conocimientos. Ejemplos de estos son los Agentes de Filtraje Cooperativo, entre otros.

Los Sistemas Multiagentes están siendo aplicados a muy diversas situaciones y utilizan muchas de las técnicas antes expuestas. Un ejemplo de Sistemas Multiagentes, y en particular de Agentcities, es el proyecto @LIS TechNet, el cual pretende crear un ambiente de enseñanza y experimentación entre varios países de Europa y Latinoamérica[14]. Otro ejemplo es la propuesta de Alfredo Rojo en su tesis de maestría, en la que se proponen diferentes tipos de agentes para recomendar recursos digitales considerando la métrica TF-IDF y el algoritmo KEA [Rojo, 2002], basado en el proyecto de Bibliotecas Digitales de Nueva Zelanda[15].

3.2 Web semántico

El concepto de Web Semántico se aplica a la idea de tener datos en la Web definidos y enlazados en una forma que puedan ser usados por los equipos de cómputo, no sólo para su visualización, sino para ser utilizados por diversas aplicaciones entre las que destacan las de Minería de Web. Un término que designa al conocimiento empleado por el Web Semántico es el de Ontología (*Ontology*) [Olivares, 2002], [Rojo, 2002], [Sowa, 2000].

El término de Ontología en un sentido formal puede ser expresado como la tupla <C, T, P, L>, donde C es un conjunto de conceptos, T un conjunto de términos (en un lenguaje dado), P un conjunto de proposiciones o relaciones entre esos conceptos y L: CxT la relación de correspondencia entre los términos y los conceptos. De esta forma, dos conceptos son "idealmente equivalentes" si se conceptualizan con los mismos términos; o sea, si las descripciones léxicas de sus contextos ontológicos son las mismas [Pazienzia, 2003].

Las Ontologías pueden ser representadas de diversas formas, siendo las jerarquías y los grafos de conceptos los más utilizados. Estas han sido empleadas con muy diversos fines, como son la descripción semántica de documentos, perfiles y actividades, así como su aplicación en problemas de clasificación, y recuperación.

El término de Web Semántico ha recibido tanta atención en los últimos años que para muchos este es una forma equivalente a la Minería de Web de Contenido cuando se considera la dimensión semántica de los documentos.

También se observa una íntima relación entre los términos Web Semántico y Ontologías junto a los de Sistemas Multiagentes. Ejemplo de ello es el trabajo de Haarsley [Haarslev, 2003].

Ejemplos de aplicación del Web Semántico es el uso de agentes para la recuperación de informaciones utilizando Ontologías [Pazienza, 2003], [Seig, 2004]; el empleo de las Ontologías para la clasificación de documentos [Bloehdorn, 2004]; y el Filtraje Cooperativo con Ontologías [Mobasher, 2004 b], [Jin, 2004], o el proyecto @Lis TechNet mencionado anteriormente.

4 Especialistas e instituciones

A continuación se relacionan algunos de los especialistas vinculados con la Minería de Web. Estos han sido ordenados alfabéticamente por su primer apellido. Junto a cada especialista se incluye su institución y, en algunos casos, sus direcciones de correo y páginas personales (*Home page*).

- Aggarwal, Charu C.
  T. J. Watson Resch. Ctr., USA. Home page: http://web.mit.edu/charu/www/home.html
- Agosti, Maristella
  Department of Information Engineering, University of Padova, Italy. E-mail: maristella.agosti@dei.unipd.it. Home page: http://ims.dei.unipd.it/members/agosti/.
- Aoki, Terumasa
  University of Toyio. E-mail: aoki@mpeg.rcast.u-tokyo.ac.jp.

[14] URL: http://alis.cs.bath.ac.uk/alis/
[15] URL: http://www.sadl.uleth.ca/nz/cgi-bin/library

- Baeza-Yates, Ricardo
  Dpto. de Ciencias de la Computación, Universidad de Chile, Chile. E-mail: rbaeza@dcc.uchile.cl, Home page: http://www.dcc.uchile.cl/~rbaeza/
- Baglioni, Miriam
  Knowledge Discovery and Delivery Laboratory, Dipartimento di Informática, Università di Pisa, Pisa Italy. E-mail: baglioni@di.unipi.it.
- Bassi, Alejandro
  University of Tokyo. E-mail: abassi@vp.ccr.u-tokyo.ac.jp.
- Berent, Bettina
  Institute of Information Systems, Faculty of Economics, Humboldt University of Berlin, Berlin, Germany. E-mail: berendt@wiwi.hu-berlin.de Home page: http://www.wiwi.hu-berlin.de/~berendt/
- Berry, Michael
  Department of Computer Science, University of Tennessee, USA. E-mail: berry@cs.utk.edu. Home page: http://www.cs.utk.edu/~berry/.
- Blockeel, Hendrik
  Department of Computer Science, Katholieke Universiteit Leuven, Belgium. E-mail: Hendrik.Blockeel@cs.kuleuven.ac.be. Home Page: http://www.cs.kuleuven.ac.be/~hendrik/
- Bloehdorn, Stephan
  University of Karlsruhe, Institute AIFB, Germany. E-mail: bloehdorn@aifb.uni-karlsruhe.de.
- Borges, José
  School of Engineering, University of Porto, Porto, Portugal. E-mail: jlborges@fe.up.pt.
- Boqin, Feng
  Computer Science Department, Xi'an Jiaotong University, China. E-mail: agentmail@xinhuanet.com.
- Bruynooghe, Maurice
  K.U. Leuven, Department of Computer Science, Leuven, Belgium. E-mail: maurice@cs.kuleuven.ac.be. Home Page: http://www.cs.kuleuven.ac.be/~maurice/
- Bussche, Jan Van den:
  Universiteit Hasselt, Belgium. E-mail: jan.vandenbussche@uhasselt.be. Home page: http://alpha.uhasselt.be/~vdbuss/
- Chakrabarti, Soumen
  Indian Institute of Technology, Bombay, India. E-mail: soumen@cse.iitb.ac.in. Home page: http://www.cse.iitb.ac.in/~soumen/
- Chen, Zheng
  Microsoft Research Asia. E-mail: zhengc@microsoft.com. Home page: http://research.microsoft.com/~zhengc/
- Coenen, Frans
  University of Liverpool, UK. E-mail: frans@csc.liv.ac.uk. Home page: http://www.csc.liv.ac.uk/~frans/.
- Davison, Brian D.
  Department of Computer Science and Engineering, Lehigh University, Bethlehem, USA. E-mail: davison@lehigh.edu. Home page: http://www.cse.lehigh.edu/~brian/
- Diaz, Elizabeth
  University of Louisiana at Lafayette, USA. E-mail: elidiaz@bellsouth.net
- Domingos, Pedro
  Department of Computer Science and Engineering, University of Washington, Seattle, USA. E-mail: pedrod@cs.washington.edu. Home page: http://www.cs.washington.edu/homes/pedrod/
- Etzioni, Oren
  Department of Computer, Science and Engineering, University of Washington, Seattle, WA, USA. E-mail: etzioni@cs.washington.edu. Home page: http://www.cs.washington.edu/homes/etzioni/
- Fanguy, Ronnie

Nicholls State University, USA. E-mail: is-raf@nicholls.edu. Home page: http://www.nicholls.edu/is/FANGUY.HTM

- Feldman, Ronen
  Department of Computer Science, Bar Ilan University Ramat Gan, Israel. E-mail: Feldman@cs.biu.ac.il. Home Page: http://www.cs.biu.ac.il/~feldman/.
- Gelbukh, Alexander
  Centro de Investigación en Computación (CIC), IPN, México. E-mail: gelbukh@cic.ipn.mx. Home page: http://www.gelbukh.com/
- Getoor, Lise
  Department of Computer Science, UMIACS, University of Maryland, USA. E-mail: getoor@cs.umd.edu. Home page: http://www.cs.umd.edu/~getoor/
- Gleich, David
  Harvey Mudd College, Claremont, USA. E-mail: dgleich@cs.hmc.edu. Home page: http://www.stanford.edu/~dgleich/
- Greiner, Russ
  Department of Computing Science, University of Alberta, Canada. E-mail: greiner@cs.ualberta.ca. Home page: http://www.cs.ualberta.ca/~greiner
- Guzmán Arenas, Adolfo
  Centro de Investigación en Computación (CIC), IPN, México. Home page: http://www.cic.ipn.mx/aguzman/
- Haarslev, Volker
  Concordia University, Montreal, Canada. E-mail: haarslev@cs.concordia.ca. Home page: http://www.cs.concordia.ca/~haarslev/
- Han, Jiawei
  Department of Computer Science, Univ. of Illinois at Urbana-Champaign, USA, E-mail: hanj@cs.uiuc.edu. Home Page: http://www-sal.cs.uiuc.edu/~hanj/pubs/research.html
- Herlocker, Jon
  School of Electrical Engineering and Computer Science, Oregon State University. E-mail: herlock@eecs.oregonstate.edu. Home Page: http://web.engr.oregonstate.edu/~herlock/
- Hotho, Andreas
  University of Kassel, Alemania. E-mail: hotho@cs.uni-kassel.de. Home page: http://www.kde.cs.uni-kassel.de/hotho/
- Jin, Xin
  Center for Web Intelligence, School of Computer Science, Telecommunication, and Information Systems, DePaul University, Chicago, Illinois, USA. E-mail: xjin@cs.depaul.edu.
- Kawamae, Noriaki
  NTT Information Sharing Platform Laboratories, Tokio, Japón. E-mail: kawamaie.noriaki@lab.ntt.co.jp.
- Kosala, Raymond
  Department of Computer Science, Katholieke Universiteit Leuven, Belgium. E-mail: Raymond@cs.kuleuven.ac.be.
- Kumar, Vipin
  Department of Computer Science, University of Minnesota, Minneapolis, USA. E-mail: kumar@cs.umn.edu. Home page: http://www-users.cs.umn.edu/~kumar/
- Levene, Mark
  School of Computer Science and Information Systems, Birkbeck College, University of London, London. E-mail: mark@dcs.bbk.ac.uk. Home page: http://www.dcs.bbk.ac.uk/~mark/
- López López, Aurelio
  Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México. E-mail: allopez@acm.org. Home page: http://ccc.inaoep.mx/~allopez/
- Liu, Tao

Nankai University, China. E-mail: liut@office.nankai.edu.cn.

- Ma, Wei-ying
  Microsoft Research Asia. E-mail: wyma@microsoft.com. Home page: http://research.microsoft.com/users/wyma/
- Madey, Gregory
  Department of Computer Science, University of Notre Dame, France. E-mail: gmadeyg@cse.nd.edu. Home page: http://www.nd.edu/~gmadey/
- Mineau, Guy
  Department of Computer Science, Université Laval, Québec, Canada. E-mail: Guy.Mineau@ift.ulaval.ca. Home page: http://www.ift.ulaval.ca/~mineau/
- Mobasher, Bamshad
  School of Computer Science, Telecommunication, and Information Systems, DePaul University, Chicago, USA. E-mail: mobasher@cs.depaul.edu. Home Page: http://maya.cs.depaul.edu/~mobasher/
- Möller, Ralf
  University of Applied Sciences, Wedel, Germany. E-mail: rmoeller@fh-wedel.de. Home page: http://www.sts.tu-harburg.de/~r.f.moeller/
- Montes y Gomez, Manuel
  Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México. E-mail: mmontesg@inaoep.mx. Home page: http://ccc.inaoep.mx/~mmontesg/
- Mutschke, Peter
  Social Science Information Centre, Lenn´estr. 30, D–53113 Bonn, Germany. E-mail : mutschke@bonn.iz-soz.de. Home page: http://www.gesis.org/IZ/Mutschke/
- Park, Laurence
  ARC Centre for Perceptive and Intelligent Machines, The University of Melbourne, Australia. E-mail: lapark@cs.mu.oz.au. Home page: http://www.cs.mu.oz.au/~lapark/main.html
- Pazienza, Maria Teresa
  Artificial Intelligence Research Group, Department of Computer Science, Systems and Production. University of Roma "Tor Vergata". E-mail: pazienza@info.uniroma2.it.
- Price, Bob
  Dept. of Computing Science, University of Alberta, Canada. E-mail: bprice@cs.utoronto.ca. Home page: http://www.cs.toronto.edu/~bprice/
- Raghavan, Vijay
  Center for Advanced Computer Studies, University of Louisiana at Lafayette, USA. E-mail: raghavan@cacs.louisiana.edu. Home page: http://www.cacs.louisiana.edu/faculty/raghavan.html
- Ramamohanarao, Kotagiri
  ARC Centre for Perceptive and Intelligent Machines, The University of Melbourne, Australia. E-mail: rao@cs.mu.oz.au. Home page: http://www.cs.mu.oz.au/~rao/
- Richardson, Matthew
  Department of Computer Science and Engineering, University of Washington, Seattle, USA. E-mail: mattr@cs.washington.edu. Home page: http://www.cs.washington.edu/homes/mattr/
- Robertson, Stephen
  Microsoft Research Cambridge and City University London, UK. E-mail: ser@microsoft.com. Home page: http://research.microsoft.com/users/robertson/.
- Ruggieri, Salvatore
  Dipartimento di Informática, Universit`a di Pisa, Pisa Italy. E-mail: ruggieri@di.unipi.it. Home page: http://www.di.unipi.it/~ruggieri/
- Shaked, Tal
  Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. E-mail: shaked@cs.washington.edu. Home page: http://www.cs.washington.edu/homes/tshaked/
- Soderland, Stephen

Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. E-mail: soderlan@cs.washington.edu.

- Soucy, Pascal
  Department of Computer Science, Université Laval, Québec, Canada. E-mail: Pascal.Soucy@ift.ulaval.ca.
- Sowa, John F
  VivoMind Intelligence, Inc. Home page: http://www.jfsowa.com/pubs/index.htm
- Sprague, Alan
  University of Alabama at Birmingham. Birmingham, USA. E-mail: sprague@cis.uab.edu. Home page: http://www.cis.uab.edu/sprague/
- Tan, Pang-Ning
  Department of Computer Science, University of Minnesota, Minneapolis, USA. E-mail: ptan@cs.umn.edu. Home page: http://www.cse.msu.edu/~ptan/
- Takahashi, Katsumi
  NTT Information Sharing Platform Laboratories, Tokio, Japón. E-mail: takahashi.katsumi@lab.ntt.co.jp.
- Turini, Franco
  Dipartimento di Informática, Universit`a di Pisa, Pisa Italy. E-mail: turini@di.unipi.it. Home page: http://www.di.unipi.it/~turini/
- Velasquez, Juan
  University of Toyio. E-mail: jvelasqu@mpeg.rcast.u-tokyo.ac.jp.
- Villaseñor Pineda, Luis
  Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México. E-mail: villasen@inaoep.mx. Home page: http://ccc.inaoep.mx/~villasen/
- Vindigni, Michele
  Artificial Intelligence Research Group, Department of Computer Science, Systems and Production. University of Roma "Tor Vergata". E-mail: vindigni@info.uniroma2.it.
- Weld, Daniel S.
  Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. E-mail: weld@cs.washington.edu. Home page: http://www.cs.washington.edu/homes/weld/
- Wheeldon, Richard
  School of Computer Science and Information Systems, Birkbeck College, University of London, London. E-mail: richard@dcs.bbk.ac.uk. Home page: http://www.rswheeldon.com/
- Wu, Gongyi
  Nankai University, China. E-mail: wgy@nankai.edu.cn.
- Xue, Gui-Rong
  Computer Science and Engineering, Shanghai Jiao-Tong University, Shanghai, China. Home page: http://apex.sjtu.edu.cn/people/grxue/
- Yao, Yi Yu
  Department of Computer Science, University of Regina, Regina, Saskatchewank, Canada. E-mail: yyaog@cs.uregina.ca. Home Page: http://www2.cs.uregina.ca/~yyao
- Yao, Jing Tao
  Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada. E-mail: fjtyao@cs.uregina.ca. Home page: http://www2.cs.uregina.ca/~jtyao/
- Yasuda, Hiroshi
  University of Toyio. E-mail: yasuda@mpeg.rcast.u-tokyo.ac.jp.
- Yu, Philip S.
  IBM Thomas J. Watson Research Center, USA. Home page: http://www.research.ibm.com/people/p/psyu/
- Zhang, Benyu
  Microsoft Research Asia. E-mail: byzhang@microsoft.com. Home page: http://research.microsoft.com/users/byzhang/

- Zhu, Tingshao
  Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada. Home page: http://www.cs.ualberta.ca/~tszhu/
- Zhukov, Leonid
  Yahoo! Research Labs, Pasadena, USA. E-mail: leonid.zhukov@overture.com. Home page: http://www.gg.caltech.edu/~zhukov/
- Wang, Ziqiang
  Computer Science Department, Xi'an Jiaotong University, China. E-mail: wzqagent@xinhuanet.com.

5 Eventos

A continuación se relacionan los principales eventos que se tendrán en el año 2005 y algunos de los anunciados para el año 2006. Estos han sido ordenados por la fecha de realización.

Abreviatura	Nombre	Lugar	Fecha
Eventos del Año 2005			
EGC 2005	5èmes journées d'Extraction et de Gestion des Connaissances	Paris, France	Jan. 19-21, 2005
ICFCA 2005	Fourth International Conference on Formal Concept Analysis	Lens, France	Feb. 14-18, 2005
OR24	Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security	Orlando, USA	Mar. 28 - Apr. 1, 2005
Salford Systems Data Mining 2005	Second International Data Mining Conference Focusing on the Contributions of Data Mining to Solving Real World Challenges	New York, USA, Barcelona, Spain	Mar. 29-30, 2005
ECIR 2005	27th European Conference on Information Retrieval	Stgo. Compost., Spain	Mar. 21-23, 2005
SETIT 2005	3rd International Conference Technologies, Electronic of Sciences of Information and Telecommunication	Tunisia	Mar. 27-31, 2005
SDM 2005	Fifth SIAM International Conference on Data Mining	Newport Beach, CA, USA	Apr. 21-23, 2005
WWW 2005	The 14th International World-Wide Web Conference	Chiba, Japan	May 10-12, 2005
PAKDD 2005	The 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining	Hanoi, Vietnam	May 18-20, 2005
Wessex Data Mining 2005	Sixth International Conference on Data Mining, Text Mining and their Business Applications	Skiathos, Greece	May 25-27, 2005
WSTST 2005	The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology	Muroran, Japan	May 25-27, 2005
ISMIS 2005	International Symposium on Methodologies for Intelligent Systems	Saratoga Springs, USA	May 25-28, 2005
JCDL 2005	The 5th ACM/IEEE Joint Conference on Digital Libraries	Denver, USA	June 7-11, 2005
SIGMOD – PODS 2005	24th ACM SIGMOD -SIGACT-SIGART Symposium on Principles of Database Systems	Baltimore, Maryland, USA	June 13-15, 2005
EA/AIE 2005	The 18th International Conference on Industrial & Engineering	Bari, Italy	June 22-25,

	Applications of Artificial Intelligence & Expert Systems		2005
CLIMA VI	Sixth International Workshop on Computational Logic in Multi-Agent Systems	City University, London, UK	June 27-29, 2005
ISCC 2005	The Tenth IEEE Symposium on Computers and Communications	Cartagena, Spain	June 27-30, 2005
COLT 2005	The Eighteenth Annual Conference on Learning Theory	Bertinoro, Italy	June 27-30, 2005
SMCia/05	2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications	Finland	June 28-30, 2005
MLDM 2005	Machine Learning and Data Mining in Pattern Recognition	Leipzig, Germany	July 9-11, 2005
AAAI 2005 – IAAI 2005	The Twentieth National Conference on Artificial Intelligence – Seventeenth Innovative Applications of AI Conference	Pittsburgh, PA, USA	July 9-13, 2005
ADMA 2005	The First International Conference on Advanced Data Mining and Applications	Wuhan, China	July 22-24, 2005
UM'2005	The 10th International Conference on User Modeling	Edinburgh, UK	July 24-30, 2005
MDAI 2005	Modeling Decisions for Artificial Intelligence	Tsukuba, Japan	July 25-27, 2005
AAMAS 2005	Fourth International Joint Conference on Autonomous Agents and Multiagent Systems	Utrecht University, The Netherlands	July 25-29, 2005
UAI 2005	21st Conference on Uncertainty in Artificial Intelligence	Edinburgh, Scotland, UK	July 26-29, 2005
IJCAI 2005	International Joint Conference on Artificial Intelligence	Edinburgh, Scotland, UK	July 30 - Aug. 5, 2005
JSM 2005	2005 Joint Statistical Meetings	Minneapolis, MN, USA	Aug. 7-11, 2005
ICML 2005	The 22nd Int. Conference on Machine Learning	Bonn, Germany	Aug. 7-11, 2005
ILP 2005	15th International Conference on Inductive Logic Programming	Bonn, Germany	Aug. 10-13, 2005
SIGIR 2005	The 28th Annual International ACM SIGIR Conference	Salvador, Brazil	Aug. 15-19, 2005
KDD 2005	The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	Chicago, IL, USA	Aug. 21-24, 2005
DaWak 2005	7th Int. Conference on Data Warehousing and Knowledge Discovery	Copenhagen, Denmark	Aug 22-26, 2005
ICIC 2005	International Conference on Intelligent Computing	Hefei, China	Aug. 23-26, 2005
ICCI 2005	2nd International Conference on Computational Intelligence	Istanbul, Turkey	Aug. 26-28, 2005
VLDB 2005	Very Large Data Bases Conference	Trondheim, Norway	Aug. 30 - Sep. 2, 2005
RSFDGrC 2005	The Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing	Regina, Canada	Sep. 1-3, 2005
CLA 2005	The 3rd international conference on Concept Lattices and Their Applications	Olomouc, Czech Republic	Sep. 7-9, 2005
ALaRT 2005	International Workshop on Automatic Learning and Real-Time	Siegen,	Sep. 7-8,

		Germany	2005
IDA 2005	6th International Symposium on Intelligent Data Analysis	Madrid, Spain	Sep. 8-10, 2005
CEEMAS 2005	4th International Central and Eastern European Conference on Multi-Agent Systems	Budapest, Hungry	Sep. 15-17, 2005
ECDL 2005	9th European Conference on Research and Advanced Technology for Digital Libraries	Vienna, Austria	Sep. 18-23, 2005
WI 2005 – IAT 2005	The 2005 IEEE/WIC/ACM International Conference on Web Intelligence – Intelligent Agent Technology	Compiegne University of Technology, France	Sep. 19-22, 2005
PRIMA 2005	Pacific Rim International Workshop on Multi-Agents	Kuala Lumpur, Maylasia	Sep. 26-28, 2005
KCAP 2005	Third International Conference on Knowledge Capture	Banff, Canada	Oct. 2-5, 2005
PKDD 2005	The 9th European Conference on Principles and Practice of Knowledge Discovery in Databases	Porto, Portugal, 2005	Oct. 3-7, 2005
ECML 2005	The 16th European Conference on Machine Learning	Porto, Portugal, 2005	Oct. 3-7, 2005
ALT 2005	The 16th International Conference on Algorithmic Learning Theory	Singapore	Oct 8-11, 2005
DS'05	8th Int. Conf. on Discovery Science	Singapore	Oct 8-11, 2005
IEEE SMC 2005	2005 IEEE International Conference on Systems, Man and Cybernetics	Big Island of Hawaii	Oct. 10-12, 2005
CIAWI 2005	Conferencia Ibero-Americana WWW/Internet	Lisbon, Portugal	Oct. 18-19, 2005
ICWI 2005	IADIS WWW/Internet 2005	Lisbon, Portugal	Oct. 19-22, 2005
M2005	SAS' Annual Data Mining Technology Conference	Las Vegas, NV, USA	Oct. 24-25, 2005
ISC 2005	The 8th IASTED International Conference on Intelligent Systems and Control	Cambridge, USA	Oct. 31 - Nov. 2, 2005
MICAI 2005	The Mexican International Conference on Artificial Intelligence	Monterrey, Mexico	Nov. 14-18, 2005
CIARP 2005	X Iberoamerican Congress on Pattern Recognition	Havana, Cuba	Nov. 15-18, 2005
ICDM 2005	The Fifth IEEE International Conference on Data Mining	Houston, USA	Nov. 26-30, 2005
IAWTIC 2005 – CIMCA 2005	International Conference on Intelligent Agents, Web Technology and Internet Commerce – International Conference on Computational Intelligence for Modelling Control and Automation	Vienna, Austria	Nov. 28-30, 2005
AI2005	The 18th Australian Joint Conference on Artificial Intelligence	Sydney, Australia	Dec. 5-9, 2005
EPIA 2005	12th Portuguese Conference on Artificial Intelligence	Covilhã, Portugal	Dec. 5-8, 2005
SOAS 2005	International Conference on Self-Organization and Adaptation of Multi-agent and Grid	Glasgow, UK	Dec. 11-13, 2005

	Systems		
AIML-05	The International Artificial Intelligence and Machine Learning Conference	Cairo, Egypt	Dec. 19, 2005
ICST 2005	The International Conference on Semantics Technology	Warsaw, Poland	Dec. 23-25, 2005
ICAI 2005	International Conference on Artificial Intelligence	Warsaw, Poland	Dec. 23-25, 2005
Eventos del Año 2006			
ICEIS 2006	First IEEE International Conference on Engineering of Intelligent Systems	Islamabad, Pakistan	Jan. 14-15, 2006
ACST 2006	The IASTED International Conference on Advances in Computer Science and Technology	Pueblo Vallarta, Mexico	Jan. 23-25, 2006
SAINT 2006	The 2006 International Symposium on Applications and the Internet	Phoenix, USA	Jan. 23-27, 2006
AIA 2006	The IASTED: Artificial Intelligence and Applications	Innsbruck, Austria	Feb. 13, 2006
ICFCA 2006	Fourth International Conference on Formal Concept Analysis	Dresden, Germany	Feb. 13-17, 2006
AIKED 2006	5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering, Data Bases	Madrid, Spain	Feb. 18-20, 2006
AC 2006	IADIS Applied Computing 2006	San Sebastian, Spain	Feb. 25-28, 2006
WBC 2006	IADIS Web Based Communities	San Sebastian, Spain	Feb. 25-28, 2006
ITW 2006	IEEE Information Theory Workshop	Punta del Este, Uruguay	March 13-17, 2006
Latin 2006	Latin American Theoretical Informatics	Valdivia, Chile	March 20-24, 2006
MDAI 2006	Modeling Decisions for Artificial Intelligence	Tarragona, Catalonia, Spain	Apr. 3-5, 2006
ICDE 2006	The 22nd International Conference on Data Engineering	Atlanta, GA, USA	Apr. 6-7, 2006
PAKDD 2006	The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining	Singapore	Apr. 9-12, 2006
SAC 2006	The 21st Annual ACM Symposium on Applied Computing	Dijon, France	Apr. 23-27, 2006
AAMAS 2006	Fifth International Joint Conference on Autonomous Agents and Multiagent Systems	Hakodate, Japan	May 8-12, 2006
WWW 2006	15th International World Wide Web Conference	Edinburgh, UK	May 22-26, 2006
ICEIS 2006	8th International Conference on Enterprise Information Systems	Paphos, Cyprus	May 23-27, 2006
JCDL 2006	Joint Conference on Digital Libraries	Chapel Hill, USA	June 11-15, 2006
ICAISC 2006	The Eighth International Conference on Artificial Intelligence and Soft Computing	Zakopane, Poland	June 25-29, 2006
PODS 2006	25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems	Chicago, USA	June 26-28, 2006
ISCC 2006	The Tenth IEEE Symposium on Computers and	Cagliari,	June 26-29,

	Communications	Sardinia, Italy	2006
IPMU 2006	Information Processing and Management of Uncertainty in Knowledge-Based Systems	Paris, France	July 2-7, 2006
ICDM-Leipzig 2006	Industrial Conference on Data Mining	Leipzig, Germany	July 14-15, 2006
AAAI 2006 – IAAI 2006	Twenty-first National Conference on Artificial Intelligence – Eighteenth Innovative Applications of Artificial Intelligence Conference	Boston, USA	July 16-20, 2006
WCCI 2006	IEEE World Congress on Computational Intelligence	Vancouver, Canada	July 16-21, 2006
WTAS 2006 – CI 2006	The IASTED International Conference on Web Technologies, Applications, and Services – The IASTED International Conference on Computational Intelligence	Calgary, Alberta, Canada	July 17-19, 2006
JSM 2006	Joint Statistical Meetings	Seattle, USA	Aug. 6-10, 2006
SIGIR 2006	29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval	Seattle, USA	Aug. 6-11, 2005
ECAI 2006	17th European Conference on Artificial Intelligence	Riva del Garda, Italy	Aug. 28 - Sep. 1, 2005
IEEE IS 2006	3rd IEEE Conference on Intelligent Systems	Varna, Bulgaria	Sep. 4-6, 2006
VLDB 2006	32nd International Conference on Very Large Data Bases	Seoul, Korea	Sep. 12-15, 2006
IEEE SMC 2006	IEEE International Conference on Systems, Man and Cybernetics	Taipei, Taiwan	Oct. 8-11 2006
ICPR 2006	International Conference on Pattern Recognition 2006	Hong Kong	Oct. 20-24, 2006
ICDM 2006	The 2006 IEEE International Conference on Data Mining	Hong Kong	Dec. 18-22, 2006
IAT 2006	2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology	Hong Kong	Dec. 18-22, 2006
WI 2006	2006 IEEE/WIC/ACM International Conference on Web Intelligence	Hong Kong	Dec. 18-22, 2006

6 Conclusiones

Como ha podido apreciarse, en este trabajo se han presentado, de una forma muy sintética, algunos de las principales tipos de aplicaciones de Minería de Web. Estas aplicaciones fueron expuestas a partir de una estructuración de los tipos de minería en este medio, tratando de considerar el criterio de la mayoría de los especialistas de esta temática.

Durante la presentación de los diferentes tipos de Minería de Web se abordaron muchas de las técnicas y métodos propuestos por los especialistas del tema. Al mismo tiempo, se analizaron diversas tecnologías asociadas con este tipo de minería, como son las Bibliotecas Digitales, los WRSS, los Sistemas Multiagentes y la Web Semántica.

Como resultado de este análisis del estado del arte de la Minería de Web, se incluyó un listado de los especialistas consultados, junto a sus instituciones, direcciones de correo y páginas personales.

Por último, se incluyó una relación de los eventos que contemplan la Minería de Web durante los años 2005 y 2006.

Como resultado de este análisis del estado del arte se observan una serie de tipos de aplicaciones con cierta permanencia y "antigüedad" en los diferentes trabajos y eventos, como son: las relativas a la Minería de Estructura y Uso de Web.

Otras, mientras tanto, se han observado fundamentalmente en los años recientes, formando parte de las principales líneas de investigación actuales, como son: el Filtraje Cooperativo y los Sistemas Recomendadores, las Bibliotecas Digitales, los WRSS, el empleo de los Sistemas Multiagentes y la Web Semántica.

Referencias

[Agrawal, 1994] Agrawal, R.; Srikant, R.: Fast algorithms for mining association rules. Proc. of the 20th Int. Conf. on Very Large Databases, Chile, 1994.

[Baeza-Yates, 2001] Baeza-Yates, R.; Ribeiro-Neto, B.: *Modern Information Retrieva*l. Addison Wesley Longman Pub. Inc., 2001.

[Baglioni, 2003] Baglioni, M.; Ferrara, U.; Romei, A.; Ruggieri, S. and Turini, F.: Preprocessing and Mining Web Log Data for Web Personalization. *LNAI 2829*, Springer-Verlag Berlin Heidelberg, pp. 237-249, 2003.

[Ben-Dov, 2004] Ben-Dov, M.; Wu, W.; Feldman, R.; Cairns, P.: Improving Knowledge Discovery by Combining Text-Mining and Link-Analysis Techniques. *SIAM International Conference on Data Mining*, 2004. URL: http://www.uclic.ucl.ac.uk/paul/research/Moty1.pdf.

[Berent, 2001] Bettina Berent, Bamshad Mobasher, Myra Spiliopoulou, and Jim Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. Proceedings of the *Web Mining Workshop*, at the First SIAM International Conference on Data Mining, Chicago, USA, 2001.

[Berry, 2003] Berry, M.: Survey of Text Mining. Clustering, Clasification and Retrieval, Springer-Verlag, 2003.

[Bloehdorn, 2004] Bloehdorn, S.; Hotho, A.: Text Classification by Boosting Weak Learners based on Terms and Concepts. *Proceedings of The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004.

[Borges, 2000] Borges, J.; Levene, M.: A Heuristic to Capture Longer User Web Navigation Patterns. Proceedings of *International Conference on Electronic Commerce and Web Technologies* (EC-Web), Greenwich, UK, 2000. URL: http://www.dcs.bbk.ac.uk/~mark/download/ecweb_jose.pdf

[Borges, 2004] Borges, J.; Levene, M.: An Average Linear Time Algorithm for Web Usage Mining. *International Journal of Information Technology and Decision Making*, pp. 307-320, 2004. URL: http://www.dcs.bbk.ac.uk/~mark/download/borges_linear_time_hpg.pdf

[Bray, 1996] Bray, T.: Measuring the web. Proceedings of *The Fifth International World Wide Web Conference*. Paris, France, 1996.

[Brin, 1998] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine. Proceedings of *Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[Budanitsky, 2001] Budanitsky A., Hirst G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources*, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), 2001.

[Bush, 1945] Bush, V.: As we may think. *The Atlantic Monthly*, July 1945. URL: http://www.javeriana.edu.co/Facultades/C_Sociales/Facultad/sociales_virtual/publicaciones/arena/aswemayth.htm

[Cadez, 2000] Cadez, I.; Heckerman, D.; Meek, C.; Smyth, P.; White, S: Visualization of navigation patterns on a web site using model based clustering. Proceedings of the *6th KDD conference*, 2000.

[Calderón, 2004] Calderón, M.; González-Caro, C.; Pérez-Alcázar, J.; García-Díaz, J.; Delgado, J.: A Comparison of Several Predictive algorithms for Collaborative Filtering on Multi-Valued Ratings. Proceedings of *The 2004 ACM symposium on Applied computing*, 2004.

[Campuzano, 2002] Campuzano, A.: ¿Qué es el permission marketing?
Sito Web: *@RompeCadenas*, En. 24, 2002, URL: http://www.rompecadenas.com.ar/campuzano2.htm

[Chakrabarti, 2000] Chakrabarti, Soumen: Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, ACM SIGKDD, Vol. 1 (2), 2000. URL: http://www.cs.berkeley.edu/~soumen/#papers.

[Chakrabarti, 2003] Chakrabarti, Soumen; Roy, Shourya; Soundalgekar, Mahesh V.: Fast and accurate text classification via multiple linear discriminant projections. *The VLDB Journal*, Springer-Verlag, 2003.

[Chakrabarti, 2004] Chakrabarti, S.; Faloutsos, C.: Graph structures in data mining. *Tutorial de SIGKDD 2004*, 2004. URL: http://www-2.cs.cmu.edu/~christos/TALKS/KDD04-tut/010-christos-foils.PDF.

[Cohn, 2001] Cohn, D. and Hofmann, T.: The missing link - a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge, MA, USA, 2001.

[Cooley, 1997] Cooley, R.; Mobasher, B and Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of *The 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997. URL: http://maya.cs.depaul.edu/~mobasher/pubs.html.

[Craven, 2004] Craven, P.: Google's PageRank Explained and how to make the most of it. Dic. 2004. URL: http://www.webworkshop.net/pagerank.html.

[Davison, 2003] Davison, B.: Overview of: WWW Search Engines. Presentation in Prof. Heflin's course *The Semantic Web*, 2003. URL: www.cse.lehigh.edu/~heflin/courses/semweb/se-overview.pdf.

[Davison, 2004] Brian D. Davison: Learning Web Request Patterns. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, pp. 435-460, Springer-Verlag, 2004. URL: http://www.cse.lehigh.edu/~brian/pubs/.

[Díaz, 2004] Diaz, E.; De, A.; Raghavan, V.V.: On Selective Result Merging in a Metasearch Enviroment. Workshop on Web-based Support Systems, 2004.

[Dongshan, 2002] Dongshan, X.; Junyi, S.: A new markov model for web access prediction. *Computing in Science & Engineering*, 4(6), pp. 34-39, 2002.

[Etzioni, 1996] Etzioni, O.: The World Wide Web: Quagmine or Gold Mine. *Communications of the ACM*, 39 (11), pp. 65-68, 1996. URL: http://www.cs.washington.edu/homes/etzioni/papers/cacm96.pdf.

[Fanguy, 2003] Fanguy, R.; Raghavan, V.: Generating Rule-Based Trees from Decision Trees for Concept-based Information Retrieval. *Proceedings of WI/IAT 2003 Workshop on Applications, Products and Services of Web-based Support Systems*, WSS 2003, Halifax, Canada, 2003. URL: http://www2.cs.uregina.ca/~wss/wss03/wss03.pdf.

[Feldman, 1996] Feldman, R.; Hirsh, H.: Mining associations in text in the presence of background knowledge. *Proc. of the Second International Conference on Knowledge Discovery from Databases*, 1996.

[Friedrich, 2004] Friedrich, S.; Lossau, N.: Search Engine Technology and Digital Libraries: Moving from Theory to Practice. *D-Lib Magazine*, Vol. 10 (9), 2004.

[Gleich, 2004] Gleich, D. and Zhukov, L.: SVD based Term Suggestion and Ranking System. *Proceedings of The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004.

[Han, 2000] Han, J.: From Data Mining To Web Mining: An Overview. Conference tutorial, *2000 International Database Systems Conference* (IDS'2000), Hong Kong, June 2000. URL: ftp://ftp.fas.sfu.ca/pub/cs/han/slides/hkw00.ppt.

[Haarslev, 2003] Haarslev, V. and Möller, R.: Racer: An OWL Reasoning Agent for the Semantic Web. *Proceedings of The 2003 IEEE/WIC International Conference on Web Intelligence*, (WI 2003), Halifax, Canada, 2003.

[Hay, 2003] Hay, B.; Wets, G.; Vanhoof, K.: Web Usage Mining by Means of Multidimensional Sequence Alignment Methods. Proceedings of *WEBKDD 2002*, *LNAI 2703*, Springer-Verlag Berlin Heidelberg, 2003.

[Heylighen, 2001] Heylighen, F.: Collaborative Filtering. Principia Cybernetica Web, Jan. 31, 2001. URL: http://pespmc1.vub.ac.be/COLLFILT.html

[Hösch, 2005] Hösch, J.: E-LEARNING: Collaborative Filtering Strategies to enhance the value of content for the student community. *Training Educatin & Simulation International (TESI 2005)*, Maastricht, The Netherlands, 2005.

[Jin, 2004] Jin, X.; Zhou, Y.; Mobasher, B.: A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content. *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*, Held at AAAI 2004, San Jose, July 2004. URL: http://maya.cs.depaul.edu/~mobasher/papers/swp04.pdf.

[Kawamae, 2004] Kawamae, N. and Takahashi, K.: Collaborative Filtering Based on Latent Classes. Proceedings of the Workshop *"Alternative Techniques for Data Mining and Knowledge Discovery"*, The Fourth IEEE International Conference on Data Mining, ICDM 2004, UK, 2004.

[Kleinberg, 1998] Kleinberg, J.M.: Authoritive sources in a hyperlinked environment. Proceedings of *ACM-SIAM Synmposium on Discrete Algorithms*, 1998. URL: www.cs.cornell.edu/home kleinber/auth.ps.

[Kosala, 2000] Kosala, R.; Blockeel, H.: Web Mining Research: A Survey. *ACM SIGKDD Explorations Newsletter*, ACM Press, Vol. 2 (1), 2000.

[Kosala, 2003] Kosala, R.; Bruynooghe, M.; Bussche, J.V.; Blockeel, H.: Information Extraction from Web Documents Based on Local Unranked Tree Automaton Inference. Proceedings of the *Eighteenth International Joint Conference on Artificial Intelligence*, 2003. URL: http://www.cs.kuleuven.ac.be/cgi-bin-dtai/publ_info.pl?person=7

[Kou, 2002] Kou H., Gardarin G.: Similarity Model and Term Association for Document Categorization. *NLDB 2002*, LNCS 2553, pp. 223-229, Springer-Verlag Berlin Heidelberg, 2002.

[Liu, 2004] Liu, T.; Chen, Z.; Zhang, B.; Ma, W.; Wu, G.: Improving Text Classification using Local Latent Semantic Indexing. Proceedings of *The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004.

[Lu, 2003] Qing Lu and Lise Getoor: Link-based Classification. Proceedings of the *Twentieth International Conference on Machine Learning* (ICML-2003), Washington DC, 2003. URL: http://www.cs.umd.edu/~getoor/publications.html

[López, 2002] López V., Aguilar R.: Minería de Textos y Aprendizaje Automático en el Procesamiento del Lenguaje Natural. *Technical report*, Dpto. Informática y Automática, Univ. Salamanca, España, 2002.

[Mobasher, 1996] Mobasher, B.; Jain, N.; Han, E. and Srivastava, J.: Web mining: Pattern discovery from world wide web transactions. *Technical Report TR96-050*, Department of Computer Science, University of Minnesota, 1996.

[Mobasher, 2002] Mobasher, B.; Dai, H.; Luo, T.; Nakagawa, M.: Using sequential and non-sequential patterns for predictive web usage mining tasks. Proceedings of the *IEEE International Conference on Data Mining*, Japan, 2002.

[Mobasher, 2004 a] Mobasher, B.: Web Usage Mining and Personalization. M. P. Singh, ed., *2004 CRC Press LLC*, 2004. URL: http://maya.cs.depaul.edu/~mobasher/papers/IC-Handbook-04.pdf.

[Mobasher, 2004 b] Mobasher, B.; Jin, X.; Zhou, Y.: Semantically Enhanced Collaborative Filtering on the Web. *Proceedings of the European Web Mining Forum*, LNAI, Springer, 2004. URL: http://maya.cs.depaul.edu/~mobasher/papers/ewmf04.pdf.

[Molina, 2002] Molina, L.: Data Mining: Torturando a los datos hasta que confiesen. Nov. 2002. URL: http://www.uoc.edu/molina1102/esp/molina1102/molina1102.html.

[Montes y Gómez, 2002] Montes y Gómez, M.; Gelbukh, A.; López López, A.: Text mining at Detail Level using Conceptual Graphs. Proc. of the 10th International Conference on Conceptual Structures, ICCS 2002, Bulgaria, LNAI, Vol. 2393, Springer, 2002.

[Mutschke, 2003] Mutschke, P.: Mining Networks and Central Entities in Digital Libraries. A Graph Theoretic Approach Applied to Co-author Networks. *IDA 2003, LNCS 2810*, pp. 155-166, 2003. Springer-Verlag Berlin Heidelberg 2003.

[Nakamura, 2003] Nakamura, A.; Kudo, M.; Tanaka, A.: Collaborative Filtering Using Restoration Operators. *PKDD 2003, LNAI 2838*, pp. 339-349, Springer-Verlag Berlin Heidelberg, 2003.

[O'Connor, 1999] O'Conner, Mark and Herlocker, Jon: Clustering Items for Collaborative Filtering. Proceedings of the ACM SIGIR Workshop on Recommender Systems, 1999.

[Olivares, 2002] Olivares, J.: Un modelo de interacción entre agentes con propósito, ontologías mixtas y eventos inesperados. *Tesis de doctorado del CIC-IPN*, DF, México, 2002.

[Oyanagi, 2003] Oyanagi, S.; Kubota, K.; Nakase, A.: Mining WWW Access Sequence by Matrix Clustering. Proceedings of *WEBKDD 2002, LNAI 2703*, Springer-Verlag Berlin Heidelberg, 2003.

[Park, 1997] Park, J.S.; Chen, M.S.; Yu, P.S.: Using a Hash-Based Method with Transanction Trimming and Database Scan Reduction for Mining Association Rules. IEEE Trans. on Knowledge and Data Engineering, Vol. 9, No. 5, pp. 813-825, October 1997.

[Pazienza, 2003] Pazienza, M.; Vindigni, M.: Agents Based Ontological Mediation in IE Systems. *Springer-Verlag Berlin Heidelberg*, pp. 92-128, 2003.

[Pons, 2004] Pons, A.: Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos. *Doctoral thesis*, University Jaume I, Spain, 2004.

[Park, 2004] Park, L. and Ramamohanarao, K.: Hybrid pre-query term expansion using Latent Semantic Analysis. Proceedings of *The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004.

[Pietracaprina, 2003] Pietracaprina, A. and Zandolin, D.: Mining frequent itemsets using Patricia Tries. Proc. of the Workshop on Frequent Itemset Mining Implementations, FIMI03, Melbourne, FL, USA, 2003.

[Rafiei, 2000] Rafiei, D. and A. Mendelzon, A.: What is this page known for? Computing web page reputations. *Proceedings of the Ninth International World Wide Web Conference*, 2000.

[Raghavan, 1986] Raghavan, V.; Wong, S.: A critical analysis of Vector Space Model for Information Retrieval. *Journal of the American Society on Information Science*, Vol. 37, No. 5, pp. 279-287, 1986.

[Ramos, 2004] Ramos, V. Abraham, A.: Evolving a Stigmergic Self-Organized Data-Mining. Fourth *International Conference on Intelligent Systems, Design and Applications (ISDA-04)*, Budapest, Hungary, 2004. URL: http://alfa.ist.utl.pt/~cvrm/staff/vramos/Vramos-ISDA04.pdf

[Richardson, 2002] Richardson, M. and Domingos, P.: The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. *Advances in Neural Information Processing Systems 14*, 2002. URL: http://www.cs.washington.edu/homes/mattr/doc/nips2002/qd-pagerank.pdf

[Rodríguez, 2003] Rodrigues, Fátima; Duarte, Jorge; Figueiredo, Vera; Vale, Zita A.; Cordeiro, Manuel: A Comparative Analysis of Clustering Algorithms Applied to Load Profiling. Proceedings of the Third International Conference, MLDM 2003, Leipzig, Germany, July 5-7, 2003.

[Rojo, 2002] Rojo García, A.: RA: Un agente recomendador de recursos digitales de la Web. *Tesis de maestría en Ciencias con Especialidad en Ingeniería en Sistemas Computacionales de la Universidad de las Américas*, Puebla, México, 2002. URL: http://www.pue.udlap.mx/~tesis/msp/rojo_g_a/.

[Seig, 2004] Seig, A.; Mobasher, B.; Burke, R.; Lytinen, S.: Using Concept Hierarchies to Enhance User Queries in Web-Based Information Retrieval. *Proceedings of the The IASTED International Conference on Artificial Intelligence and Applications*, Innsbruck, Austria, February 2004.

[Salton, 1971] Salton, G.: *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[Simón, 2004] Simón A., Rosete A., Panucia K., Ortiz A.: Aproximación a un método para la representación en Mapas Conceptuales del conocimiento almacenado en textos, con beneficios para la Minería de Texto. *Proceedings of I Simposio Cubano de Inteligencia Artificial*, Convención Informática 2004, Cuba, 2004.

[Soderland, 2004] Soderland, S.; Etzioni, O.; Shaked, T.; Weld, D.S.: The Use of Web-based Statistics to Validate Information Extraction. *American Association for Artificial Intelligence* (AAAI) 2004. URL: http://www.ai.sri.com/~muslea/atem-04/soderland.pdf.

[Sowa, 2000] Sowa, J.F.: Ontology, Metadata, and Semiotics. *Conceptual Structures: Logical, Linguistic, and Computational Issues*, Lecture Notes in AI #1867, Springer-Verlag, Berlin, 2000.

[Soucy, 2003] Soucy, P.; Mineau, G.: Feature Selection Strategies for Text Categorization. *AI 2003, LNAI 2671*, pp. 505-509, Springer-Verlag Berlin Heidelberg, 2003.

[Spiliopoulou, 2001] Spiliopoulou, M.; Pohle, C.: Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, Vol. 5, pp. 85-114, 2001.

[Sprague, 2003] Sprague, A.: Clustering for Text Mining. *Proceedings of MinDat 2003*, Pachuca, México, 2003.

[Tan, 2002] Tan, P.N.; Kumar, V.: Mining Indirect Associations in Web Data. *WEBKDD 2001, Lecture Notes in Artificial Intelligence* (LNAI 2356), Springer-Verlag Berlin Heidelberg, pp. 145-166, 2002.

[Tang, 2003] Hong Tang, Yu Wu, J.T. Yao, Gouyin Wang, Y. Y. Yao: CUPTRSS: A Web-based Research Support System. URL: http://www2.cs.uregina.ca/~wss/wss03/wss03.pdf.

[Velasquez, 2004] Velasquez, J.; Bassi, A.; Yasuda, H. and Aoki, T.: Mining web data to create online navigation recommendations. Proceedings of *The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004

[Xiang, 2003] Xiaorong Xiang, Yingping Huang, Gregory Madey: A Web-based Collaboratory for Supporting Environmental Science Research. URL: http://www2.cs.uregina.ca/~wss/wss03/wss03.pdf.

[Xu, 2003] Xu, J.; Huang, Y. and Madey, G.: A Research Support System Framework For Web Data Mining. Proceedings of *WI/IAT 2003 Workshop on Applications, Products and Services of Web-based Support Systems*, WSS 2003, Halifax, Canada, 2003. URL: www.nd.edu/~oss/Papers/WIC_webmin_final. pdf.

[Xue, 2004] Xue, G.R. *et al*.: IRC: An Iterative Reinforcement Categorization Algorithm for Interrelated Web Objects. Proceedings of *The Fourth IEEE International Conference on Data Mining*, ICDM 2004, UK, 2004.

[Yao, 2003] Yao, Y.Y.: A Framework for Web-based Research Support Systems. Proceedings of *Computer Software and Application Conference*, COMPOSAC 2003, Dallas, Texas, 2003. URL: http://www2.cs.uregina.ca/~yyao/wss/afwrss.ps.

[Yao, J.T., 2003] Yao, J.T.; Yao, Y.Y.: Web-based Information Retrieval Support Systems: building research tools for scientists in the new information age. *Proceedings of The 2003 IEEE/WIC International Conference on Web Intelligence*, (WI 2003), Halifax, Canada, 2003.

[Yu, 2004] Yu, H.; Han, J. and Chang, K.C.C.: PEBL: Web Page Classification without Negative Examples. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 16, No. 1, January 2004.

[Zhong, 2003] Zhong, S. and Ghosh, J.: A Comparative Study of Generative Models for Document Clustering. *Proc. of SDM Workshop on Clustering High Dimensional Data and Its Applications* May 2003.

[Zhou, 2004] Zhou, Y.; Jin, X.; Mobasher, B.: A Recommendation Model Based on Latent Principal Factors in Web Navigation Data. Proceedings of the *3rd International Workshop on Web Dynamics.* Held at the WWW 2004 Conference, New York, 2004. URL: http://maya.cs.depaul.edu/~mobasher/papers/webdyn04.pdf

[Zhu, 2003] Zhu, T.: Learning Browsing Behavior Model for Web Recommendation. *Doctor of Philosophy Thesis*, University of Alberta, Edmonton, Canada, 2003.

[Zhu, 2004] Zhu, T.: *Tingshao Zhu's Home Page Repository*. URL: http://www.cs.ualberta.ca/~tszhu/webmining.htm.

[Zhu, 2005] Zhu, T.; Greiner, R.; Häubl, G.; Jewell, K.; Price, B.: Off-line Evaluation of Web User Models. Proceedings of The 10th International Conference on User Modelling (UM'2005), will be held in Edinburgh, UK, 2005. URL (Working Paper) http://www.cs.ualberta.ca/~tszhu/publication.html.

[Ziqiang, 2004] Ziqiang, W.; Boqin, F.: Collaborative Filtering Algorithm Based on Mutual Information. *APWeb 2004, LNCS 3007*, pp. 405-415. Springer-Verlag Berlin Heidelberg, 2004.

[Walls, 1999] Walls, F.; Jin, H.; Sista, S.; Schawrtz, R.: Topic Detection in Broadcast News. *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[WEBKDD, 2004] WEBKDD 2004: *Workshop on Web Mining and Web Usage Analysis*. Seattle, USA, August 2004. URL: http://maya.cs.depaul.edu/webkdd04/.

[Wheeldon, 2003] Richard Wheeldon and Mark Levene: The Best Trail Algorithm for Assisted Navigation of Web Sites. Proceedings of *1st Latin American Web Congress*, Santiago, Chile, November, 2003 URL: http://www.dcs.bbk.ac.uk/~mark/download/besttrail.pdf.

[Weiss, 2000] Weiss, G.: *Multiagent Systems: a Modern approach to Distributed Artificial Intelligence*. The MIT Press, 2000.
[Wong, 1985] Wong, S.K.M., Ziarko, W. and Wong, P.C.N.: Generalized Vector Space Model in Information Retrieval. *Proc. of the 8th Int. ACM SIGIR*, Conference on Research and Development in Information Retrieval, New York, ACM 11, 1985.

RTMD_001, Agosto 2007

Aprobado por el Consejo Cientifico CENATAV

Editor: Lic. Arturo Mesa Imbernó
Diseño de Portada: DCG Matilde Galindo Sánchez

RNPS No. 0552
ISSN Solicitado

Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu

C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200

Impreso en Cuba