Informe Técnico – Technical Report
DPTOIA-IT-2006-003
junio, 2006

Minería Web para el Comercio Electrónico

Saddys Segrera Francia
María N. Moreno García

Departamento de Informática y Automática
Universidad de Salamanca

Revisado por:

Dr. Francisco J. García Peñalvo

Dra. Vivian F. López Batista

Aprobado en el Consejo de Departamento de 29 de junio de 2006

Información de los autores:

Saddys Segrera Francia

Estudiante de Doctorado

Departamento de Informática y Automática

Facultad de Ciencias – Universidad de Salamanca

Plaza de la Merced S/N – 37008 – Salamanca

saddys@usal.es

María N. Moreno García

Área de Lenguajes y Sistemas Informáticos

Departamento de Informática y Automática

Facultad de Ciencias – Universidad de Salamanca

Plaza de la Merced S/N – 37008 – Salamanca

mmg@usal.es

Resumen

Internet ha crecido aceleradamente en los últimos años permitiendo la proliferación del comercio electrónico. Se hace necesario que los sitios web sean más intuitivos y accesibles para los usuarios y a su vez, permitan a los proveedores que los negocios se logren exitosamente. La minería web aplicada al comercio electrónico forma parte de este estudio. Se describen las etapas de la minería web, su importancia para mejorar y viabilizar el comercio electrónico, cómo contribuye a la clasificación de información en la web y el uso de multiclasificadores en esta actividad, así como el rol de la personalización en el comercio electrónico para atender usuarios con diferentes perfiles.

Abstract

In the last years Internet has grown quickly allowing the proliferation of the electronic commerce. It is necessary that the web sites are more intuitive and accessible for the users and allow to the providers that the businesses are achieved successfully. The web mining technology applied to the electronic commerce (e-commerce), is part of this study. It describes the stages of web mining, its importance to improve and make possible the electronic commerce, how it contributes to the classification of information in the web and the multi-classifiers use in this activity, as well as the role of the personalization in the electronic commerce to assist users with different profiles.

Tabla de Contenidos

Tabla de Figuras

1 Introducción

Este trabajo presenta un estudio de la forma en que la minería web ha contribuido al desarrollo del comercio electrónico. El estudio recoge los requisitos en cuanto diseño, contexto, funcionalidad y personalización que deben tener los sistemas web para hacer más eficiente el comercio electrónico. De ahí que en [86] se haya expresado que el desarrollo web es una mezcla entre la publicación y el desarrollo de software, entre el *marketing* y la computación, entre las comunicaciones internas y las relaciones externas, y entre el arte y la tecnología. Uno de los aspectos fundamentales es la aplicación de técnicas de minería de datos a información procedente de la web. Es por ello que la minería web es el tema principal de este estudio, en especial su aplicación en la extracción de mayor conocimiento de los clientes, entidades, empresas y relaciones entre ellos.

El trabajo está encaminado a profundizar en el área de minería de uso de la web, aunque también se describen las áreas de minería del contenido y de la estructura de la web, de manera general. Con el auge del uso de Internet se ha favorecido el desarrollo del comercio electrónico a través de sitios web, mediante los cuales tanto los vendedores como los compradores pueden realizar transacciones de negocios de forma ágil y rápida a través de la web. Actualmente existen actividades comerciales, gracias al desarrollo de Internet, que previamente no estaban disponibles o eran muy ineficientes, entre ellas: las interacciones con los clientes, incluyendo la personalización del contenido, las campañas a través del correo electrónico, los servicios al cliente en línea y los estudios en línea, que proporcionan nuevos canales de comunicación. Todas estas actividades son imperativas para que las organizaciones y compañías optimicen sus negocios electrónicos [58].

Con el objetivo de hacer más eficiente el comercio electrónico, los vendedores deben prestar atención al conocimiento sobre el cliente para el establecimiento de soluciones viables que contribuyan a mejorar el grado de satisfacción de los clientes y además, se obtengan incrementos en las ventas. La minería web para el comercio electrónico es la aplicación de minería de datos para adquirir este conocimiento y mejorarlo. El uso de estas técnicas en el comercio electrónico puede ayudar a mejorar las ventas, mostrar características de los productos y/o servicios, ofrecer recomendaciones de compra o de información que puedan resultar de interés a los distintos usuarios, según su perfil.

Muchas compañías en la actualidad emplean las técnicas de minería de datos para anticiparse a las demandas de los clientes, y como resultado estas compañías son capaces de reducir producciones innecesarias y coste de inventario, de modo que pueden variar su estrategia de negocio para obtener estos resultados [69].

En este informe se presenta el estudio de los sistemas de recomendación que han permitido el perfeccionamiento del comercio electrónico. A través de la personalización basada en los perfiles de los usuarios que visitan el sitio web y teniendo en cuenta los intereses de los clientes y criterios de la comunidad, los sistemas de recomendación producen diferentes tipos de salidas y adoptan distintos métodos de recomendación. A pesar de que el tema principal de este trabajo es la minería web aplicada al comercio electrónico, incluyendo la aplicación de multiclasificadores, también se contempla cómo el uso de métodos de computación *soft* ha contribuido al desarrollo de esta disciplina.

Este documento se ha organizado de la siguiente forma: La sección 2, describe la problemática del comercio electrónico y expone la importancia que para el desarrollo del comercio electrónico posee la construcción de modelos web eficientes. La sección 3 recoge la aplicación de la minería web, sus categorías, estado del arte de otras técnicas como la computación *soft* en la minería web, así como las ventajas y desventajas de la aplicación de la

minería web en el comercio electrónico. La sección 4 está dedicada a la aplicación de la minería del uso de la web y los sistemas de recomendación al comercio electrónico. En la sección 5 se explica el uso de multiclasificadores en la minería web. La sección 6 muestra los resultados obtenidos en un caso de estudio. Finalmente, en la sección 7 se presentan las conclusiones.

2 Problemática del comercio electrónico

Internet es un medio de comunicación que permite el intercambio de información entre los usuarios conectados a la red, que según las Estadísticas de Usuarios Mundiales de Internet actualizadas en Noviembre de 2005, conecta a más de 8 millones de servidores encargados de servicios de información y de todas las operaciones de comunicación y de retransmisión. Internet llega a alrededor de 900 millones de usuarios en más de 230 países, ofreciendo una oportunidad única, especial y decisiva a organizaciones de cualquier tamaño.

La rápida difusión y el gran interés en el mundo de la informática han permitido la creación de la tecnología Internet/web, una herramienta fundamental para redes de computadoras y sus usuarios. Se abre así un nuevo mercado que define la "economía digital".

En la práctica, muchas empresas usan Internet como un nuevo canal de ventas, sustituyendo las visitas personales, correo y teléfono por pedidos electrónicos, ya que gestionar un pedido por Internet cuesta 5% menos que hacerlo por vías tradicionales. Nace entonces el comercio electrónico, como una alternativa de reducción de costes y una herramienta fundamental en la actividad empresarial.

El comercio electrónico es una metodología moderna para hacer negocios que responde a las necesidades de las empresas, comerciantes y consumidores de reducir costos, así como mejorar la calidad de los bienes y servicios, además de reducir el tiempo de entrega de los bienes o servicios. Es el uso de la tecnología para mejorar la forma de llevar a cabo las actividades empresariales. El comercio electrónico se puede entender como cualquier forma de transacción comercial en la cual las partes involucradas interactúan de manera electrónica en lugar de hacerlo de la manera tradicional con intercambios físicos o trato físico directo [13].

2.1 Ventajas y desventajas del comercio electrónico

La aparición del comercio electrónico obliga claramente a replantearse muchas de las cuestiones del comercio tradicional, surgiendo nuevos problemas, e incluso agudizando algunos de los ya existentes. En ese catálogo de problemas, se plantean cuestiones que van, desde la validez legal de las transacciones y contratos sin papel, la necesidad de acuerdos internacionales que armonicen las legislaciones sobre comercio, el control de las transacciones internacionales, incluido el cobro de impuestos; la protección de los derechos de propiedad intelectual, la protección de los consumidores en cuanto a publicidad engañosa o no deseada, fraude, contenidos ilegales y uso abusivo de datos personales, hasta otros provocados por la dificultad de encontrar información en Internet, comparar ofertas y evaluar la fiabilidad del vendedor y del comprador en una relación electrónica, la falta de seguridad de las transacciones y medios de pago electrónicos, la falta de estándares consolidados, la proliferación de aplicaciones y protocolos de comercio electrónico incompatibles y la congestión de Internet [13].

No obstante, las ventajas del comercio electrónico y el estudio de los datos producidos de la actividad comercial a través de Internet permiten, mediante la minería web, hacer que el comercio electrónico sea más eficiente y cómodo, tanto para clientes como para proveedores.

Entre las principales ventajas del comercio electrónico, se pueden citar:

- **Para los Clientes**: Permite el acceso a más información, facilita la investigación y comparación de mercados, abarata los costes y precios, aumenta la comodidad en las actividades de negocio.
- **Para las empresas**: Mejora la distribución, permite comunicaciones de mercadeo y obtiene beneficios operacionales.

2.2 Importancia de modelos web eficientes para el comercio electrónico

Un sitio web es a menudo el primer punto de contacto entre un cliente potencial y una compañía. Es por consiguiente esencial que el proceso de exploración/uso del sitio web se realice de tal forma que logre ser tan simple y agradable para el cliente como sea posible. Las páginas web cuidadosamente diseñadas constituyen la principal pieza para un buen "gancho" y éstas pueden reforzarse con la información a la que se accede. El progreso del cliente se supervisa por el registro del servidor web, donde se almacenan detalles de cada página web visitada. Durante un período de tiempo se puede recoger un conjunto útil de estadísticas, que puede usarse para afrontar ciertos problemas, por ejemplo: ¿cuándo el cliente abandonó el carrito de la compra? o ¿por qué los clientes visitan una página pero finalmente no compran? Por tanto, pueden descubrirse algunos problemas que pueden estar relacionados con páginas que están mal diseñadas.

De ahí que los desarrolladores de sitios web deben brindar especial atención a la usabilidad en la web, utilizar un modelo de proceso para sistemas interactivos y usar además, la ingeniería de software. Un sistema es usable, si es eficiente su uso, fácil de aprender, fácil de recordar, tolerante a los errores y subjetivamente agradable [20]. Por lo que es un aspecto que debe ser atendido desde el primer momento.

Por otra parte, en [78] se definen otras características de calidad de una aplicación web:

- **Funcionalidad**: Determina si se tiene el conjunto de funciones apropiadas para las tareas especificadas, verifica si la aplicación hace lo que fue acordado en forma esperada y correcta, si interactúa con otros sistemas especificados, comprueba que lo desarrollado esté acorde con las leyes, normas y estándares, u otras prescripciones y previene accesos no autorizados a los datos y programas.
- **Fiabilidad**: Permite conocer con qué frecuencia el sistema presenta fallos por defectos o errores, y si se presentan fallos, cómo se comporta en cuanto al rendimiento especificado y cuál es su capacidad de recuperación.
- **Eficiencia**: Indica cuál es el tiempo de respuesta y rendimiento en la ejecución de la función y cuántos recursos usa y durante cuánto tiempo.
- **Capacidad de Mantenimiento**: reconoce si es fácil diagnosticar un fallo o identificar partes a modificar, si es fácil de modificar y adaptar, si hay riesgos o efectos inesperados cuando se realizan cambios y si son fáciles de validar las modificaciones.
- **Portabilidad**: determina si es fácil de adaptar a otros entornos las herramientas y tecnologías que en ese momento se poseen, si es fácil de instalar en el ambiente especificado, si se adhiere a los estándares y convenciones de portabilidad y si es fácil usarlo en lugar de otro software para ese ambiente.

Siguiendo estos requisitos se puede crear un sitio web de buena calidad, pero para medir si se cumplieron las expectativas de su funcionamiento y eficiencia, e incluso efectuar mejoras, la aplicación de la minería web sería un elemento a tener en cuenta para conocer qué factores "ocultos" podrían afectar a aquellos sitios web de actividad comercial para brindar servicios de excelencia.

2.3 Personalización en el comercio electrónico

El comercio electrónico permite, como ya se ha mencionado, incrementar la oferta de productos y servicios a los clientes. No obstante, el posible consumidor antes de efectuar una compra debe sentirse cómodo y seguro de realizar esta acción directamente a través de un sitio web. Es por ello, que se hace necesario que el suministrador garantice que la información que se brinda a los

clientes en el sitio web sea correcta, adecuada y fiable, además de gestionar medidas de seguridad, de modo que el sitio sea confiable para el cliente y se respeten las condiciones de privacidad.

Debido al gran número de visitantes que recibe un sitio web, se puede concluir que existen diferentes tipos de clientes con distintas necesidades y prioridades (modelado del usuario) por lo que se requiere personalizar. Según [39] personalizar una aplicación consiste en dotarla de mecanismos que permitan al usuario manejar la información que contiene y los métodos de acceso a ésta, de manera que dicha aplicación se reconvierta a su medida en función de sus expectativas y necesidades. La personalización puede traducirse en la construcción de diferentes interfaces (adaptadas a cada aplicación y cliente particular), en proporcionar caminos de navegación personalizados según las preferencias y expectativas del cliente y en ofrecer distintas políticas de negocio en cuanto a precios, opciones de pago, políticas de fidelidad (promociones personalizadas, club de clientes, formas de atención al cliente, entre otras).

Una alternativa en la que se puede aplicar la minería web en el comercio electrónico sería contribuyendo a la construcción sistemas de recomendación basados en el conocimiento del perfil del usuario. Estos sistemas se utilizan en el área del comercio electrónico para proponer productos a los visitantes del sitio web y ofrecerles información adecuada al tipo de usuario para ayudarle a decidir los productos que van a comprar. Este aspecto será analizado en otro epígrafe con más profundidad. La recomendación se puede realizar sugiriendo productos al consumidor, suministrando los productos con información personalizada o resumiendo la opinión de otros visitantes, entre otras formas [92].

3 Minería web

La minería web o *Web mining* consiste en aplicar las técnicas de minería de datos para descubrir y extraer automáticamente información de los documentos y servicios de la web [26]. En particular, la creación, extracción y mantenimiento de los modelos de usuarios en Sistemas de Recomendación en Internet mejora la experiencia del usuario en relación con la información que le es relevante reduciendo el problema conocido como sobrecarga de la información [40].

Sin embargo, las técnicas de minería de datos no son fácilmente aplicables a datos de la web debido a problemas relacionados tanto con la tecnología subyacente como con la ausencia de estándares en el diseño e implementación de páginas web. La información contenida en archivos *log* y otras fuentes de información debe ser procesada previamente a la obtención de los modelos [67].

En [26] se señala que la minería web está compuesta por tres tareas: descubrimiento de las fuentes, selección y preprocesamiento de la información y descubrimiento de patrones generales desde los sitios web, en esta última es donde se realiza el proceso de minería en sí. En investigaciones posteriores se consideró una cuarta etapa dirigida al análisis para la validación y/o interpretación de los patrones minados [80].

La minería web puede dividirse en tres áreas o categorías principales: minería de contenido, minería de estructura y minería de uso, en función de los datos utilizados para inducir los modelos.

3.1 Etapas de la minería web

A continuación se describen las cuatro tareas que forman parte de la minería web (Figura 1).

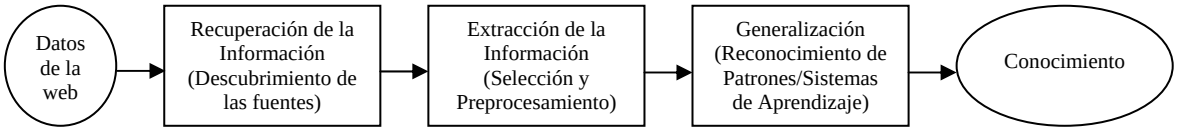

Figura 1. Fases de la minería web, tomado de [80].

3.1.1 Descubrimiento de las fuentes

El descubrimiento de las fuentes o la recuperación de la información (RI) consiste en la recuperación automática de los documentos relevantes, mientras que se asegura al mismo tiempo, tanto como sea posible, que los no relevantes no sean considerados. El proceso de RI principalmente incluye la representación del documento, uso de índices y búsqueda de documentos.

El gran número de páginas en la web, su dinamismo, y la actualización frecuente hace a las técnicas de uso de índices aparentemente imposible. En la actualidad, existen cuatro enfoques para poner índices a los documentos en la web: índice humano o manual, índice automático, índice inteligente o basado en agentes e índice basado en metadatos.

Los futuros sistemas de descubrimiento de las fuentes harán uso de la tecnología de categorización de texto automática para clasificar los documentos de la web en categorías. Esta tecnología podría facilitar la construcción automática de directorios de la web como *Yahoo*, que localiza documentos y los presenta en categorías. Alternativamente, esta tecnología podría ser usada para filtrar los resultados de consultas a los índices de búsqueda.

Los estudios en RI incluyen la modelación, desarrollo de interfaces con el usuario, visualización de los datos, y filtros [3]. Un estudio más detallado de RI en la web está disponible en [55].

3.1.2 Selección/Extracción y Preprocesamiento

Una vez los documentos se han recuperado, el desafío es extraer conocimiento automáticamente y otras informaciones requeridas sin la interacción humana. La extracción de la información (EI) es la tarea de identificar fragmentos específicos de un documento que constituyen su contenido semántico fundamental. Hasta ahora, los principales métodos de EI involucran *wrappers* de escritura (codificación de la escritura) que asigna los documentos a algún modelo de datos.

Los *wrappers* actúan como interfaces de cada fuente de datos, proporcionando una semi-estructura a aquellas fuentes no estructuradas o bien mapean la estructura de datos original en la búsqueda de un patrón común [1]. Los métodos *wrapper*, aunque son eficaces para eliminar atributos irrelevantes y redundantes, son muy lentos, variando en cada ejecución el número de atributos, siguiendo algún criterio de búsqueda y de parada [57].

Otro método para la EI del hipertexto se describe en [30] donde cada página responde a un conjunto de preguntas estándares. Por consiguiente, el problema es la identificación de los fragmentos de texto que responden a esas preguntas específicas.

La EI tiene como objetivo extraer el nuevo conocimiento de los documentos recuperados en la estructura y representación del documento mediante la conversión en mayúsculas, teniendo en cuenta que los expertos de RI consideran que el texto del documento es una bolsa de palabras y no prestan atención a la estructura del documento. La escalabilidad es el mayor desafío más para los expertos de EI; no es factible construir sistemas de EI que sean escalables al tamaño y dinamismo de la web. Por tanto, la mayoría de los sistemas de EI extraen información de sitios específicos y se enfocan en áreas definidas.

Algunos de los agentes inteligentes de la web han sido desarrollados para buscar información relevante usando características de un dominio particular (y posiblemente un perfil de usuario) para organizar e interpretar la información descubierta.

Es necesario un sistema de procesamiento robusto para extraer cualquier tipo de conocimiento, incluso a partir bases de datos medianas. Cuando un usuario solicita una página web se accede a una variedad de archivos como imágenes, sonido, video, *cgi* ejecutables y *html*. Como resultado, el *log* del servidor contiene muchas entradas que son redundantes o irrelevantes para las tareas de minería. Esto significa, que estas entradas deben eliminarse a través del preprocesamiento. Una de las técnicas de preprocesamiento usadas para EI es el índice semántico latente, del inglés *latent semantic index*, que busca transformar los vectores del documento original a un espacio dimensional más bajo mediante el análisis de la estructura correlacional en esa colección del documento de modo que documentos similares que no comparten los mismos términos se colocan en la misma categoría (tema).

3.1.3 Generalización

Una vez que se ha automatizado el descubrimiento y la extracción de la información procedente de los sitios web, el siguiente paso es tratar de generalizar a partir de la experiencia acumulada. Para ello, la minería web ha adaptado técnicas de minería de datos (reglas de asociación, *clustering*, entre otras), de la RI (algunas técnicas para la categorización y la clasificación de textos) y ha desarrollado algunas técnicas propias, como por ejemplo el análisis de caminos, que ha sido usado para extraer secuencias de patrones de navegación desde archivos *log* [44].

Actualmente, la mayoría de los sistemas de aprendizaje desplegados en la web se dedican más a aprender acerca de los intereses de sus usuarios, en lugar de aprender sobre el propio contenido y organización de la web. El problema del etiquetado es un obstáculo importante cuando se aprende sobre la web, ya que los datos son abundantes pero no están etiquetados [26].

Algunas técnicas, como las pruebas de incertidumbre, reducen la cantidad de datos no etiquetados necesarios, pero no eliminan el problema del etiquetado. Un enfoque para resolver este problema se basa en el hecho de que la web es mucho más que una colección enlazada de documentos, es un medio interactivo. Por ejemplo, Ahoy [27] toma como entrada el nombre de una persona y su afiliación y se intenta localizar el *homepage* de esa persona, de esta forma se le pregunta a los usuarios acerca de las páginas recuperadas, para etiquetar sus respuestas como correctas o incorrectas [80].

Las técnicas de *clustering* no requieren las entradas etiquetadas y se han aplicado con éxito a las grandes colecciones de documentos. De hecho, la web ofrece un terreno fértil para investigaciones de *clustering* y clasificación de documentos.

Las reglas de asociación también son una parte íntegra de esta fase. Básicamente, las reglas de asociación son expresiones del tipo $X \Rightarrow Y$, donde X e Y son conjuntos de elementos. $X \Rightarrow Y$ expresa que siempre que una transacción T contenga a X entonces probablemente T también contiene a Y. La probabilidad o confianza de la regla se define como el porcentaje de transacciones que contienen a Y y además a X en relación con el total de transacciones que contienen a X.

3.1.4 Análisis

El análisis es un problema de manipulación de datos que requiere que existan datos suficientes disponibles para que la información potencialmente útil se pueda extraer y analizar. Los humanos juegan un papel importante en el proceso de descubrimiento del conocimiento en la web, considerando que la web es un medio interactivo. Esto es especialmente importante para la validación y/o interpretación de los patrones minados que tienen lugar en esta fase. Una vez que los patrones se han descubierto, los analistas necesitan herramientas apropiadas como el sistema Webviz [85], para entender, visualizar e interpretar estos patrones. Algunos usan el

Procesamiento Analítico en Línea, del inglés *Online Analytical Processing* (OLAP) con el propósito de simplificar el análisis de las estadísticas a partir de los *logs* de los servidores, y otros mecanismos SQL, como el sistema WEBMINER [74] que propone un lenguaje de consultas, similar a SQL, que posee un mecanismo de consultas para preguntar acerca del conocimiento descubierto (en forma de reglas de la asociación y modelos secuenciales).

3.2 Áreas o categorías de la minería web

En [2] se explica que en el caso de la minería web los datos pueden ser obtenidos desde el lado del servidor, del cliente, de los servidores *proxy* o de la base de datos corporativa de la entidad a la cual pertenece el sitio. Desde este punto de vista, los datos encontrados en un sitio web en particular, pueden ser clasificados en tres tipos predominantes, véase la figura 2:

- **Contenido**: Son los datos reales que se entregan a los usuarios. Es decir, los datos que almacenan los sitios web, los cuales consisten generalmente en textos e imágenes u otros medios. Este tipo de dato es el más importante y difícil de procesar, por ser multimedial.
- **Estructura**: Son los datos que describen la organización del contenido en el interior de un sitio. Esto incluye, la organización dentro de una página, la distribución de los enlaces tanto internos al sitio como externos, y la jerarquía de todo el sitio.
- **Uso o Utilización**: Son aquellos datos que describen el uso al cual se ve sometido un sitio, registrado en los *logs* de acceso de los servidores web.

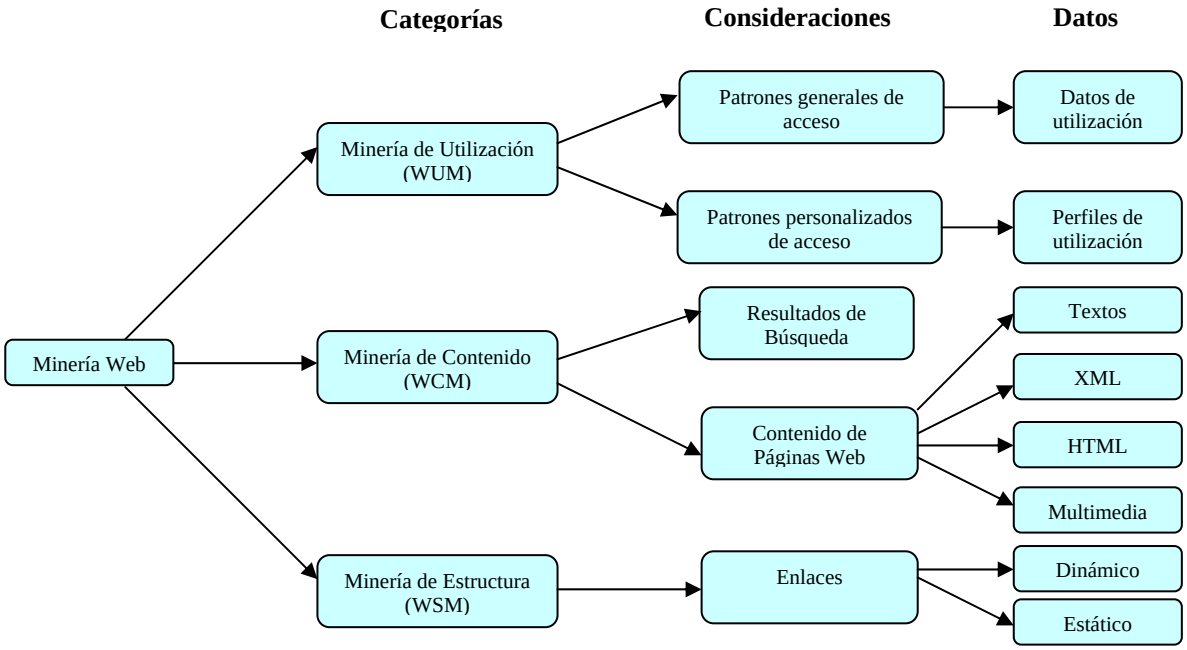

Figura 2. Mapa conceptual de la minería web, tomado de [24].

3.2.1 Minería del contenido de la web

Web Content Mining (minería de contenido web) se centra en el contenido, por lo que se pueden obtener datos acerca de la forma de escribir que sea más atractiva para el usuario, de si la catalogación que se usa sirve para mejorar la relevancia del sitio, si los temas que se tratan interesan o no.

Esta área de la minería web tiene dos vertientes: recuperación de la información y base de datos.

Como se conoce, los sitios de web están compuestos de colecciones de documentos de hipertexto. La recuperación de la información se realiza a través de la exploración semántica de los documentos, mediante dos enfoques: la minería de textos y el análisis semántico de los textos.

Considerando que los sitios de web también son colecciones de documentos semi-estructurados, se pueden descubrir y extraer esquemas para formularios que capturen información semántica relevante de fuentes de datos heterogéneas. Los enfoques están basados en lenguajes de consultas para web (XML, WebSQL, WebML), base de datos múltiples y descubrimiento de jerarquías.

3.2.2 Minería de la estructura de la web

Web Structure Mining (minería de estructura web) se refiere al grado de dificultad que tienen los usuarios para encontrar la información, si la estructura del sitio es simple o muy profunda, si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación es comprensible, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central.

La minería de estructura web revela más información que simplemente la información contenida en los documentos. Por ejemplo, enlaces o eslabones que apuntan a un documento indican su nivel de popularidad, mientras que los enlaces o eslabones que salen de un documento indican la riqueza o quizás la variedad de temas que se abarcan en el documento. Esto fue resaltado por [38] en el algoritmo HITS, (del inglés *Hypertext Induced Topic Selection*), es un algoritmo diseñado para valorar y de paso clasificar la importancia de una página web.

3.2.3 Minería de uso de la web

Web Usage Mining (minería de uso web) tiene como objetivo la extracción de patrones de navegación que se pueden descubrir en los usuarios que visitan un sitio y que pueden ser útiles para mejorar la navegación.

Para llevar a cabo un proyecto de minería de uso de la web, como en todo proyecto de minería de datos, es necesario seguir un proceso perfectamente definido, véase la figura 3. En [21] se explica detalladamente cada una de las fases de la minería de uso de la web.

En la fase inicial, se establecen los objetivos desde el punto de vista del negocio, así como las estrategias de validación de estos objetivos.

En la fase siguiente se reúnen los datos que formarán parte del análisis, pudiendo ser ficheros históricos de *logs* del servidor o servidores del sitio web a analizar, datos de los clientes/usuarios, datos demográficos, datos de facturación y *marketing*, entre otros. Una vez recopilados los datos, se llevarán a cabo tareas de limpieza y selección de los mismos, donde se identificarán las sesiones y transacciones de usuario.

Debido a la existencia de *caché*, en distintos niveles de la conexión del cliente/usuario con el servidor web, algunas páginas que el cliente/usuario recibe no quedan registradas en el fichero *log*. Esto dificulta el proceso de minería de uso de la web, al no tener en este fichero un reflejo fidedigno del recorrido realizado en el servidor. Para resolver este problema hay distintas posibilidades: hacer uso de páginas que no queden almacenadas en la *caché* (páginas activas), incluir en las páginas estáticas convencionales un elemento de este tipo, de manera que siempre quede registrado en los ficheros *log* o, si ninguna de las soluciones anteriores es posible, reconstruir la secuencia real de páginas visitadas a partir de los rastros que queden en el fichero *log* y del mapa del servidor web.

Una sesión de usuario está formada por todas las páginas consultadas por un usuario durante una sola visita al sitio. Una transacción es un conjunto de páginas homogéneas que han sido visitadas en una sesión. El tamaño de una transacción puede variar desde una sola página consultada hasta todas las consultadas en una sesión de usuario. Como resultado de esta fase, se construirán los ficheros o almacenes de datos sobre los se aplican las diferentes herramientas de extracción de información.

Una vez identificadas las transacciones y/o sesiones de usuario, es posible comenzar a buscar patrones de acceso y comportamiento de los usuarios de la web.

El trabajo en la fase de descubrimiento de patrones de acceso consiste en exponer qué técnicas y métodos se puede emplear para realizar minería de uso de la web. Dependiendo del problema que se intenta resolver y de los datos de los que se dispone, serán más adecuadas unas técnicas que otras.

Las técnicas que más se emplean para realizar minería de uso de la web son: agrupamiento y clasificación, detección de reglas de asociación, análisis de caminos y detección de patrones secuenciales.

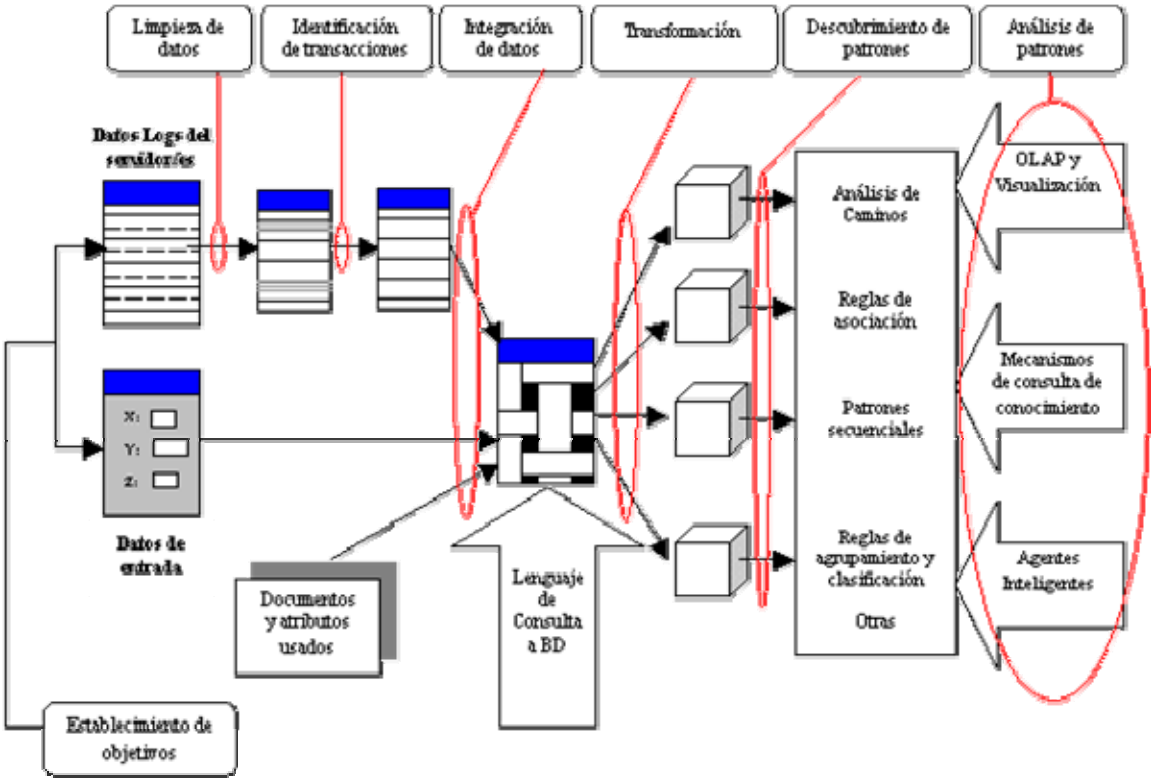

Figura 3. Proceso de minería de uso de la web, tomado de [21].

Las técnicas de agrupamiento o *clustering* distribuyen comportamientos de individuos similares en grupos homogéneos. Esto significa que dos elementos con características similares pertenecerán al mismo grupo y las características de un grupo (definidas por un elemento prototipo o ideal) serán diferentes a las de otro grupo. Dependiendo de la información almacenada en los ficheros *log*, es posible detectar grupos de usuarios como:

- Aquellos que visitan gran cantidad de páginas con un tiempo de estancia muy similar en todas ellas.

- Los que visitan un número pequeño de páginas en sesiones cortas.
- Los que visitan un numero pequeño-mediano de páginas con tiempo variable en cada una de ellas.

Una vez descubiertos los prototipos o perfiles de cada grupo, se pueden usar las características de cada uno de ellos para realizar clasificación.

En minería de uso de la web, las técnicas de clasificación permiten desarrollar un perfil para clientes/usuarios que acceden a ficheros concretos del servidor, en función de sus patrones de acceso. El agrupamiento y clasificación de clientes/usuarios, puede facilitar el desarrollo y la ejecución de estrategias de mercado futuras, tanto *online* como *offline*, tales como envío de correo automático a aquellos clientes/usuarios que se encuentren dentro de un cierto grupo, reasignación dinámica de servidor para un cliente (por ejemplo: menos sobrecargado, para darle un mejor servicio), o presentación de contenidos específicos según el tipo de cliente.

En algunos casos, la información extraída puede ser mezclada con datos de la compañía que permitan reconocer a un usuario entre distintas sesiones de acceso, que recuerden sus actuaciones y sus compras, etc. Esta posibilidad incrementa la calidad de la información que obtenga el proceso de minería de datos.

Las técnicas de descubrimiento de reglas de asociación se aplican a los conjuntos de sesiones y/o transacciones. En este caso, el problema consiste en descubrir todas las asociaciones y correlaciones entre los accesos y usos de la web por parte de los usuarios.

Cada sesión o transacción consiste en un conjunto de URLs accedidas por un cliente en una visita al servidor.

Esta información puede ser obtenida por un paquete estadístico cualquiera, siempre y cuando se sospeche que existe alguna relación entre las páginas señaladas o entre la oferta y la compra realizada. A diferencia del análisis estadístico, que es corroborativo, el análisis de asociaciones descubre las relaciones sin que exista intervención alguna por parte del operador.

El descubrimiento de estas reglas de asociación para organizaciones dedicadas al comercio electrónico puede ayudar al desarrollo de estrategias de mercado efectivas. Pero además, las reglas de asociación descubiertas a partir de los históricos de acceso web, dan una indicación de cómo disponer mejor el espacio web de una organización.

El análisis de caminos supone la generación de algunas formas de grafos orientados que representan relaciones entre páginas web. Este grafo puede ser un esquema físico en el que las páginas web son los nodos del grafo y los hiperenlaces entre las páginas son las flechas dirigidas entre nodos.

Pueden formarse otros grafos a partir de los tipos de páginas web, con arcos que representen la similitud entre páginas, o creando arcos que muestren el número de usuarios que van desde una página a otra. El análisis de caminos podría utilizarse para determinar los caminos más frecuentemente visitados en un sitio web.

El problema de descubrir patrones secuenciales se centra en localizar la presencia de un conjunto de elementos seguida por otro elemento en un conjunto de transacciones o visitas ordenadas en el tiempo. En un histórico de transacciones de un servidor web, la visita de un cliente se guarda por un período de tiempo asociado. El descubrimiento de patrones secuenciales en los históricos de acceso al servidor web permite a las organizaciones predecir los patrones de visita de usuarios y ayudar en el destino de anuncios dirigidos a grupos de usuarios en función de estos patrones. Analizando esta información, el sistema de minería de uso de la web puede determinar relaciones temporales entre elementos de datos.

3.3 Estado del arte y diferentes direcciones de la minería web

La computación *soft* es un consorcio de metodologías que trabajan de forma sinérgica y proporcionan, de una forma u otra, la capacidad de procesamiento de información flexible para tratar situaciones ambiguas de la vida real. Su objetivo es aprovechar la tolerancia a la imprecisión, la incertidumbre, el razonamiento aproximado y la verdad parcial para lograr robustez, soluciones económicas y acercarse a la manera en que el hombre toma las decisiones [110].

El principal objetivo es crear métodos de computación que logren una solución aceptable de bajo coste mediante la búsqueda de una solución aproximada a un problema formulado de forma precisa o imprecisa. En la actualidad, las principales herramientas de computación *soft* incluyen los conjuntos difusos, las redes neuronales artificiales y los algoritmos genéticos [80].

Las aplicaciones de conjuntos difusos en minería web están dirigidas principalmente a las etapas de recuperación de la información y de generalización.

En [106] se describe un marco para formular consultas lingüísticas y jerárquicas. En esta investigación se detalla un idioma de recuperación de la información que permite a los usuarios que especifiquen las relaciones mutuas entre los atributos deseados de documentos buscados usando los cuantificadores lingüísticos. Los ejemplos de cuantificadores lingüísticos incluyen “la mayoría”, “por lo menos”, “la mitad”.

En [36] los autores tratan el problema de la búsqueda automática y la recuperación de documentos, en los casos donde no se garantiza que las consultas de los usuarios con palabras específicas devuelvan documentos que no incluya a estas palabras.

En [61] se han usado métodos difusos para realizar *clustering* de documentos web o fragmentos de éstos. Por otro lado, en [49] una técnica de *clustering* difusa se ha descrito para minar datos de *logs*.

Algunos algoritmos para reglas de asociación que usan técnicas de lógica difusa han sido sugeridos en [42]. Éstos tratan el problema de emplear reglas de asociación difusas comprensibles para los humanos mediante una base de datos que contiene tanto atributos cuantitativos como nominales. Las reglas de asociación de la forma si *X* es *A*, entonces *Y* es *B*, donde *X*, *Y* son atributos y *A*, *B* son conjuntos difusos minados. Un algoritmo de aprendizaje desarrollado en [77] crea reglas difusas “mezcladas” que incluyen atributos numéricos y nominales.

Las redes neuronales, por su parte, han sido aplicadas con más frecuencia en las tareas de recuperación de la información, en la extracción de la información y en *clustering* en minería web.

Mercure [9] es un sistema de recuperación de la información basado en redes multicapa, que permite la recuperación de documentos usando un proceso de activación y la optimización de consultas. Este modelo está formado por una capa de entrada, que representa la información que requiere el usuario, una capa de neuronas de términos, una capa de neuronas de documentos y una capa de salida que representa el resultado de la evaluación de la consulta.

Las redes neuronales también pueden ser usadas para el aprendizaje de perfiles de usuarios como en [25]. Una red neuronal puede calcular el grado de popularidad[1] de una página (*page rank*) a partir de una combinación de parámetros como conexión, autoridad, vigencia o validez, interés y preferencia del usuario, asignándoles pesos a cada uno, de modo que se refina la red de acuerdo a un interés personalizado por usuario [80].

Los algoritmos genéticos en minería web han sido utilizados con menos frecuencia que la lógica difusa y las redes neuronales. Generalmente, su uso en esta área va dirigido a la búsqueda, la optimización y la descripción.

La recuperación de documentos web mediante el aprendizaje genético de factores de importancia de etiquetas HTML ha sido descrita en [51]. En este caso, el método aprende la importancia de las etiquetas a partir de un conjunto de entrenamiento de textos. Los algoritmos genéticos también se han aplicado con el propósito de seleccionar características en minería de texto [68].

Algunos investigadores están construyendo agentes inteligentes o agentes de Internet que ayudan al uso individual de la web. Por ejemplo, muchos agentes fueron construidos para el filtrado de la información que se recolectaba en la web. WARREN [98] es un sistema multiagente para compilar información financiera. WebSifter [50] es un agente meta-buscador que utiliza la taxonomía[2] para mejorar la búsqueda en la web. Este agente le permite al usuario crear categorías y sub-categorías de sus problemas de información de modo individual en una representación de taxonomías. Cuando un usuario está desarrollando la búsqueda por taxonomía, se consultan las taxonomías almacenadas para realizar sugerencias basadas en trabajos anteriores. Durante la construcción de la búsqueda de taxonomías se consulta el diccionario para encontrar las categorías similares semánticamente. La taxonomía resultante se nutre de las páginas web apropiadas ya encontradas por los múltiples motores de búsqueda.

En [111] se propuso un método para construir un agente software que usa técnicas de minería de datos tales como reglas de asociación con el objetivo de construir un modelo que represente los comportamientos de los usuario *online*, y se utiliza este modelo para sugerir actividades o atajos. Estas sugerencias pueden ayudar a los usuarios a mejorar su navegación con los materiales *online* mediante el hallazgo de recursos relevantes de forma más rápida a través de los atajos recomendados y permite además, asistir al usuario en la elección de las actividades más adecuadas de aprendizaje basado en el comportamiento *online* de usuarios con hábitos correctos durante la navegación.

En [89] se describe una metodología para mejorar las prestaciones de *Adaptive Systems for Web-Based Education* – ASWE (Sistemas Adaptativos Educativos en la Web). Se emplearon técnicas de minería de datos para el descubrimiento de relaciones interesantes en los datos de uso. Las reglas descubiertas ayudaron a realizar modificaciones en ASWE para hacer más eficiente la toma de decisiones. En el descubrimiento de las reglas se utilizó la Programación Genética basada en la Gramática (*Grammar-Based Genetic Programming*) aplicando técnicas de optimización multiobjetivo. También se creó una herramienta que permitiese el pre-procesamiento de los datos de los estudiantes para asignar las restricciones en los tipos de

[1] La popularidad de enlaces se refiere al número de ellos que apuntan de una determinada página a otra, es decir, cuántos enlaces de otros sitios están apuntando hacia el suyo, pero influye además la relevancia propia de esos enlaces, o sea, poco relevante será un enlace de un sitio de mascotas para otro de venta de ordenadores. La mayoría de los buscadores analizan la popularidad y en función de ella incrementan el grado de popularidad de una página web. Mientras mayor sea la popularidad de los enlaces, mayor será la popularidad de la página.

[2] Una taxonomía, directorio o catálogo es una división de un conjunto de objetos (documentos, imágenes, productos, géneros, servicios, etc.) dentro de un conjunto de categorías.

relación que se desean descubrir, así como el uso de los algoritmos de minería de datos para la extracción y visualización de las reglas extraídas.

3.4 Ventajas y desventajas de aplicar minería web al comercio electrónico

Un análisis descrito en [56], permitió señalar algunos aspectos que resultaban beneficiosos para el éxito de la aplicación de la minería de datos en comercio electrónico, otros que eran insuficientes y otros engorrosos.

Para que la minería de datos tenga éxito, deben satisfacerse varios elementos. En el comercio electrónico, algunos están satisfechos y algunos pueden satisfacerse con el diseño apropiado [59]:

1. **Gran cantidad de datos**. Teniendo muchos registros, y haciéndolos coincidir con un patrón se asegura la importancia estadística de los patrones encontrados y reduce la probabilidad de *overfitting* o sobreajuste. Los datos *clickstream* que aparecen en la información de las páginas visitadas pueden recolectarse de forma rápida. Por ejemplo, los servidores de *Yahoo!* almacenan alrededor de 1 billón de páginas visitadas al día [105], lo que implica que sólo los archivos *log* requieren 10 gigabytes por hora. Con tantos datos, la aplicación de estrategias válidas de muestreo, como muestras de usuarios (basado en las *cookies*[3]) bien individualmente o por sesiones.
2. **Muchos atributos (los *records* o registros son extensos)**. Las entidades que se minan deben tener muchos atributos. Si los registros están compuestos de pocos atributos, entonces son suficientes para entender los datos técnicas simples, como los gráficos de barras, los gráficos de dispersión y hojas de cálculo. Cuando existen varias docenas o centenares de atributos, se necesitan de técnicas automatizadas para el análisis e identificación los factores importantes. Con el diseño apropiado de un sitio, pueden recogerse datos extensos y tener disponibles un gran número de atributos.
3. **Datos limpios**. Los datos con ruido y corruptos pueden esconder las predicciones y hacerlas más difíciles. La entrada de los datos de forma manual y la integración con los sistemas de envío pueden introducir inconsistencias y anomalías. La recopilación electrónica directa de la fuente proporciona una calidad superior y que los datos sean muy fiables. Con la arquitectura apropiada, se pueden anotar eventos automáticamente, en un sitio web, en un centro de llamada o en un kiosco.
4. **Dominio procesable**. En el comercio electrónico, pueden hacerse muchos descubrimientos procesables en minería de datos cambiando los sitios web, incluso modificando el esquema del sitio, su diseño, ventas cruzadas (*cross-sells*[4]) y su personalización. Las campañas a través de *emails* son relativamente fáciles de ejecutar

[3] Una cookie es un *string* que se pasa en una cabecera HTTP y que el navegador puede guardar en un pequeño fichero de texto (por ejemplo, en Windows/Archivos temporales de Internet para Microsoft Internet Explorer o en Users/usuario/cookies.txt para Netscape Navigator).

[4] Esta técnica se basa en el mercadeo concéntrico, esto es, en múltiples ofertas alrededor de un mismo cliente. A mayor cantidad de transacciones o relaciones que sostenga una cuenta con los usuarios, mayor será la capacidad de la empresa de retenerlos con el paso del tiempo. Esto requiere de la segmentación de los clientes para adaptar la oferta a las necesidades del cliente o grupo de éstos y de la existencia de alguna matriz que identifique qué productos se le han colocado a qué clientes (y cuáles no) para facilitar eventuales ofertas.

y automatizar. El sistema operacional y el sistema de análisis pueden unirse en uno solo de forma más fácil que en otros tipos de dominios.

5. **Retorno de la Inversión medible**. Evaluar los cambios y seguir su efecto en las tiendas es caro y toma un tiempo largo. En la web cambia todo, pueden recogerse *clickstreams* y eventos electrónicamente y pueden traducirse los efectos de cambios y descubrimientos rápidamente en forma de dinero al sitio web.

Los servidores de web pueden generar *logs* que detallan las interacciones con los clientes, típicamente los navegadores web. Los servidores web generan los *logs* en *Common Log Format* (Formato Común de *Logs*) [99] o *Extended Common Log Format* (Formato Extendido Común de *Logs*), que incluye los campos siguientes: *remote host* (el cliente), *remote logname* (la información de identidad de cliente), *username* para la autenticación, fecha y tiempo de demanda, código del estado HTTP, el número de *bytes* transferidos, URL del servidor consultado, y el agente usuario (el nombre y versión del cliente). La mayoría de los servidores web apoyan opciones para anotar campos adicionales, como *cookies*.

Los esfuerzos por aplicar la minería de datos a la web y al comercio electrónico se inician con los *logs* de los servidores web como fuente primaria [52], [17]. Muchos sitios usan servidores web que apoyan la generación de *logs*, pero esto limita que los datos estén disponibles y crea barreras mayores cuando se necesita recoger información adicional. Los *logs* se diseñaron para poner a punto los servidores web, no para la minería de los datos. Seguidamente, se describen algunos de los problemas de los *logs* y cómo pueden ser solucionados mediante aplicaciones que registran los *clickstreams* en los servidores de aplicaciones:

1. **No identifican sesiones o usuarios**. HTTP no reconoce las sesiones, término que es crucial para la minería de datos y que no existe en el nivel del servidor web. La unión de peticiones que forman una sesión es en la actualidad un tema de investigación activo [17], [6], [14]. Las técnicas comunes confían en las *cookies*, el tiempo, las IPs, y agentes de usuario de navegación. Los problemas ocurren debido a *proxy caches*, a la reasignación de IP, a los navegadores que rechazan las *cookies*, etc. En [6] se señala que las herramientas de sesiones están basadas en reglas heurísticas y son, por lo tanto, propensas al error. Un servidor de aplicaciones controla las sesiones, los registros de usuarios, *login* y *logout*, por lo que es no es necesario el uso de reglas heurísticas.

2. **Necesitan ser combinados con los datos transaccionales**. Los sitios de comercio electrónico almacenan las órdenes en una base de datos de tipo transaccional. La acción de combinar los datos de las órdenes y otros datos transaccionales con los *logs* del servidor web es un proceso complejo de extracción-transformación-carga (*extract-transform-load*) y requiere identificar los *clicks* relevantes para cada transacción. Por ejemplo, uno de los informes más comunes en los sitios dedicados al comercio electrónico es la ganancia por las ventas atribuibles al sitio. Mientras que los *logs* del servidor web pueden contener el URL de la referencia del campo "referer" del HTTP, procesar esta información requiere de la información del *clickstream* y la información de la orden. Sin embargo, un servidor de aplicaciones registra los datos de las órdenes y si éstos también registran eventos *clickstreams*, es posible generar un registro simple y comprensible con las identificaciones constantes (IDs) entre las tablas.

3. **Carecen de eventos críticos**. Los eventos como: "agregar al carrito", "eliminar un artículo" o "modificar la cantidad", no están disponibles en los *logs*. Una de las métricas más importantes para el comercio electrónico es el valor de los carritos abandonados y todavía estos eventos no son calculables a partir de los *logs*. El servidor de aplicaciones debe conocer de este tipo de eventos y puede registrarlos, además también puede registrar eventos específicos e interesantes como los de conocer cuando

el usuario ha seleccionado el botón "Actualizar", cuando la descarga de una página ha finalizado, entre otros.

4. **No guardan la información de formularios web**. Cuando un usuario rellena un formulario, como el formulario de búsqueda, es importante saber qué información se introdujo para mejorar el sitio. Conociendo las palabras claves usadas se puede ayudar a las compañías a agregar sinónimos y pueden mejorar su mezcla de producto. Esto es algo que se puede hacer en un servidor de aplicaciones que analice el formulario. Una idea interesante es registrar las ocasiones en que fallan los formularios debido a errores del usuario. Este nivel de registro de errores puede ayudar a mejorar los formularios.
5. **Contienen URLs, no la información semántica de qué URLs contiene**. Los URLs necesitan ser asignados a la información semántica que describe lo que ellos contienen. ¿Qué producto se presenta cuándo un URL dado se muestra? ¿Qué páginas son parte del proceso de chequeo o registro? La misma página puede tener versiones múltiples en idiomas diferentes. En un servidor de aplicaciones de un sitio dinámico, la información semántica significativa está disponible sobre el contenido de la página que es exhibida.
6. **Carecen de información para sitios modernos que generan contenido dinámico**. Se construyen sitios dinámicos en un conjunto de plantillas de URL que se reutilizan para presentar información diferente, haciendo más difícil extraer la información de los *logs*. ¿Qué producto se presentó al usuario si todos los productos se presentan con la misma plantilla? ¿Se presentó algún elemento dinámico, por ejemplo una promoción? El URL en sí mismo llega a ser menos importante al poder registrar la información en el servidor de aplicaciones.
7. **Son archivos *flat* en los sistemas de archivos múltiples, posiblemente en diferentes zonas horarias**. Los sitios grandes tendrán servidores web múltiples, cada uno registra los datos en archivos separados, normalmente en sistemas de archivos diferentes. La situación es más compleja si los servidores web están geográficamente distribuidos en diferentes zonas horarias. Los *logs* están normalmente en ASCII, que es una forma ineficaz de guardar grandes cantidades de datos estructurados. Los registros del servidor de aplicaciones se pueden generar directamente en la base de datos, de modo que almacena el nivel de integridad de la transacción; significa que independientemente de la zona horaria desde donde se accedió a servidor, se logra almacenar todos los registros en un mismo archivo y en una base de datos estructurada. Los tiempos se pueden almacenar en GMT, posiblemente con otro campo para la compensación de la hora local del navegador del usuario. Se debe realizar la sincronización de los servidores de aplicaciones.
8. **Contienen información redundante**. La mayoría de las entradas en el *log* no son de interés para aplicar técnicas de minería. Por ejemplo, contienen el número de veces que cada imagen existente en la página fue solicitada. Redundancias como éstas se eliminan trivialmente cuando se registran los controles en un servidor de aplicaciones. Normalmente se recorta el 90% de los archivos *log*.
9. **Carecen de información importante que se pueden recopilar usando otros medios**. El encabezamiento HTTP [37] que es la fuente de información de los *logs*, no contiene información importante como el tiempo local del usuario. Con un servidor de aplicaciones cualquier información que se recoge también puede ser registrada dentro de la misma base de datos con las claves (*keys*) apropiadas.

Pese a que la minería de uso de la web se ha enfocado principalmente hacia el descubrimiento de patrones de acceso a partir de los ficheros *log*, existen muchos enfoques diferentes de la solución del problema. Éstas pueden clasificarse de acuerdo con los pasos y técnicas seguidas para la obtención de patrones. Una vez que el *log* del servidor ha sido

preprocesado y se dispone de información de sesiones, una posible alternativa es aplicar una u otra técnica de computación *soft*. Muchas soluciones se han centrado en la aplicación de técnicas inteligentes para proporcionar recomendaciones personales al usuario [67].

4 Aplicaciones de la Minería de Uso de la Web

Los principales procedimientos para minar los patrones de navegación de los usuarios a partir de los datos *log*, según [44] son:

- Transformación de los datos a notación tabular y aplicación de técnicas estándar de minería de datos [16]
- Desarrollo de técnicas *ad-hoc* para el trabajo directo con los datos *log* [97]

Las aplicaciones de la minería de uso de la web van dirigidas a dos campos: el aprendizaje de patrones de navegación y el aprendizaje de perfiles de usuario para proporcionar recomendaciones personales al cliente.

Los patrones de navegación definen el modo en que el usuario accede a la información en la web. Son patrones particularmente relevantes en el ámbito del hipertexto [11], porque permiten modelar la navegación mediante el comportamiento de los visitantes al sitio web.

El perfil de usuario es utilizado en muchas situaciones, entre ellas la personalización de aplicaciones y servicios, en la automatización de tareas, etc. La creación de perfiles de usuario permite conocer mejor a los clientes (sus preferencias, sus patrones de comportamiento, sus características, en resumen, cualquier información que se pueda conseguir de sus transacciones *online* u *offline*) y trazar su comportamiento (páginas visitadas, búsquedas realizadas, productos o servicios adquiridos) [82].

El perfil de usuario contiene información modelada sobre el usuario, representada explícita o implícitamente, y cuya explotación permite al sistema incrementar la calidad de sus adaptaciones.

En la obtención de un perfil más actual y preciso, es necesario acompañar las acciones del usuario de la forma más cercana posible. Por eso se recoge, procesa y guarda información de las acciones del usuario, que sirve para, entre otras cosas, determinar que perfiles de otros componentes del sistema interactúan con el perfil actual, así como para proceder a las depuraciones y actualizaciones que se tengan que realizar [19].

Los perfiles de los consumidores no sólo aseguran incrementar los beneficios de los vendedores a partir del conocimiento de las preferencias de los clientes sino que también aumentan la satisfacción del consumidor ya que facilita de forma más sencilla la compra y logra que los clientes repitan la visita para nuevas adquisiciones [22].

Las aplicaciones de la minería de uso en la actualidad tienen un campo amplio en la construcción de sistemas de recomendación.

4.1 Aprendizaje de Patrones de Navegación

El uso de técnicas de minería de datos para analizar los datos *log* fue propuesto por [16] y [109].

En [16] se describió una técnica para minar los patrones de acceso de los usuarios durante la navegación. Los datos *log* se convierten a un formato o a una estructura que sea más flexible para la aplicación de las técnicas de minería de datos y se implementan dos algoritmos para tratar reglas en este contexto. Por su parte, en [109] se propuso un método donde para cada

sesión del usuario, se infiere, a partir de los datos *log*, un vector para almacenar el número de visitas a cada página. Se aplica un algoritmo para encontrar cluster de vectores similares.

El uso de datos *log* para crear sitios web adaptativos fue recogido en [84], permitiendo la creación de índices de las páginas de forma automática. Se dice que un sitio web es adaptativo cuando automáticamente mejora su organización y presentación mediante el aprendizaje de los patrones de acceso de los visitantes.

En [8] los autores han propuesto un modelo de hipertexto para recoger las preferencias de navegación de los usuarios a través de la web. Este trabajo se basa en gramáticas probabilísticas. El conjunto de sesiones de navegación de usuario se modela como una Gramática Probabilística de Hipertexto (*Hypertext Probabilistic Grammar*[5]) y las cadenas de texto (*strings*) que se generan con la más alta probabilidad se corresponden a los caminos de navegación preferidos por el usuario. El modelo propuesto a los datos *log* en esta investigación toma forma de grafo, donde los pesos de los arcos son las probabilidades que refleja la interacción del usuario a través de las transacciones entre páginas.

El uso de los datos *log* para predecir la siguiente URL que será visitada por el usuario para que el servidor pueda generar el contenido dinámico de antemano, y reducir el intervalo de tiempo que media entre la presentación de la solicitud y el inicio de la respuesta a ésta (latencia), fue recogido en [93]. Se genera además, un árbol que contiene los caminos del usuario a partir de los datos *log* y un algoritmo se propone para predecir la próxima solicitud hecha al árbol y la sesión del usuario actual.

En [67] se propuso el estudio de las sesiones (secuencias de clicks) de los usuarios para encontrar subsesiones o subsecuencias de clicks que están semánticamente relacionadas y que reflejan un comportamiento específico de un determinado usuario, lo que permite analizar los datos de las sesiones con diferentes niveles de granularidad. La propuesta que se realiza en ese trabajo es calcular caminos frecuentes que se pueden utilizar para obtener subsesiones dentro de una sesión con el fin de obtener perfiles de los usuarios, para que cuando éstos estén navegando por el sitio web se puedan precargar páginas que seguramente visitarán y también se pueden utilizar para ofrecer al usuario nuevos elementos que le agradarán en función de su perfil.

Para resolver este problema se ha planteado un algoritmo basado en estructuras auxiliares tipo árbol. La estructura en la que está basado el algoritmo se ha denominado FBP-tree (*Frequent Behavior Paths - Tree*). El FBP-tree representa caminos dentro del sitio web. Después de construir el FBP-tree se pueden obtener reglas de comportamiento habituales que servirán para analizar subsesiones dentro de la sesión de un usuario. El descubrimiento de estas subsecciones permite analizar con diferente nivel de granularidad el comportamiento de los usuarios sobre la base de las páginas visitadas y a las subsesiones. Las subsesiones se pueden ver también como un método probabilístico para obtener el estado de una sesión de navegación. Adicionalmente, también se puede predecir el camino que un usuario recorrerá para llegar a una cierta página.

4.2 Sistemas de Recomendación y Personalización

Los sistemas de recomendación están relacionados directamente con la personalización de los sitios web y con el desarrollo del comercio electrónico.

[5] Es una gramática regular (una tupla V, Σ, S, P) con una relación uno a uno entre el conjunto de símbolos no terminales (V) y el conjunto de símbolos terminales (Σ). Cada símbolo no terminal corresponde a una página web y cada regla de producción de la gramática (P) a un enlace entre páginas. Existen dos estados artificiales S y F que representan los estados de inicio y fin de las sesiones de navegación [44]

La personalización engloba una serie de procesos fundamentales e interdependientes [35]:

- **Adquisición de datos del usuario**. Se trata de ampliar y utilizar la información contenida en los ficheros de *log* del sitio web para enriquecer los datos que se tienen acerca de la interacción que realiza el usuario con la plataforma.
- **Construcción de modelos**. Se trata de ampliar la información y las técnicas para la construcción de los modelos que sustentan las tareas de adaptación previstas para el sistema.
- **Identificación de las tareas de adaptación**. Relacionado con los modelos construidos y la definición de tareas adaptativas, se trata de identificar para cada tarea de aprendizaje cooperativo que se plantee qué tipo de ayuda puede ser de utilidad para su realización.

Un modo de construir un sistema de recomendación usando un clasificador sería a través del uso de la información acerca de un producto y un consumidor como datos de entrada y hacer que la categoría de salida represente en qué grado se puede recomendar el producto al cliente. Los clasificadores pueden implementarse a través de distintas técnicas de aprendizaje y atendiendo a distintas estrategias. A pesar de las ventajas que ofrece el uso de multiclasificadores no abundan trabajos que describan su implementación en los sistemas de recomendación.

En la mayoría de los casos, los problemas de recomendación en el comercio electrónico van dirigidos a resolver tres situaciones (1) si los clientes para quienes se quieren realizar las recomendaciones (clientes activos) son todos los clientes o un conjunto de ellos previamente seleccionados, (2) si el objetivo de las recomendaciones es predecir cuanto a un cliente particular le gustará un producto particular (problema de predicción), o identificar una lista de productos que serán de interés a un cliente dado (problema de *top-n* recomendaciones), y (3) si la recomendación se cumple en un momento específico o es constante. Por ejemplo, los sistemas de recomendación de filtrado colaborativo presentan las predicciones o *top-n* recomendaciones a los clientes siempre que ellos visiten el sitio. Por otro lado, la mayoría de los sistemas de gestión de campañas de *marketing* hacen *top-n* recomendaciones para los clientes particulares en un momento especificado [53].

En [92] se señala una taxonomía de sistemas de recomendación de acuerdo a las funcionalidades de entradas del cliente-objetivo, las entradas de la comunidad, a los métodos de recomendación, el grado de personalización, la forma en que se ofrecen las salidas (Figura 4).

Las entradas de los clientes-objetivo participan en el proceso de recomendación para proporcionar recomendaciones personalizadas. Una aplicación que no usa ninguna entrada sobre el cliente-objetivo puede producir solamente recomendaciones no-personales. La existencia de diferentes tipos de entradas permite la aplicación de un sistema de recomendación para personalizar recomendaciones basadas en la actividad actual del cliente, las preferencias a largo plazo del cliente o ambos. También existen distintas maneras de categorizar las entradas del cliente objetivo.

La interpretación del comportamiento de los clientes en estas entradas, incluye dos tipos de acciones, aquellas en las que el cliente actuaría exactamente de igual forma aunque no lo alertara el sistema de recomendación y acciones en las que el único objetivo es la mejora de las propias recomendaciones. Las entradas de la navegación implícita generalmente se infieren a partir del comportamiento del cliente sin que él conozca el uso que éstas puedan tener en el proceso de recomendación. Esta entrada puede incluir uno o varios productos que el cliente actualmente está visualizando o aquellos productos que en ese momento están en el carro de la compra.

A diferencia de éstas, las entradas de la navegación explícita son intencionalmente provocadas por el cliente con el propósito de informar al sistema de recomendación de sus preferencias.

En algunos casos, las entradas procedentes del cliente no pueden estar limitadas a una sola categoría o producto de interés. Una opción para ampliar las entradas es el uso de palabras claves y atributos de productos, ya sea a partir de una búsqueda explícita o implícita o como derivación de los productos que se visualizan.

El cliente-objetivo puede proporcionar las entradas más útiles y explícitas en el formulario de valoraciones de los productos que se han consumido. En una situación ideal, se presentan los clientes con una muestra representativa de productos a partir de la base de datos del suministrador electrónico y se pide indicar al cliente su preferencia para cada uno de los productos representativos.

En algunos sitios, en lugar de ir pidiendo a los consumidores que proporcionen las valoraciones explícitas, utilizan el historial de la compra del cliente-objetivo como un formulario implícito de valoraciones. Éstos proporcionan listas de productos para que el cliente exprese una preferencia de manera muy concreta.

Las entradas de la comunidad poseen un intervalo ancho de datos que consideran cómo los diversos individuos en la comunidad, o la comunidad en conjunto, perciben los productos. Las entradas que reflejan las opiniones globales de la comunidad incluyen las asignaciones de atributos de productos. De igual forma, la popularidad externa del artículo puede reflejar la popularidad en las comunidades. Finalmente, así como se usa el historial de compra de un cliente individual como un conjunto de valoraciones implícitas sobre los productos, se puede utilizar el historial de compra de la comunidad para hacer lo mismo.

Los comentarios son útiles, el cliente debe leer cada párrafo y debe interpretar en qué grado le es positivo o negativo. La mayoría de los sitios que ofrecen la oportunidad a la comunidad de escribir comentarios también animan a que los miembros realicen valoraciones a través de algún formulario.

Las salidas de las recomendaciones de productos varían en el tipo y cantidad de información proporcionadas al cliente. El tipo más común de salida puede ser una sugerencia. Esto a menudo toma la forma de “intente esto” o simplemente poniendo “esto” en la página web visitada por el usuario.

Varios algoritmos de los sistemas de recomendación presentan a los clientes una predicción de la valoración que ellos darían a un producto. Estas estimaciones pueden presentarse como las estimaciones personalizadas para los clientes individuales o como estimaciones no-personalizadas para los miembros de la comunidad.

Cuando las comunidades son pequeñas o los miembros de la comunidad son bien conocidos, puede ser útil desplegar las valoraciones individuales de miembros de la comunidad para permitirle al cliente-objetivo predecir su propia conclusión acerca del peso de una recomendación. Esta técnica es particularmente valiosa cuando el cliente puede seleccionar a los miembros conocidos de la comunidad o cuando las valoraciones se acompañan por revisiones. Las revisiones son un ejemplo de recomendaciones que contienen evaluaciones.

La recuperación de la información les proporciona a los clientes una interfaz de búsqueda a través de consultas a una base de datos de productos. Existen sistemas que valoran la personalidad por encima de la personalización y pueden crear conjuntos de recomendaciones que se han seleccionado de forma manual por editores, artistas, críticos y otros expertos. Estos emisores humanos de recomendaciones identifican productos basados en sus propios gustos, intereses y objetivos, y crean una lista disponible de productos recomendados a los miembros de la comunidad.

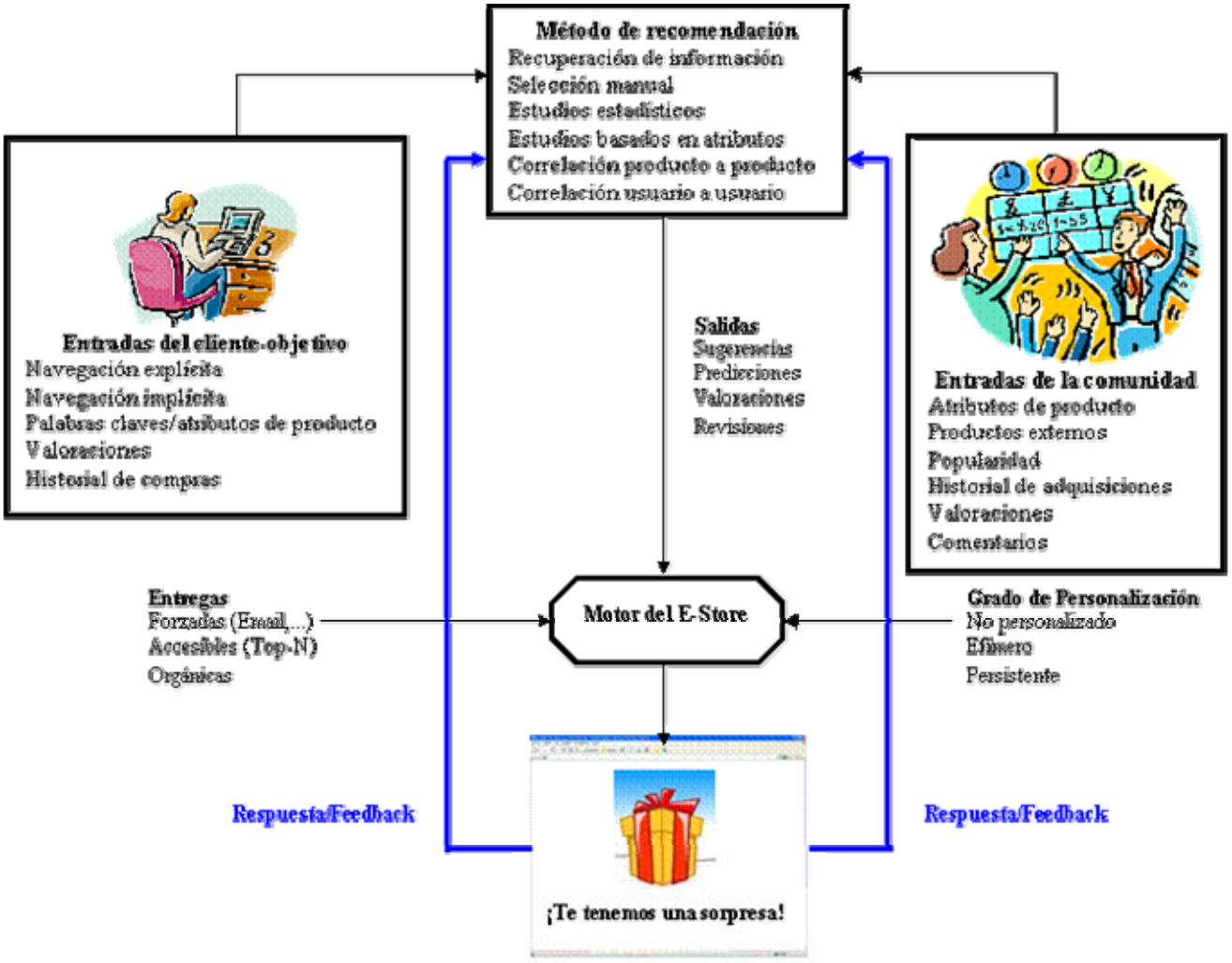

Figura 4. Taxonomía de los sistemas de recomendación, basado en [92] y [34].

En casos donde la personalización no es practicable o innecesaria, los sistemas de recomendación pueden proporcionar muy eficazmente estudios estadísticos de opinión de la comunidad. Las recomendaciones basadas en las propiedades sintácticas de los productos y los intereses del cliente, también pueden utilizar tecnologías de recomendación basadas en atributos.

Otras aplicaciones usan la correlación producto a producto para identificar productos encontrados frecuentemente en "asociación" con productos sobre los que un cliente ha expresado su interés. La asociación puede estar basada en los datos de compra, en la preferencia de clientes comunes o en otras medidas.

Los sistemas de recomendación que usan la correlación usuario a usuario recomiendan los productos a clientes basándose en la correlación entre ese cliente y otros clientes que han comprado productos del sitio a través del comercio electrónico.

Por otra parte, los sistemas de recomendación pueden producir recomendaciones en distintos grados de personalización. El grado de personalización abarca varios factores que incluyen la exactitud y la utilidad de las recomendaciones.

Cuando los sistemas de recomendación proporcionan recomendaciones idénticas a cada cliente, la aplicación es clasificada como no-personalizada. Las recomendaciones específicas pueden estar basadas en la selección manual, estudios estadísticos u otras técnicas.

Como ya se ha mencionado, la personalización es fundamental en el desarrollo de un sistema de recomendación. Estos sistemas pueden ser dotados de personalización, en mayor o menor grado, de ahí que los sistemas de recomendación pueden ser de tipo efímeros o persistentes. El término efímero se puede entender como limitado y persistente como de amplia aplicación. La personalización persistente se basa en un perfil de usuario que dura con el tiempo y se guarda en una estructura mientras que la personalización efímera no se basa en ningún perfil de usuario persistente.

Las diferencias principales son los rasgos temporales del proceso a construir y manejar con el perfil de usuario. En la personalización persistente, el perfil de usuario se desarrolla incrementalmente con el tiempo y al final de cada sesión se guarda para ser usado posteriormente en las sesiones siguientes. Normalmente, pero no necesariamente, la información utilizada para construir el perfil proviene de distintas fuentes, involucra diferentes aspectos del usuario, y está a menudo extendida por medio de procesos de razonamiento y de aprendizaje. En la personalización efímera, la información usada para construir el perfil de usuario sólo se recoge durante la sesión actual, y se utiliza inmediatamente para ejecutar algún proceso adaptable para personalizar la interacción actual. Al final de cada sesión, el perfil del usuario se pierde, y no se guarda ninguna información sobre el usuario de una manera persistente para el uso posterior [72].

Los sistemas de recomendación que usan al cliente actual para personalizar la recomendación de los intereses actuales del cliente ofrecen una personalización efímera. Éste es un paso superior a los sistemas no-personalizados porque proporciona recomendaciones que son sensibles a la navegación del cliente y a la selección. La personalización efímera normalmente se basa en la correlación producto a producto, la recomendación basada en atributos, o ambos.

La mayoría de los sistemas de recomendación altamente personalizados usan la personalización persistente para crear recomendaciones y difieren para clientes diferentes, incluso cuando ellos están visualizando los mismos productos. Estos sistemas persistentes emplean la correlación usuario a usuario, la recomendación basada en atributos, que usan las preferencias de productos persistentes, o la correlación producto a producto basándose en las preferencias de productos persistentes.

La unión de la entrega de recomendaciones a la actividad del cliente es una decisión crítica de diseño en los sistemas de recomendación en comercio electrónico. Existen tres tipos de formas de entrega de las recomendaciones: forzadas, accesibles y orgánicas.

Las tecnologías *push* (forzadas) tienen la ventaja de asistir a un cliente cuando el cliente no está interactuando en ese momento con el sitio.

En los sistemas de recomendación que usan las tecnologías *pull* (accesibles) le permiten al cliente controlar en qué momento se muestran las recomendaciones. Éstas no se muestran de forma activa hasta que el cliente lo solicita.

Por otro lado, están las llamadas recomendaciones “orgánicas”, que realizan la entrega pasiva de la recomendación en el contexto natural del resto del sistema de recomendación para el comercio electrónico. La recomendación pasiva tiene la ventaja de asistir al cliente en el momento en que éste es ya receptivo a la idea. El comercio electrónico usa este tipo de recomendación como parte del proceso de compra, sugiriendo en tiempo real las opciones de compra.

Existen distintos métodos o aproximaciones tecnológicas a la recomendación:

- **Filtrado colaborativo (FC), del inglés *collaborative filtering***. Productos recomendados a los consumidores basados en la opinión de otros consumidores denominados "vecinos".
- **Redes Bayesianas**. Crean modelos basados en entrenamientos de árboles de decisión donde los nodos representan la información del cliente. El modelo generalmente es construido *offline*; operan en los entornos donde el conocimiento de las preferencias del usuario cambia más lentamente que el tiempo necesario para la construcción de los modelos.
- **Técnicas de *Clustering***. Identifican grupos de consumidores con similares preferencias, generan predicciones individuales promediando otros usuarios del *cluster*. Se aplican como primer paso en los algoritmos de elección de vecinos próximos.
- **Filtrado y recuperación de la información**. Se emplean en la selección y asociación de términos semánticos que interesen al usuario. Utilizados en sistemas de búsquedas de información de productos.
- **Reglas de asociación**. Analizan patrones de preferencia entre productos, recomiendan productos al consumidor basándose en otros productos seleccionados. Se utilizan en la aplicación de estrategias de ventas cruzadas.
- **Agentes**. Pueden emplearse en cualquier vertiente de la recomendación debido a la facilidad de su uso en sistemas dinámicos y adaptativos.

A continuación algunos ejemplos de sistemas de recomendación de diferente naturaleza:

- **GroupLens [70]**. Usa métodos colaborativos para predecir las preferencias de los usuarios. Estas predicciones se derivan a partir de las valoraciones, basadas en los datos navegacionales como el número de veces que un producto fue visto en la página web, la transacción histórica de un producto y los atributos del usuario.

- **Frontmind de Manna Inc. [29]**. Usa modelos de redes bayesianas para trabajar con casos de datos incompletos y determinar la distribución de probabilidad para un atributo o combinación de ellos. Los modelos también se actualizan dinámicamente con los datos *online* recogidos del cliente.

- **Learn Sesame [12]**. Define modelos de los dominios compuestos de objetos (documentos, aplicaciones, entre otros), atributos de objetos y tipos de eventos. La información del cliente se categoriza en los modelos de los dominios. Los modelos de los dominios se analizan usando *clustering* para detectar el comportamiento y patrones de preferencias de los clientes.

- **Personalization Server [60]**. Se basa en el procesamiento *offline* de clientes de distintas zonas demográficas. Clasifica a los clientes en diferentes grupos basándose en reglas definidas *a priori* por el vendedor. Las reglas obedecen a patrones de comportamiento de clientes que incluyen comportamiento en la navegación, medios de software, hardware y redes empleadas por los clientes. Una regla de clasificación se puede emplear para determinar lo que el usuario registró en una cierta clasificación a partir de los valores de propiedad del usuario y del grupo actual o de las propiedades de la solicitud o sesión. Por ejemplo, si un usuario es mayor de 35 años y menor de 65 y además de género masculino, entonces el usuario es considerado un hombre de la edad madura. Si una regla de clasificador se evalúa como correcta, entonces se devuelve un objeto de la clasificación con el mismo nombre de la regla. Los resultados de esta regla pueden ser usados por un diseñador de páginas para variar el contenido mostrado basado en una o más clasificaciones.

Actualmente y dada la importancia que va teniendo la personalización de los sitios de negocios, se hacen cada vez más deseables los sistemas adaptativos. Un sistema adaptativo es capaz de ajustar automáticamente los distintos parámetros al usuario, basándose en una

combinación de conocimientos e hipótesis tanto del usuario como de su entorno en función de unos objetivos concretos. Aunque en ocasiones, basta con que sean adaptables, de modo que se le permita al usuario manipular y modificar distintos parámetros del sistema adaptándolo a sus necesidades o expectativas [39].

Los sistemas de recomendación aplicados a las ventas en el comercio electrónico adoptan tres formas según [92]:

- **Convertir a los navegadores en vendedores**. Los visitantes a un sitio web frecuentemente exploran el sitio sin realizar compras. Los sistemas de recomendación pueden ayudar a los clientes a encontrar los productos que ellos desean adquirir.
- **Incrementar las ventas cruzadas (*cross sells*)**. Los sistemas de recomendación mejoran las ventas cruzadas mediante la sugerencia de productos adicionales que el cliente puede adquirir. Un sitio puede recomendar productos adicionales o complementarios en un proceso de chequeo, basándose en los productos que están colocados en el carrito de la compra. Por ejemplo, si el cliente ya posee en el carrito de la compra una cámara digital de fotos, el sistema podrá recomendar distintos accesorios para este producto (tarjetas de memorias, fundas de protección, baterías, cargadores de baterías, cables de interface, etc.).
- **Construir lealtad**. En el mundo del comercio electrónico ganar la lealtad del consumidor es una estrategia esencial en los negocios. Los sistemas de recomendación favorecen la lealtad mediante la creación de un valor añadido en las relaciones entre el sitio web y el cliente. Los sitios invierten en el aprendizaje sobre sus clientes, usan sistemas de recomendación para analizar las experiencias y muestran interfaces que acercan a los clientes a las necesidades que poseen. Creando relaciones entre consumidores, permitiendo que los clientes recomienden el sitio web a otros posibles compradores, se favorece la lealtad.

4.2.1 Métodos Colaborativos (Collaborative filtering)

Muchas aproximaciones se han centrado en la aplicación de técnicas inteligentes para proporcionar recomendaciones personales al usuario. *Collaborative filtering* se basa en el supuesto de que encontrando usuarios similares y examinando sus patrones de uso se pueden realizar recomendaciones útiles.

Collaborative filtering (CF) es el método de realizar predicciones automáticamente (filtrado) acerca de los intereses de un usuario a partir de la colección de información sobre los gustos de muchos usuarios (colaboración). La idea de este método es: aquellos quienes estuvieron de acuerdo en el pasado acerca de un tema tienden a estar también de acuerdo en el futuro. Por ejemplo, un sistema de recomendación para gustos sobre música podría predecir sobre qué música a un usuario le pudiese atraer mediante una lista parcial de los gustos de muchos otros usuarios.

El problema puede verse como una matriz formada por los valores de cada usuario para los distintos elementos en el conjunto de documentos, por ejemplo, la matriz contiene un conjunto de valores u_{ij}, correspondientes a los del usuario i para un elemento j. Mediante el uso de esta matriz el objetivo del filtrado colaborativo es predecir los valores de un usuario particular i, para uno o un conjunto de elementos del conjunto de documentos.

Los pasos para la predicción de estos valores para un usuario específico i, según [41] son:

- Seleccionar un conjunto de usuarios con intereses/preferencias similares al usuario i, por ejemplo, usuarios quienes tengan valores similares en los elementos o parámetros como el usuario i.

- Predecir las recomendaciones para el usuario *i* a partir del conjunto seleccionado en el paso anterior, por ejemplo, si estos usuarios puntuaron un elemento *j* con un valor alto (favorable), este elemento será recomendado al usuario *i*.

Existen distintas técnicas que se pueden usar para implementar estos dos pasos.

Los algoritmos basados en la vecindad son los más empleados. Un subconjunto de usuarios es escogido basándose en un usuario actual o activo. Estos métodos comprenden tres pasos principales:

1. Se seleccionan los usuarios con gustos similares al usuario activo (se calcula la correlación del usuario). Entre las técnicas para realizar este cálculo, están: la correlación de Pearson [83], fue utilizada en el sistema original GroupLens [87], la correlación modificada de Pearson que fue usada en el sistema Ringo [95], entre otras.
2. Un subconjunto de estos usuarios se selecciona como un conjunto de predictores (selección de vecindad). De acuerdo a los valores de correlación existen también varias técnicas para determinar el número de vecinos seleccionados: umbral de correlación y correlaciones Best-n.
3. Se calcula una predicción a partir de los valores de estos vecinos seleccionados (generación de una predicción). Una vez que la vecindad ha sido generada, se producen las predicciones. Entre las técnicas usadas para cumplir este objetivo está calcular el promedio ponderado de los valores de usuarios usando las correlaciones como pesos. Este promedio ponderado hace la suposición de que todos los valores de los elementos de usuarios tienen la misma distribución. Otra técnica es el cálculo de la media ponderada de todos los valores de los vecinos, en lugar de tomar el valor numérico explícito.

4.2.2 Métodos Basados en el Contenido (Content-Based Filtering)

Los sistemas que implementan las recomendaciones basados en el contenido son muy populares para el tratamiento de datos de tipo texto. Analizan un conjunto de documentos, usualmente textos previamente valorados por un usuario individual y construyen un modelo o perfil de los intereses del usuario basándose en las características de los objetos valorados por ese usuario [73]. El perfil se usa para recomendar los nuevos elementos de interés.

Estos métodos tienen sus inicios en la recuperación de la información. A diferencia de los métodos colaborativos éstos tienen dificultades en la recogida de diferentes aspectos del contenido (música, películas, imágenes, entre otros).

De ahí que existan sistemas de recomendación que combinan tanto métodos colaborativos como de contenido, como los sistemas Fab [4] y WebCobra [102].

4.2.3 Métodos Basados en Refuerzo

Los métodos de aprendizaje por refuerzo son aquellos que permiten mejorar el comportamiento de un sistema frente a un problema sobre la base de una señal que indica si la realización de este comportamiento ha sido adecuada o no para resolver el problema. Esta señal, llamada refuerzo o crítica), puede consistir en un valor discreto (correcto, incorrecto), o en un valor numérico que indica el grado de éxito obtenido con el comportamiento realizado [18]. Se entiende por refuerzo un estímulo que modifica un reflejo condicionado o una conducta aprendida aumentando su arraigo (refuerzo positivo o premio) o disminuyéndolo (refuerzo negativo o castigo) [5].

El aprendizaje por refuerzo es un método online, es decir, principalmente útil para sistemas que interaccionan con un entorno de forma continuada (como los sitios web dedicados al comercio electrónico) y que deben aprender a comportarse correctamente a partir de esta

interacción. Además, es adecuado cuando no existe un conocimiento a priori del entorno o éste es demasiado complejo como para utilizar otros métodos [18].

Uno de los sistemas de recomendación con este método es *News Dude*. Este sistema tiene como objetivo hallar noticias que aun no hayan sido leídas y que puedan ser interesantes para el usuario. Estas noticias se presentan en cuatro categorías (valores de refuerzo): no apropiado pero interesante, no apropiado por redundante, apropiado, muy apropiado [44]. Una noticia no leída podrá clasificarse en una de estas cuatro categorías, distinguiendo dos perfiles si la noticia tiene interés a corto plazo o a largo plazo.

5 Multiclasificadores en Minería Web

En los ambientes basados en páginas web existe heterogeneidad en los datos recogidos. Cuando existe una amplia variedad de datos, la actuación de clasificadores de modo individual falla en algunas regiones de los datos de entrenamiento aunque a veces funciona de forma correcta. De ahí, a que en muchas ocasiones sea más factible el uso de multiclasificadores. Los multiclasificadores son el resultado de combinar varios clasificadores individuales. Los métodos de construcción de multiclasificadores se dividen en dos grupos: métodos de ensamble y métodos híbridos. Los primeros métodos, como *Bagging* [10] y *Boosting* [32], inducen modelos que combinan clasificadores con el mismo algoritmo de aprendizaje, y en el caso de estos dos ejemplos, introducen modificaciones en el conjunto de datos de entrenamiento. El segundo tipo de métodos, como *Stacking* [104], crea nuevas técnicas de aprendizaje híbrido a partir de diferentes algoritmos de base..

En un estudio anterior [94] se describieron las arquitecturas y principales métodos de los multiclasificadores y se demostró mediante un caso de estudio su capacidad de aumentar la precisión con respecto a los clasificadores individuales que los conforman.

Una condición necesaria y suficiente para que una combinación de clasificadores obtenga un resultado más preciso que cualquiera de los clasificadores que la componen, es que los clasificadores sean a su vez suficientemente precisos y diversos [43]. Un clasificador es preciso si tiene un error menor que el que se obtendría eligiendo una clase arbitrariamente. Dos clasificadores son diversos si cometen diferentes errores en los datos de entrada (errores que ya hemos visto pueden estar provocados por datos o por representaciones o atributos no adecuados).

La mayoría de los trabajos realizados con el uso de multiclasificadores en minería web están dirigidos a la clasificación de textos (minería de textos) y enmarcados en el contenido de la web.

En la actualidad la categorización de textos ha sido tema de muchas investigaciones mediante el uso de diferentes tipos de algoritmos. Con este objetivo se han desarrollado estudios sobre el estado del arte de algoritmos de clasificación como máquinas de soporte vectorial, vecino más cercano y redes neuronales para obtener buenos desempeños en comparación a otros clasificadores existentes en la literatura [107], [108], [100]. Con el propósito de combinar las ventajas de diferentes enfoques de clasificación y así incrementar la precisión general, también se han estudiado artículos que proponen el aprendizaje multi-estratégico o combinación de clasificadores como [91].

En [108] se propuso un Sistema Generador de Mejores Resultados Globales (*Best Overall Results Generator System*) que combina linealmente métodos de clasificación usando el mismo peso para cada clasificador individual en el descubrimiento del tema y el dominio de rastreo. Algunos de los métodos de clasificación usados fueron: Rocchio [88], vecino más cercano y modelación del lenguaje. En [64] se demostró el incremento del rendimiento en el contenido de

la categorización del texto mediante el uso de una nueva formulación de consultas y los métodos de pesos, combinando tres clasificadores independientes (vecino más cercano, retroalimentación de relevancia y un clasificador bayesiano). Varios investigadores [48] examinaron distintas estrategias de combinación en el filtrado del contenido de los documentos con algoritmos de aprendizaje como Rocchio, vecino más cercano, análisis de discriminante lineal y red neuronal. En [15] se presentó la evaluación de métodos de votación y meta-aprendizaje en los datos divididos a través del aprendizaje inductivo. En [101] se demostró la efectividad de la generalización del método de *Stacking* (método de acumulación) para combinar diferentes tipos de algoritmos de aprendizaje.

Más recientemente, en [81] se presenta un nuevo método para la clasificación de páginas web que usan datos no-etiquetados para completar el número limitado de datos etiquetados existente. El método de aprendizaje propuesto primero entrena a un clasificador con un pequeño conjunto de entrenamiento de datos etiquetados y se determina la confianza del clasificador. Posteriormente, se construye una serie de clasificadores de forma secuencial con datos no etiquetados.

En [31] se describe BWI (*Boosted Wrapper Induction*) un enfoque que construye un sistema de extracción de la información, donde se combinan las técnicas de inducción *wrapper* [62], [45], [63], [76] con el algoritmo AdaBoost [32].

En [66] se utilizó un multiclasificador de Máquinas de Soporte Vectorial (MSV) para la clasificación de documentos web. Aunque los MSV están limitados por el número de clases, en esta investigación se intenta diseñar un clasificador MSV apropiado para problemas multiclase.

El problema de clasificación de páginas web es de gran escala tanto por la dimensión de las características como por el número total de muestras, lo que influye en el tiempo de entrenamiento y en el espacio de almacenamiento por el diseño de clasificadores. Por esta razón, es un problema multiclase. Algunos clasificadores tradicionales y muy complejos, como las redes neuronales multicapa y *clustering*, son muy difíciles de implementar bajo condiciones experimentales. Se requiere de un clasificador que pueda garantizar una alta precisión y al mismo tiempo sea factible de diseñar. Los clasificadores MSV son clasificadores ideales para este problema pero se necesita que puedan adoptar el esquema multiclase. La implementación clásica para la clasificación multiclase de MSV es la estrategia “uno contra todos” (*one-against-all*) y otra estrategia útil es “uno contra uno” (*one-against-one*) [46].

En [113] se realizó también una investigación dirigida a la clasificación de páginas web. Los autores propusieron un algoritmo llamado tri-training que aborda el problema de determinar cómo etiquetar los ejemplos sin etiqueta y cómo producir la hipótesis final. Por otra parte, se alcanza una capacidad mejor de generalización al combinar tres clasificadores. Se denota como L al conjunto de ejemplos etiquetados y U al conjunto de ejemplos no etiquetados. Entonces para cualquier clasificador, un ejemplo no etiquetado puede serlo por uno de los clasificadores si los otros dos clasificadores están de acuerdo en etiquetar este ejemplo. Es decir, si se tiene que los clasificadores h_2 y h_3 están de acuerdo en etiquetar un ejemplo x en U, entonces x puede ser etiquetado por h_1. Esto significa que los tres clasificadores participan en un proceso de refinamiento. La hipótesis final se produce mediante el voto mayoritario. La generación de los clasificadores iniciales es similar al entrenamiento de un algoritmo ensamble como *Bagging* a partir de ejemplos etiquetados.

En la web semántica son muy utilizadas las taxonomías. Existe un número elevado de taxonomías en la web y frecuentemente se necesitan integrar los objetos a partir de varias taxonomías en una taxonomía *master*. En [112] se describe la técnica Co-Bootstrapping, donde el algoritmo AdaBoost y las Máquinas de Soporte Vectorial se mejoran para la integración de taxonomías.

A pesar de la abundancia de investigaciones dirigidas al uso de multiclasificadores para minería de textos, uno de los principales objetivos de este trabajo es conocer el uso de combinaciones de clasificadores en el comercio electrónico para contribuir al descubrimiento de patrones de usuarios que visitan los sitios web o como parte de sistemas de recomendación.

A continuación se describen algunos algoritmos que han sido utilizados para construir combinaciones de clasificadores en las investigaciones antes mencionadas y en minería del uso de la web.

5.1 Algoritmo Rocchio

El algoritmo Rocchio tiene como idea la formulación y la ejecución de una consulta inicial. El usuario examina la información recuperada y determina cuál le resulta relevante y cuál no. Con estos datos, el sistema genera automáticamente una nueva consulta, basándose en los documentos que el usuario señaló como relevantes o no. El algoritmo Rocchio proporciona un sistema para construir el vector de la nueva consulta, recalculando los pesos de los términos de ésta y aplicando un coeficiente a los pesos de la consulta inicial, otro a los de los documentos relevantes y otro distinto a los que no lo son.

En el ámbito de la categorización, el mismo algoritmo Rocchio proporciona un sistema para construir los patrones de cada una de las clases o categorías de documentos. Así, partiendo de una colección de entrenamiento, categorizada manualmente de antemano, y aplicando el modelo vectorial, se puede construir vectores patrón para cada una de las clases, considerando como ejemplos positivos los documentos de entrenamiento de esa categoría, y como ejemplos negativos los de las categorías restantes [28].

Una vez que se tienen los patrones de cada una de las clases, el proceso de entrenamiento o aprendizaje concluye. La categorización de nuevos documentos se realiza mediante la estimación de la similitud entre el nuevo documento y cada uno de los patrones. El que consigue un índice mayor indica la categoría a la que se debe asignar ese documento.

Rocchio emplea la retroalimentación de relevancia que es una técnica usada en los sistemas tradicionales de recuperación de la información (fundamentalmente textos) y que consiste en ajustar automáticamente una pregunta o consulta existente que usa la retroalimentación de la información del usuario sobre la relevancia de objetos previamente recuperados [90]. Es un proceso en el cual el usuario juzga la calidad de la información recuperada, posteriormente esta información es refinada, y se repite su refinamiento hasta que el usuario esté satisfecho de los resultados [47].

En [103] se compararon dos sistemas, uno con retroalimentación de relevancia explícita (donde los usuarios explícitamente marcan los documentos que consideran relevantes) y otro que utiliza la retroalimentación de relevancia implícita (donde el sistema se esfuerza por estimar la relevancia a través de aplicar la minería a la consulta emitida por el usuario). La retroalimentación se utiliza para actualizar la pantalla de acuerdo a la interacción del usuario.

5.2 Clasificación por Modelos del Lenguaje (*Language Modeling*)

El proceso de clasificación de preguntas se hace más dinámico y automático mediante la modelación del lenguaje, este algoritmo estadístico ha ganado mucha atención recientemente en el área de la recuperación de la información. En este algoritmo los modelos se pueden construir automáticamente a partir del conjunto de entrenamiento y su funcionamiento es competitivo en relación a otros enfoques [65].

La idea básica de la modelación del lenguaje es que cada fragmento de texto puede verse como que ha sido generado a partir de un modelo del lenguaje. Si se tienen dos fragmentos de texto, se puede definir el grado de relevancia entre ellos como la probabilidad con que ellos se generan por el mismo modelo del lenguaje. En el área de recuperación de información, se puede construir un modelo del lenguaje para cada documento *D*. Dada una consulta, se puede decidir si un documento es relevante basándose en la probabilidad con la que su modelo del lenguaje genera tal consulta. Si la consulta *Q* se compone de *n* fichas (*tokens*): w_1, w_2,… w_n, se puede calcular la probabilidad como:

$$P(Q|D) = P(w_1|D) * P(w_2|D,w_1) * ... * P(w_n|D, w_1, w_2, ..., w_{n-1})$$

Se han introducido ideas similares en tareas de clasificación de preguntas. Se construye un modelo del lenguaje para cada categoría *C* de preguntas. Cuando una nueva pregunta *Q* se realiza, se calcula la probabilidad $P(Q|C)$ para cada *C* y se escoge la de probabilidad más alta. La mayor ventaja del modelo del lenguaje es su flexibilidad. La expresión regular del modelo está compuesta de reglas *hard-coded* (fuertemente codificadas) que necesitan que se modifiquen para poder aplicarse a los nuevos casos. El modelo del lenguaje, sin embargo, se puede mantener automáticamente. Con conjuntos grandes de datos de entrenamiento la actuación del modelo del lenguaje se puede mejorar.

5.3 Combinación de Clasificadores en Minería del Uso de la Web

Una evaluación empírica de esquemas de combinación de clasificadores para predecir el comportamiento de navegación de los usuarios se presentó en [75]. La información puede usarse, en primer lugar, para automatizar la evaluación de la utilidad del sitio web con el objetivo de mejorar su arquitectura y el diseño para interacción con el cliente y en segundo lugar para propiciar la navegación web adaptativa [96].

La principal idea en esa investigación fue combinar diferentes clasificadores que se crearon a partir de distintos conjuntos de entrenamiento para predecir el comportamiento del usuario y para estudiar la evolución de tales predicciones. Se estudiaron dos paradigmas clásicos: combinación a través del voto [54] y *Cascading* (método en cascada) [33], con el objetivo de construir un sistema de clasificación dinámico.

En minería web existen dos formas de realizar la combinación de múltiples clasificadores según [71]:

- ***Offline***: Es la forma más sencilla para encontrar los clasificadores y seleccionar el de menor error para un conjunto de datos específico.
- ***Online***: Usa a los clasificadores y luego aplica el voto mayoritario. La clase que obtenga el máximo número de votos entre los clasificadores individuales será la asignada para el ejemplo de prueba en curso. No obstante, también se realizaron estudios para la clase con más del 50% de los votos y también para la que obtuviese más del 75% de los votos.

En [71] se describió un sistema para la clasificación de estudiantes con el objetivo de predecir su calificación o nivel final, basándose en las características extraídas de los datos *logs* del sistema *online* creado por la Universidad de Michigan, *Learning Online Network with Computer-Assisted Personalized Approach* (Red de Aprendizaje en Línea con Enfoque Personalizado y Computación Asistida). Para ello utilizaron una combinación de múltiples clasificadores que permitieron mejorar significativamente la precisión de la clasificación y posteriormente, usaron un algoritmo genético para optimizar la combinación. El algoritmo genético intenta encontrar una población con los mayores pesos para cada vector de las

características, que está compuesto por un conjunto de seis variables para cada estudiante. Los clasificadores que se incluyen en este estudio son de distintas técnicas de aprendizaje, como son: clasificador bayesiano, vecino más cercano, aproximación gaussiana, perceptrón multicapa y árbol de decisión.

En el sistema *News Dude* [7] se enfocan dos problemas relacionados con la inducción automatizada de perfiles de usuario para la clasificación de noticias; motivan la inducción de un modelo híbrido de usuario compuesto por modelos separados para los intereses de un usuario a corto y a largo plazo. Se usa el algoritmo del vecino más cercano para atender a los intereses a corto plazo y un clasificador bayesiano para los intereses a largo plazo.

En [79] se presenta una propuesta para combinar las predicciones dadas por múltiples modelos de usuario que se denomina inducción de la aplicabilidad del modelo (MAI, del inglés *Model Applicability Induction*). Este método consiste en caracterizar las situaciones en las que cada modelo es capaz de hacer predicciones correctas. Esto es posible gracias a la construcción de un árbitro o *referee* para cada modelo disponible. Este árbitro es el encargado de determinar para cada instancia de entrada si su modelo correspondiente lo clasificará correctamente. Si es así, se tomará en cuenta su predicción y no se hará en caso contrario.

6 Caso de estudio

MovieLens es un sistema de recomendación de películas, accesible a través de Internet y basado en la tecnología GroupLens, que como ya se ha mencionado, pertenece a la categoría de filtrado colaborativo. MovieLens es una fuente experimental de datos. En el sitio web http://www.cs.umn.edu/Research/GroupLens/ aparecen disponibles actualmente dos bases de datos. Una de ellas, ha sido seleccionada en este estudio para la predicción de las valoraciones (de 1 a 5) de películas mediante 100000 registros para 1682 películas por 943 usuarios. Finalmente, en este trabajo se procesaron solamente 1240 registros, debido a que se tuvieron que eliminar registros en la fase de procesamiento.

Los datos que aparecen en el sitio web, se encuentran en distintos archivos por lo que se realizaron las operaciones pertinentes para agruparlos en uno que sirvieron de entrada a las herramientas informáticas *Mineset* de *Silicon Graphics, Inc.* y *WEKA* (*Waikato Environment for Knowledge Analysis*) versión 3.4 de la Universidad de Waikato.

La valoración (“rating”) es un dato en una escala de 1 a 5 sobre de la opinión que emiten los usuarios acerca una película, donde el valor 1 se corresponde con la más baja valoración o preferencia por una película y el valor 5 representa la máxima valoración. De cada usuario se conoce su género (“gender”), su edad (“age”), su ocupación (“occupation”) y código postal en el que vive (“zipcode”). Entre los atributos correspondientes a las películas están el título (“movietitle”), la fecha de su lanzamiento (“releasedat”) y un atributo para cada uno de los géneros o categorías (unknown, action, adventure, animation, children, comedy, crime, documentary, drama, fantasy, film-noir, horror, musical, mystery, romance, science-fiction, thriller, war y western) que toman valor 1, si la película pertenece a ese género y 0, en caso contrario. Esto significa que la película puede pertenecer a distintos géneros cinematográficos. En la base de datos original aparece el atributo “videoreleasedat”, que ha sido directamente excluido por no tener información registrada. Existe además, el atributo “timestamp” que se refiere al momento en que se realizó la valoración en formato GMT.

Antes de realizar el análisis de los atributos que finalmente tomaron parte en la construcción de los modelos, se puede inferir que los datos presentan ruido, pues no se entiende que existan valoraciones superiores a 2 cuando no se conoce el género al que pertenece la película ni su título, existen además otros registros en los que los usuarios no poseen código postal o que presentan menos de cinco dígitos. No obstante, en *Mineset* se generaron los

gráficos de cajas e histogramas (Figura 5), según el tipo de variable para realizar un análisis más exhaustivo sobre la distribución de valores de los atributos.

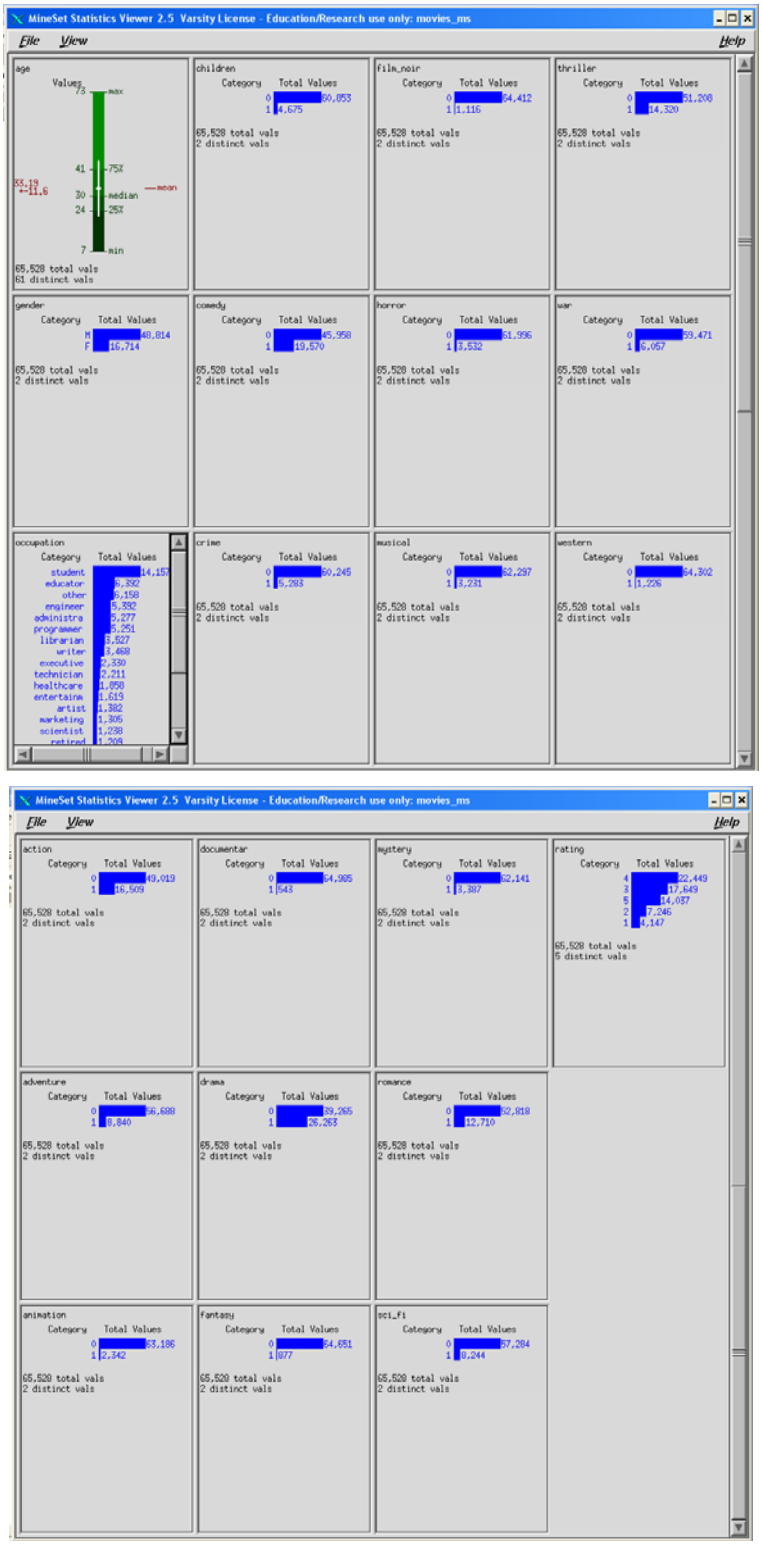

Figura 5. Visualización estadística de los datos.

En la figura 5 se muestra que los datos al parecer no tienen ruido, debido a que el comportamiento de cada variable está acorde a su naturaleza. Por ejemplo, la variable numérica "age", que corresponde a la edad del usuario va desde un rango de 7 a 73 años. Las variables nominales que representan los posibles géneros (18) en los que puede estar catalogada cada película muestran una distribución que no es anormal, una película pertenece o no a un género. Por citar dos ejemplos, es lógico encontrar que exista una buena distribución en el género "drama" entre los dos valores y que exista un porcentaje bajo de películas que sean documentales ("documentar") ya que es un género que no abunda.

En cuanto a la variable a predecir en el estudio ("rating"), de cinco categorías, se realizaron las modificaciones necesarias para convertirla en la variable ("recom") con sólo dos posibles valores: "No recomendar" para los valores de 1 y 2, y "Sí recomendar" para los valores entre 3 y 5, véase la figura 6. Estos cambios se produjeron para simplificar el problema, ya que lo importante es predecir si una puede ser de interés para los usuarios o no, además como el número de variables es alto (22) se reduce el tiempo de construcción de los modelos.

Figura 6. Representación de la modificación de categorías de la etiqueta.

Con vistas a realizar un estudio en el que los multiclasificadores podrían ser utilizados en la recomendación de películas, se analizó el comportamiento de *Bagging*, *Boosting* y *Stacking* con distintas técnicas de aprendizaje.

En *WEKA* para este caso de estudio los clasificadores individuales a través de distintos algoritmos mostraron altas precisiones (Figura 7), por lo que pudiera no justificarse el uso de multiclasificadores que aumentan el tiempo de construcción y evaluación de los modelos para sólo superar por centésimas a la mayoría de los clasificadores individuales y en otros ni tan siquiera eso, véase el Cuadro 1.

Los tiempos de construcción y evaluación de los algoritmos individuales son cortos respecto a los que muestran *Bagging* y *Boosting* (Cuadro 2). Este último aumenta su tiempo de ejecución significativamente para la técnica de aprendizaje del vecino más cercano. A través del análisis de la figura 7, donde el algoritmo individual del vecino más cercano presenta el menor valor de precisión en relación con el método de aprendizaje bayesiano y al de árbol de decisión, se puede descartar tanto al clasificador individual del vecino más cercano como a los métodos ensamble *Bagging* y *Boosting* con este aprendizaje para utilizarlo en un sistema de recomendación de películas.

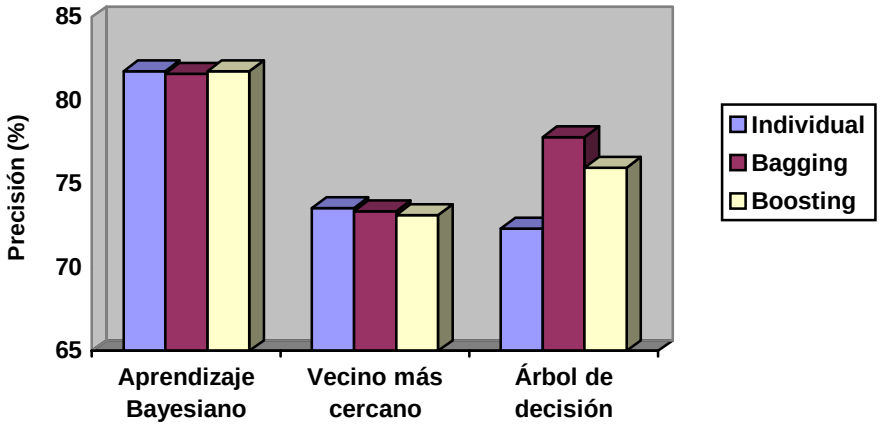

Figura 7. Precisión de tres algoritmos individuales y los multiclasificadores Bagging y Boosting.

Cuadro 1. Valores de la precisión (%) para los algoritmos individuales y los multiclasificadores Bagging y Boosting.

	Individual	Bagging	Boosting (AdaBoot)
Aprendizaje bayesiano (BayesNet)	81,77	81,61	81,77
Vecino más cercano (IB1)	73,54	73,39	73,14
Árbol de decisión (RandomTree)	72,34	77,82	75,97

Cuadro 2. Valores de los tiempos de ejecución (segundos) para los algoritmos individuales y los multiclasificadores Bagging y Boosting.

	Individual	Bagging	Boosting (AdaBoot)
Aprendizaje bayesiano (BayesNet)	3	7	10
Vecino más cercano (IB1)	8	24	1354
Árbol de decisión (RandomTree)	8	8	24

Sin embargo, con árbol de decisión se obtuvieron multiclasificadores que mejoraron considerablemente la precisión, tanto *Bagging* como *Boosting* aumentaron la precisión en relación con el método individual. A la inversa del caso de estudio de evaluación de tierras descrito recientemente en [94] los datos actuales sí presentan ruido, pues como se explicó anteriormente existen irregularidades en los datos que hacen pensar que hubo variación en los datos reales en el momento de hacer la limpieza y durante el preprocesamiento de los datos por parte de los suministradores. En este estudio, con árboles de decisión se demuestra que *Bagging* se comporta mejor que *Boosting* (*AdaBoost*), lo que ratifica los estudios hechos en [23].

El método híbrido, *Stacking*, a su vez, incrementó en un 8% la precisión alcanzada por el clasificador individual de árbol de decisión. Las técnicas de aprendizaje que se utilizaron para

construir el clasificador *Stacking* fueron dos árboles de decisión como algoritmos de base (*RandomTree* y *J48*) y para aprender el meta-modelo se usó el algoritmo *IB1* correspondiente al algoritmo del vecino más cercano, véase el Cuadro 2.

Cuadro 2. Valores de precisión (%) de RandomTree y Stacking.

RandomTree	Stacking
72,34	80,08

A pesar del incremento de la precisión mediante *Stacking*, no se logra superar la precisión alcanzada por el método individual del aprendizaje bayesiano, además el tiempo de ejecución de este modelo es el más rápido, solamente 3 segundos, por lo que para este caso de estudio no se recomienda el uso de métodos multiclasificadores.

Los multiclasificadores son sensibles a la calidad de los datos procedentes de la web. Su aplicación en sistemas de recomendación debe tenerse en cuenta si el tiempo que se emplea construcción de los modelos no es prolongado, ya que para este tipo de sistema la inmediatez es uno de los principales factores a considerar como requisito indispensable. Esto representa un problema importante para algoritmos como el vecino más cercano que deben ejecutarse cada vez que se realiza una recomendación, sin embargo cuando los modelos se construyen previamente (*offline*) el tiempo en que se tarda en inducir el clasificador no repercute en el tiempo de respuesta al usuario. Estos últimos modelos son adecuados cuando las preferencias de los usuarios no cambian muy rápidamente con el tiempo, con lo cual los modelos construidos son aplicables durante un período de tiempo aceptable.

7 Conclusiones

Las características de la *World Wide Web* y la necesidad de obtener distintos tipos de información de la misma, presentan un nuevo desafío relativo tanto a los métodos de recuperación de información tradicionales como a las técnicas aplicadas para analizar los datos registrados por los servidores. Se puede observar que en ambos aspectos todavía existen bastantes deficiencias en la recolección y análisis de los datos.

La minería de uso de la web puede aportar una información valiosa, tanto a los gestores de servicios de información como a los proveedores de comercio electrónico. Esta tecnología, utilizada ampliamente en el mundo del *marketing* convencional tiene en la web unas aplicaciones potentes e inmediatas. La gran cantidad de datos que se acumulan en los ficheros *log* de los servidores, unidos a los datos de ventas es una fuente de información que no se debe despreciar. Técnicas como el agrupamiento automático de clientes, la clasificación de los usuarios y la personalización de servicios, permiten tomar una posición en el mercado que puede diferenciar a una empresa de sus competidores.

El comercio electrónico induce a la introducción de las nuevas tecnologías como un paso natural del comercio tradicional, entre ellas la minería web que analiza los patrones de navegación de los usuarios. Los sistemas de recomendación integran gran variedad de características requeridas en el negocio del comercio electrónico (*marketing*, tratamiento personalizado del cliente, acceso a grandes bases de datos). La personalización contribuye a incrementar las relaciones de fidelidad y confianza del usuario en el negocio virtual. La privacidad de los datos de los usuarios constituye un punto importante a resolver en los sistemas de recomendación de ahí la importancia del análisis de los perfiles de usuarios.

El uso de agentes, conjuntos difusos, algoritmos genéticos, entre otros mecanismos, han sido aplicados con éxito en la minería web.

El uso de los multiclasificadores en minería web es más limitado que en la minería de datos tradicional. La combinación de clasificadores en la web es más frecuente en el área de minería del contenido de la web, aunque también se han aplicado en la predicción del comportamiento del usuario y para estudiar la evolución de tales predicciones.

El caso de estudio incluido en este trabajo corroboró que si los datos presentan ruido *Bagging* muestra su superioridad con relación a *Boosting* cuando se construyen a partir de árboles de decisión y que se puede construir un método híbrido a través de *Stacking* que a la vez sea aun mejor que los dos anteriores.

En la decisión de usar multiclasificadores en los sistemas de recomendación se debe considerar cuándo se construyen los modelos y qué algoritmos se deben elegir para emitir las recomendaciones a los usuarios en el ambiente web. Si se procesan muchos datos, algoritmos como el vecino más cercano, que efectúan la inducción del modelo *online*, no son idóneos debido a que aumentaría mucho el tiempo de respuesta al usuario. En esos casos, y cuando las preferencias de los usuarios no cambian muy rápidamente se pueden utilizar modelos construidos *offline*, ya que el tiempo empleado en su inducción no va a repercutir en el tiempo de respuesta al usuario. Es necesario tener en cuenta estas consideraciones a la hora de decidir si usar o no multiclasificadores en sistemas de recomendación debido a que, en general, los multiclasificadores aumentan la precisión en comparación con los clasificadores individuales, pero su ejecución es más lenta.

Referencias

[1] Aldana, J.F., Gómez, A.C., Roldán, M.M., Moreno, N., Nebro, A.J.: Metadata functionality for semantic web integration. Proceedings of the Seventh International of the International Society of Knowledge Organization (ISKO'02) Conference, July 10-13, 2002, Granada.

[2] Baeza-Yates, R., Poblete, B.: Una herramienta de minería de consultas para el diseño del contenido y la estructura de un sitio Web. I Congreso Español de Informática, Actas del III Taller de Minería de Datos y Aprendizaje (TAMIDA'2005), 13-16 septiembre 2005, Granada, España, Thomson, 39-48, ISBN 84-9732-449-8.

[3] Baeza-Yates, R., Ribiero-Neto, B.: Modern Information Retrieval. Readings, MA, Addison-Wesley Longman, 1999.

[4] Balabanovic, M., Shoham, Y.: Fab: Content- Based, Collaborative Recommendation. Communications of the ACM, vol. 40, 3, 66-72, 1997.

[5] Balbotín, D.: Aprendizaje por Refuerzo en Tablas. Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla, 1998. http://www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web10/indice.html

[6] Berendt, B., Mobasher, B., Spiliopoulou, M., Wiltshire, J.: Measuring the accuracy of sessionizers for web usage analysis. In Workshop on Web Mining at the First SIAM International Conference on Data Mining, April 5-7, 7-14, Chicago, USA, 2001.

[7] Billsus, D., Pazzani, M.: A Hybrid User Model for News Story Classification. Proceedings of the Seventh International Conference on User Modeling, 99-108, Banff, Canada, June 20-24, 1999.

[8] Borges, J., Levene, M.: Mining navigation patterns with hypertext probabilistic grammars. Research Note RN/99/08, Department of Computer Science, University College London, Gower Street, London, UK, February 1999.

[9] Boughanem, M., Dkaki, T., Mothe, J., Soule-Dupuy, C.: Mercure at trec7. Proceedings of the 7th International Conference on Text Retrieval, TREC7, NIST SP 500-236, 355-360, November, 1997.

[10] Breiman, L.: Bagging predictors. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 24, 2, 123-140, 1996.
[11] Cachero, C., Gómez, J., Pastor, O.: Modelando aspectos de navegación y presentación en aplicaciones hipermediales. Actas de la V Jornada de Ingeniería del Software y Base de Datos, Valladolid, España, Noviembre 2000. http://www.dlsi.ua.es/~cachero/pPublicaciones.htm
[12] Caglayan, A., Snorrason, M., Jacoby, J., Mazzu, J., Kumar, R.: Learn Sesame: A Learning Agent Engine. Applied Artificial Intelligence, 11, 393-412, 1997.
[13] Campitelli, A., Rosso, C.L.: Comercio electrónico. Monografías. Revisado 2005. http://www.monografias.com/trabajos12/monogrr/monogrr.shtml
[14] Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, vol. 27, 6, 1065-1073, 1995.
[15] Chan, P., Stolfo, J.: Comparative evaluation of voting and meta-learning on partitioned data. Proceedings of the International Conference on Machine Learning (ICML '95), 90-98, 1995.
[16] Chen, M.S., Park, J.S., Yu, P.S.: Efficient data mining for path traversal patterns. IEEE Transactions on Knowledge and Data Engineering, vol. 10, 2, 209-221, March/April 1998.
[17] Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, vol. 1, 1, 5-32, 1999.
[18] Cortés, U., Moreno, A., Armengol, E., Béjar, J., Belanche, L., Gavaldà, R., Gimeno, J.M., López, B., Martín, M., Sánchez, M.: Aprendizaje Automático. Ediciones UPC, 1994. http://www.edicionsupc.es
[19] Cruz, R.A.P.P., García, F.J., Alonso, L.: Perfiles de Usuario: En la Senda de la Personalización. Informe Técnico – Technical Report DPTO-IT-2003-001. Departamento de Informática y Automática, Universidad de Salamanca, España, Enero 2003.
[20] Cueva, J.M.: Panorámica actual de la Ingeniería Web. Departamento de Informática. Universidad de Oviedo. OOTLab, 2004. http://www.ootlab.uniovi.es
[21] DAEDALUS – Data, Decisions and Language, S.A.: Minería Web: Documentos básico DAEDALUS. White Paper, C-26-AB-6002-010, Noviembre 2002. http://www.daedalus.es
[22] Dasgupta, P., Melliar-Smith, P.M.: Dynamic Consumer Profiling and Tiried Pricing Using Software Agents. Electronic Commerce Research, Kluwer Academic Publishers, Manufactured in The Netherlands, 3, 277-296, 2003.
[23] Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 40, 2, 139-157, 2000.
[24] Dürsteler, J.C.: Minería Web. Revista digital de InfoVis.net, 2005. http://www.infovis.net/printMag.php?num=172&lang=1
[25] Eliassi-Rad, T., Shavlik, J.: A System for Building Intelligent Agents that Learn to Retrieve and Extract Information. International Journal on User Modeling and User-Adapted Interaction, Special Issue on User Modeling and Intelligent Agents, 13, 35-88, 2003.
[26] Etzioni, O.: The World Wide Web: quagmire or gold mine? Communications of the ACM, vol. 39, 11, 65-68, 1996.
[27] Etzioni, O., Shakes, J., Langheinrich, M.: Ahoy! The homepage finder. Proceedings of the 6th WWW Conference, April 1997, Santa Clara, California.
[28] Figuerola, C.G., Alonso, J.L., Zazo, A.F., Rodríguez, E.: Algunas Técnicas de Clasificación Automática de Documentos. Cuadernos de Documentación Multimedia, vol. 15, 2004. http://multidoc.rediris.es/cdm/
[29] Fink, J., Kobsa, A.: A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web. User Modeling and User-Adapted Interaction, Kluwer Academic Publishers, The Netherlands, 10, 209-249, 2000.

[30] Freitag, D.: Information Extraction from Html: Applications of a General Machine Learning Approach. Proceedings of the 15th Conference of American Association for Artificial Intelligence AAAI-98, 517-523, 1998.
[31] Freitag, D., Kushmerick, N.: Boosted wrapper induction. Proceedings of the 17th National Conference on Artificial Intelligence AAAI-2000, 577-583, 2000.
[32] Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning, 148-156, 1996.
[33] Gama, J., Brazdil, P.: Cascade Generalization. Machine Learning, Kluwer Academic Publishers, Boston, Manufactured in The Netherlands, vol. 41, 3, 315-343, 2000.
[34] García, F.J., Gil, A.: Personalización de Sistemas de Recomendación. Actas del Workshop de Investigación sobre nuevos paradigmas de interacción en entornos colaborativos aplicados a la gestión y difusión del Patrimonio cultural (COLINE'02), Granada, Noviembre 2002.
[35] Gaudioso, E.: Contribuciones al Modelado del Usuario en Entornos Adaptativos de Aprendizaje y Colaboración a través de Internet mediante técnicas de Aprendizaje Automático. Tesis Doctoral. Dpto. de Inteligencia Artificial, Facultad de Ciencias, Universidad Nacional de Educación a Distancia, Madrid, 2002.
[36] Gedeon, T., Koczy, L.: A model of intelligent information retrieval using fuzzy tolerance relations based on hierarchical co-occurrence of words. Soft Computing in Information Retrieval: Techniques and Applications, F. Crestani and G. Pasi (Eds), Heidelberg, Germany: Physica-Verlag, vol. 50, 48–74, 2000.
[37] Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext transfer protocol - http/1.1. RFC2616, 1999. http://www.w3.org/Protocols/rfc2616/rfc2616.html
[38] Gibson, D., Kleinberg, J., Raghavan, P.: Inferring Web Communities from Link Topologies. Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space - Structure in Hypermedia Systems, 225-234, June 1998.
[39] Gil, A., García, F.J.: Recomendación y Personalización en Aplicaciones de Comercio Electrónico. Proceedings of European E-Commerce Workshop 2002 (EECW'02), Salamanca, Octubre 2002.
[40] González, G., Delfín, S., Lluís, J.: Preprocesamiento de bases de datos masivas y multi-dimensionales en minería de uso web para modelar usuarios: comparación de herramientas y técnicas con un caso de estudio. I Congreso Español de Informática, Actas del III Taller de Minería de Datos y Aprendizaje (TAMIDA'2005), 13-16 septiembre 2005, Granada, España, Thomson, 193-202, ISBN 84-9732-449-8.
[41] Griffith, J., O'Riordan, C.: Collaborative Filtering. Technical Report of the Department of Information Technology, National University of Ireland, Galway, 2000.
[42] Gyenesei, A.: A Fuzzy Approach for Mining Quantitative Association Rules. Acta Cybernetica, vol. 15, 2001. http://www.cs.aue.auc.dk/datamining/papers/tr336.ps
[43] Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, 10, 993-1001, 1990.
[44] Hernández-Orallo, J., Ramírez, M.J., Ferri, C.: Introducción a la minería de datos. Pearson Educación, S.A., Madrid, 2004.
[45] Hsu, C.N., Dung, M.T.: Generating finite-state transducers for semistructured data extraction from the web. Information Systems, vol. 23, 8, 521-538, 1998.
[46] Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, vol. 13, 2, 415-425, March 2002.
[47] Huang, X., Chen, S.C., Shyu, M.L., Zhang, C.: User Concept Pattern Discovery Using Relevance Feedback and Multiple Instance Learning For Content-Based Image Retrieval. Proceedings of the Third International Workshop on Multimedia Data Mining, MDM/KDD'2002, 100-108, Edmonton, Alberta, Canada, July 23rd, 2002.

[48] Hull, D.A., Pedersen, J.O., Hinrich, S.: Method combination for document filtering. Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, 279-288, 1996.
[49] Joshi, A., Krishnapuram, R.: Robust fuzzy clustering methods to support web mining. Proceedings of Workshop in Data Mining and Knowledge Discovery, SIGMOD, 15-1–15-8, 1998.
[50] Kerschberg, L., Kim, W., Scime, A.: WebSifter II: A Personalized Meta-Search Agent Based on Weighted Semantic Taxonomy Tree, International Conference on Internet Computing (IC'2001), Las Vegas, NV, 14-20, June, 2001.
[51] Kim, S., Zhang, B.T.: Web document retrieval by genetic learning of importance factors for html tags. Proceedings of PRICAI 2000 Workshop on Text and Web Mining, Melbourne, Australia, 13-23, 2000.
[52] Kimball, R., Merz, R.: The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. John Wiley & Sons, 2000.
[53] Kim, J.K., Cho, Y.H., Kim, W.J., Kim, J.R., Suh, J.H.: A personalized recommendation procedure for Internet shopping support. Electronic Commerce Research and Applications, vol. 1, 301-313, 2002.
[54] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, 3, 226-239, March 1998.
[55] Kobayashi, M., Takeda, K.: Information retrieval on the web. ACM Computing Surveys (CSUR), vol. 32, 2, 144-173, June 2000.
[56] Kohavi, R.: Mining e-commerce data: The Good, the Bad, and the Ugly. Proceedings of the Seven ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2001.
http://robotics.Stanford.EDU/users/ronnyk/goodBadUglyKDDItrack.pdf
[57] Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence, vol. 97, 1-2, 273-324, 1997.
[58] Kohavi, R., Masand, B., Spiliopoulou, M. and Srivastava, J.: Web Mining. Data Mining and Knowledge Discovery. Kluwer Academic Publishers. Manufactured in The Netherlands, 6, 5-8, 2002.
[59] Kohavi, R., Provost, F.: Applications of data mining to electronic commerce. Data Mining and Knowledge Discovery, vol. 5,1-2, 5-10, 2001.
[60] Krame, M.I.: ATG's Dynamo Scenario Server: Scenario-Based e-Commerce, CRM, and Merchandising. White Paper from ATG.com Enterprice Search. http://www.atg.com/repositories/ContentCatalogRepository_en/white_papers/PSG_Report.pdf
[61] Krishnapuram, R., Joshi, A., Yi, L.: A fuzzy relative of the k-medoids algorithm with application to document and snippet clustering. Proceedings of the IEEE International Conference FUZZIEEE 99, Korea, 1999.
[62] Kushmerick, N., Weld, D., Doorenbos, R.: Wrapper Induction for Information Extraction. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), 729-737, 1997.
[63] Kushmerick, N.: Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence Journal, vol. 118, 1-2, 15-68, 2000.
[64] Larkey, L., Croft, W.: Combining classifiers in text categorization. Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, 289-297, 1996.
[65] Li, W.: Question Classification Using Language Modeling. Center for Intelligent Information Retrieval Technical Report IR-259. University of Massachusetts, Amherst, 2002. http://ciir.cs.umass.edu/pubfiles/ir-259.pdf

[66] Liang, J.Z.: SVM multi-classifier and Web document classification. Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 3, 1347-1351, 26-29 August, 2004.
[67] Marbán, O., Menasalvas, E.: Estudio de perfiles de visitantes de un website a partir de los logs de los servidores web aplicando técnicas de Data mining (Webmining). Departamento de Lenguajes, Sistemas e Ingeniería del Software, Universidad Politécnica de Madrid, 2002. http://is.ls.fi.upm.es/doctorado/Trabajos20012002/OMarban.doc
[68] Martín-Bautista, M.J, Villa, M.A.: A Survey of Genetic Feature Selection in Mining Issues. Proceedings of the Congress on Evolutionary Computation – CEC99, IEEE Press, 6-9 July 1999, vol. 2, 1314-1321, 1999.
[69] McGarry, K., Martin, A., Addison, D. and MacIntyre, J.: Data Mining and User Profiling for an E-Commerce System. FSDK'02, Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, November 18-22, 2002, Orchid Country Club, Singapore, 682-686, ISBN 981-04-7520-9.
[70] Miller, B., Riedl, J., and Konstan, J.: Experiences with GroupLens: Making Usenet useful again. Proceedings of the 1997 Usenix Winter Technical Conference, January 1997.
[71] Minaei-Bidgoli, B., Punch, W.F.: Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. GECCO'2003 Genetic and Evolutionary Computation Conference, Springer-Verlag, 2252-2263, Chicago, IL, July, 2003.
[72] Mizzaro, S., Tasso, C.: Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web. Adaptive Hypermedia and Adaptive Web-Based Systems, Second International Conference AH2002, Springer, 306-316, 2002.
[73] Mladenic, D.: Text-learning and related intelligent agents: a survey. IEEE Intelligent Systems and Their Applications, vol. 14, 4, 44-54, July-August 1999.
[74] Mobasher, B., Jain, N., Han, E.H., Srivastava, J.: Web mining: Patterns from WWW Transactions. Department of Computer Science & Engineering, University of Minnesota, Technical Report, TR96-050, 1996.
[75] Mor, E., Minguillón, J.: An empirical evaluation of classifier combination schemes for predicting user navigational behavior. Proceedings of the International Conference on Information Technology: Computers and Communications, ITCC'03, 467- 471, 2003.
[76] Muslea, I., Minton, S., Knoblock, C.: Hierachical wrapper induction for semistructured information sources. Autonomous Agents and Multi-Agent Systems Journal, vol. 4, 1-2, 93-114, 2001.
[77] Nauck, D.: Using symbolic data in neuro-fuzy classification. Proceedings of the 18th meeting of North American Fuzzy Information Processing Society (NAFIPS'99), New York, 536-540, June 1999.
[78] Olsina, L.A. Ingeniería de Software en la Web. Metodología Cuantitativa para la Evaluación y Comparación de la Calidad de Sitios Web. Tesis Doctoral. Facultad de Ciencias Exactas de la Universidad Nacional de La Plata (UNLP), Argentina, 1999.
[79] Ortega, J.: Exploiting multiple existing models and learning algorithms. Working Notes of the AAAI Workshop on Integrating Multiple Learned Models, 101-106, 1996.
[80] Pal, S.K., Talwar, V., Mitra, P.: Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions. IEEE Transactions on Neural Networks, vol. 13, 5, 1163-1177, September 2002.
[81] Park, S.B., Zhang, B.T.: Automatic Webpage Classification Enhanced by Unlabeled Data. Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, vol. 2690, 821-825, 2003.
[82] Pavón, J.: Personalización de servicios en la Web. Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, 2001. http://grasia.fdi.ucm.es/SP/cursos/personalizacion.pdf

[83] Pearson, K.: Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society of London, 187, 253-318, 1896.
[84] Perkowitz, M., Etzioni, O.: Adaptive Web Sites: Automatically Synthesizing Web Pages. Proceedings of AAAI98, 727-732, 1998.
[85] Pitkow, J.: In Search of Reliable Usage Data on the WWW. Computer Networks and ISDN Systems, vol. 29, 8-13, 1343-1355, 1997.
[86] Powell, T. A.: Web Site Engineering: Beyond Web Page Design. Prentice-Hall.
[87] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Reidl, J.: Grouplens: An open architecture for collaborative filtering of netnews. Proceedings of ACM 1994 Conference on CSCW, 175-186, 1994.
[88] Rocchio, J.J.: Relevance feedback in information retrieval. The SMART Retrieval System Experiments in Automatic Document Processing, Prentice Hall, 313-323, 1971.
[89] Romero, C., Ventura, S., De Bra, P.: Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors. User Modeling and User-Adapted Interaction, vol. 14, 5, 425-464, 2004.
[90] Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. IEEE Transactions on Circuits and Video Technology, vol. 8, 5, 644-655, September, 1998.
[91] Saleeb, H.: Information Retrieval: A Framework for Recommending Text-based Classification Algorithms. Submitted in Partial Fulfillment of the Requirements for the degree of Doctor of Professional Studies In Computing At School of Computer Science and Information Systems, Pace University, June 2002
www.pace.edu/library/pages/pdf/saleeb_thesis.pdf
[92] Schafer, J.B., Konstan, J.A., Riedl, J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, The Netherlands, 5, 115-153, 2001.
[93] Schechter, S.E., Krishnan, M., Smith M.D.: Using Path Profiles to Predict HTTP Requests. Computer Networks, vol. 30, 457-467, 1998.
[94] Segrera, S., Moreno, M.N.: Multiclasificadores: Métodos y Arquitecturas. Informe Técnico – Technical Report, DPTOIA-IT-2006-001, Departamento de Informática y Automática, Universidad de Salamanca, Mayo 2006.
[95] Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating word of mouth. Proceedings of the Annual ACM SIGCHI on Human Factors in Computing Systems (CHI'95), 210-217, 1995.
[96] Spiliopoulou, M., Faulstich, L.C., Wilkler, K.: A data miner analyzing the navigational behaviour of web users. Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI99, Greece, 588-599, July 1999.
[97] Spiliopoulou, M., Pohle, C., Faulstich, L.: Improving the effectiveness of a web site with web usage mining. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD99), 142-162, 1999.
[98] Sycara, K., Pannu, A., Willamson, M., Dajun Zeng: Distributed intelligent agents. IEEE Expert Magazine, vol. 11, 6, 36-46, December 1996.
[99] The Common Logfile Format. 1995.
http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format
[100] Thorsten, J.: Text categorization with Support Vector Machines: Learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1398, Springer Verlag, Heidelberg, DE, 137-142, 1998.
[101] Ting, K.M., Witten, I.H.: Stacked Generalizations: When Does It work? Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, 866-873, 1997.
[102] Vel, O., Nesbitt, S.A.: A Collaborative Filtering Agent System for Dynamic Virtual Communities on the Web. Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98, 1998.

[103] White, R.W., Ruthven, I., Jose, J.M.: The use of implicit evidence for relevance feedback in web retrieval. 24th BSC-IRSG European Colloquium on IR Research (ECIR 2002), Glasgow, Scotland, United Kingdom, March 2002.
[104] Wolpert, D.H.: Stacked Generalization. Neural Networks, Pergamon Press, vol. 5, 241-259, 1992.
[105] Yahoo! Reports First Quarter 2001 Financial Results, Release April 11, 2001. http://biz.yahoo.com/bw/010411/0403.html
[106] Yager, R.: A framework for linguistic and hierarchical queries for document retrieval. Soft Computing in Information Retrieval: Techniques and Applications. F. Crestani and G. Pasi (Eds), Heidelberg: Physica-Verlag, vol. 50, 3-20, 2000.
[107] Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, vol. 1, Kluwer Academic Publishers, 69-90, 1999.
[108] Yang, Y. Ault, T., Pierce, T.: Combining multiple learning strategies for effective cross Validation. Proceedings of ICML-00, 17th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US, 1167-1182, 2000.
[109] Yan, T.W., Jacobsen, M., García-Molina, H., Dayal, U.: From user access patterns to dynamic hypertext linking. Computer Networks and ISDN Systems, vol. 28, 1007-1014, 1996.
[110] Zadeh, L.A.: Fuzzy logic, neural networks, and soft computing. Communications of the ACM, vol. 37, 3, 77-84, 1994.
[111] Zaïane, O.R.: Building a Recommender Agent for e-Learning Systems. Proceedings of the 7th International Conference on Computers in Education (ICCE 2002), 55-59, Auckland, New Zealand, December 3-6, 2002.
[112] Zhang, D., Lee, W.S.: Learning to Integrate Web Taxonomies. Journal of Web Semantics, vol. 2, 2, 131-151, 2004.
[113] Zhou, Z.H., Li, M.: Tri-training: exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering, vol. 17, 11, 1529-1541, November 2005.