


TESIS DOCTORAL

Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa

Víctor Heughes Escobar Jeria

Granada, 2007

Editor: Editorial de la Universidad de Granada
Autor: Víctor Heughes Escobar Jeria
D.L.: Gr. 2868 - 2007
ISBN: 978-84-338-4707-2


Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa

memoria que presenta

Víctor Heughes Escobar Jeria

para optar al grado de

Doctor en Informática

2007

DIRECTORES

Dra. María José Martín Bautista

Dra. María Amparo Vila Miranda

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
E INTELIGENCIA ARTIFICIAL

La memoria titulada ”Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa”, que presenta D. Víctor Heughes Escobar Jeria para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de las Doctoras Dra. María José Martín Bautista y Dra. María Amparo Vila Miranda.

Granada, 2007.

El Doctorando

D. Víctor H. Escobar Jeria

Los Directores

Dra. María José Martín Bautista

Dra. María Amparo Vila Miranda

Agradecimientos

Muchas personas me han apoyado en este largo recorrido, muchos me han ayudado a levantarme cuando me he caído y me han animado a seguir adelante pese a las dificultades.

Quiero agradecer a María José por la excelente labor desarrollada en la dirección de esta tesis. Por todos estos años de paciencia y apoyo que sin duda me han servido para crecer profesionalmente, y que sin su ayuda no hubiera sido capaz de realizar. Quisiera extender mi agradecimiento a Amparo por su ayuda y apoyo constante en este trabajo.

También me gustaría agradecer a mis amigos y compañeros, en especial a Verónica, Mariola, Óscar, Jesús, Javi y Carlos, que me apoyaron y me dieron su amistad durante este largo período.

Agradecer a mi Familia por su constante interés, apoyo moral y amor que me entregaron en la distancia durante este largo camino; y a todos mis amigos que siempre me entregaban ánimo y me entregaron grandes tardes de alegría.

Quisiera dedicar este trabajo en especial a Patricia Rojas Labarca que es la persona más importante de mi vida, es una mujer maravillosa, la cuál me ha dado su apoyo incondicional y ha tenido una gran paciencia durante todo este tiempo que hemos estado separados. Gracias mi amor por todo lo que me has entregado durante esta dura travesía.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La información hoy en día es una materia prima muy valiosa, tanto para empresas u organizaciones como para simples usuarios, ya que para todos es importante obtener información de buena calidad y oportuna. Es por esto que en la sociedad de la información se destina gran cantidad de recursos en adquirir, almacenar y procesar enormes volúmenes de datos; se estima que cada 20 meses se duplica la información en todo el mundo.

Existen factores importantes que han llevado a este aumento de la información como la acumulación rápida de datos, el desarrollo de sistemas gestores de base de datos más poderosos y el constante desarrollo tecnológico donde Internet y las bases de datos dinámicas, entre otras, pasan a ser las principales fuentes de extracción de información.

Aunque la especie humana posee habilidades extremadamente sofisticadas para detectar patrones y descubrir tendencias, es evidente que el Hombre no puede realizar con la misma eficiencia la tarea de analizar los trillones de datos almacenados electrónicamente al monitorear las transacciones comerciales de una o varias bases de datos. Por ejemplo, los satélites de observación de la Tierra generan del orden de un petabyte de datos (1015 bytes) diariamente; otros sistemas menos sofisticados, como centros de información de turismo, las transacciones realizadas en un supermercado, operaciones de tarjetas de créditos, etc. también son susceptibles de generar un volumen de datos imposible de analizar de forma manual o sólo utilizando técnicas estadísticas tradicionales, lo cual resulta lento, costoso y altamente subjetivo.

La explosión en el número de fuentes de información disponibles en Internet ofrece una nueva oportunidad de búsqueda y extracción de información útil a partir de bases de datos dinámicas y crecientes. Sin embargo, aunque aparentemente parezca que se presenta

un problema que es la ”sobre-información”, en el sentido de recibir más información de la que podamos asimilar, en la realidad, el fenómeno puede ser lo contrario, que padezcamos una crónica falta de información. Pondremos un ejemplo sencillo para explicar lo que esta sucediendo: el Diabético. Como se sabe, la diabetes es una enfermedad que se caracteriza por impedir la absorción de glucosa por las células para su transformación en energía; sin embargo hay suficiente en la sangre, solo que falta el elemento que hace posible la transferencia de la glucosa a la célula sedienta. De la misma manera, con la información, existe mucha y sólo faltan aquellos elementos de transformación para poder conseguir de ella el conocimiento que se necesita.

Pero, ¿qué entendemos por *conocimiento*?. Si queremos dar la definición de conocimiento, primero debemos referirnos a los diversos elementos con los que está relacionado, tales como los *datos*, la *información*, la *inteligencia* y el *aprendizaje*. Los *datos* son la representación de elementos abstraídos más o menos aislados de la realidad a partir de los modelos mentales de un individuo o conjunto de ellos. Es decir, los *datos* no aportan por sí solos ninguna explicación sobre los sucesos que describen, por lo tanto, tienen que ser interpretados por las personas para tener significado y poder ser útiles. La *información* se genera a partir de datos seleccionados, organizados y procesados de acuerdo a criterios pre-establecidos. Los métodos básicos para convertir datos en información y, al mismo tiempo, dotarlos de significado son la categorización, la contextualización, el análisis y la síntesis [Mol02]. La *inteligencia* se puede definir como la capacidad de plantear y resolver problemas de forma rutinaria. El *aprendizaje* es el proceso mediante el cuál se adquiere conocimiento.

Entonces, ¿cómo podemos encontrar la información que necesitamos?, más aún, ¿cómo podemos adquirir conocimiento de tanta información existente?. Para poder responder a estas preguntas, el primer punto de partida es definir qué es el conocimiento y así luego poder aplicar procesos para adquirir dicho conocimiento [Mol02].

Por lo tanto podemos definir el *conocimiento* como *el resultado de la aplicación de la inteligencia sobre la información en un contexto y con un propósito determinado*. Y el conocimiento en una base de datos es *el conjunto de los patrones o modelos encontrados que atienden a determinadas metas*.

Para poder extraer el conocimiento se pueden utilizar técnicas de Minería de Datos y también otras técnicas relacionadas con otras áreas del conocimiento como son la Minería de Texto, Minería Stream y Minería Web, las cuales se enmarcan dentro del Proceso de Extracción de Conocimiento (en inglés *Knowledge Discovery in Database (KDD)*) [FPSSU96].

La Minería de Datos es un área multidisciplinaria, donde se pueden utilizar diferentes técnicas fuertemente ligadas al aprendizaje automático, tales como redes neuronales, técnicas de agrupamiento, algoritmos genéticos, árboles de decisión, entre muchas otras. Dentro de este proceso se emplean algoritmos de aprendizaje clásicos, métodos estadísticos o técnicas avanzadas de bases de datos [FPSSU96], [Dan02], [CMS99], [VDS00].

La aplicación de la Minería de Datos ha diversos tipos de información generalmente provenientes de bases de datos textuales o documentales y de internet ha propiciado la aparición de nuevas áreas de estudio específicas para la explotación de este tipo de datos.

Una de estas áreas es la Minería de Texto. La Minería de Texto se centra en el descubrimiento de patrones interesantes y nuevo conocimiento en un conjunto de textos, es decir, su objetivo es descubrir cosas tales como tendencias, desviaciones y asociaciones entre “grandes” cantidades de información textual. Como la información que se analiza carece de estructura, es necesario aplicar técnicas adecuadas para la obtención de una estructura que permita el análisis de la información, es por ello que la etapa de preprocesamiento juega un rol de suma importancia en este proceso [XKPS02], [Jus04], [Tan99].

Otra de estas áreas es la Minería Web. Podemos decir de forma general que la Minería Web trata de descubrir patrones interesantes en el contenido, en la estructura y en la utilización de sitios Web. Dentro de la Minería Web podemos destacar tres áreas de estudio importantes, las cuales son la Minería Web de Contenido, la Minería Web de Estructura y la Minería Web de Uso. Todas estas áreas serán analizadas a lo largo de este trabajo, siendo esta última área, la Minería Web de Uso, nuestro principal objetivo de estudio [Etz96], [KB00],[NFJK99].

En general podemos decir que la Minería Web de Uso consiste en la aplicación de técnicas de minería para descubrir patrones de uso de la información Web con el objetivo de entender y satisfacer las necesidades de los usuarios. La principal fuente de información en este caso son los archivos Web log que se encuentran almacenados en los servidores Web.

La Minería de Datos ha utilizado durante la última década técnicas procedentes de la Computación Flexible (Soft Computing), tales como la *Lógica Difusa*. La lógica difusa cumple un rol importante en el desarrollo de sistemas inteligentes, permitiendo mayor flexibilidad y el manejo de impresiones subjetivas que pueden estar presentes en la información que se esté procesando [Zad75], [MPM02], [AM04].

En el campo del KDD, la lógica difusa puede ayudar en los problemas tales como la manipulación de asuntos relacionados con la comprensibilidad de patrones, datos ruidosos

e incompletos y la interacción humana. Concretamente, nosotros nos centraremos en las aplicaciones de lógica difusa a la Minería Web, que nos permitan mejorar y optimizar diferentes procesos de la Web.

Como en la Minería Web de Uso existen diferentes elementos que tienen cierto grado de imprecisión o de incertidumbre, la lógica difusa nos permitirá manipular esos elementos para poder representar de mejor forma la realidad. Un ejemplo significativo son los perfiles de los usuario. El tratamiento de los perfiles de usuario se basa principalmente en analizar la información que es registrada de diversas fuentes de información y se busca obtener información sobre las preferencias y características de los usuarios, muchas de estas preferencias o intereses tienen cierto grado de incertidumbre, por lo que la lógica difusa nos va a permitir modelar y manejar dicha información de una manera flexible.

1.1. Objetivos

Como se ha comentado anteriormente los objetivos de este trabajo incluyen el estudiar una serie de problemas que aparecen en la Minería Web de Uso, utilizando especialmente la lógica difusa como herramienta. Nos centraremos en el análisis del comportamiento de los usuarios en la Web, para poder obtener una representación de perfiles de usuarios.Trataremos de comprobar el comportamiento de estas técnicas y la validez de nuestros modelos mediante una experimentación sobre un caso real Para alcanzar estos objetivos, las tareas a desarrollar son las siguientes:

- Estudiar el potencial de las distintas técnicas que más se utilizan en la Minería Web centrándonos principalmente en el uso de la lógica difusa, las reglas de asociación difusas y el clustering [1] difuso.

- Realizar un estudio para la obtención de patrones de navegación, y utilizar la técnica de reglas de asociación difusas para este fin.

[1] A lo largo de este trabajo, cuando estemos hablando de grupos de elementos utilizaremos la palabra cluster/s en vez de la palabra grupo/s, y cuando estemos hablando de agrupamiento de tales elementos utilizaremos el término ya conocido por todos, que es el clustering. El propósito de esta terminología es evitar la confusión y la ambigüedad que suponen las palabras grupo/agrupamiento, ya que por el amplio significado de las mismas, no implican que se hayan obtenido de un proceso de clustering propiamente dicho. Sin embargo, el uso de las palabras cluster/clustering sí determina claramente la obtención de los grupos, lo cuál nos permite además utilizar la palabra grupos para denominaciones mucho más generales no relacionadas con la técnica de clustering.

- Realizar un estudio demográfico utilizando la técnica del clustering difuso para la agrupación y caracterización de sesiones de usuarios.

- Realizar un estudio sobre la personalización web, principalmente todo lo relacionado con los perfiles de los usuarios, para plantear un modelo general de obtención y representación de los mismos.

1.2. Estructura de la tesis

Tras el presente capítulo introductorio, pasaremos a un **Segundo Capítulo** en el cuál describiremos brevemente la problemática del crecimiento de la información, para poder recoger los principales conceptos del Proceso de Extracción del Conocimiento (KDD) en cada una de sus etapas. Dentro de este área de investigación, vamos a referirnos en primer lugar a la Minería de Datos, donde veremos principalmente las técnicas más utilizadas dentro de este proceso. También hablaremos brevemente sobre un área de investigación que hoy en día está en pleno crecimiento denominada Minería Stream. A continuación, veremos los conceptos relacionados con la Minería de Texto donde describiremos sus etapas y las técnicas utilizadas más habituales. Por último en este capítulo, analizaremos el área de la Minería Web donde también analizaremos las etapas y técnicas utilizadas, así como los principales tipos de Minería Web.

En el **Tercer Capítulo** nos detendremos para analizar con una mayor profundidad la Minería Web de Uso, identificaremos y describiremos sus etapas y comentaremos algunos problemas asociados a esta área de investigación. También hablaremos sobre los ficheros Log de los servidores Web, siendo estos una de las fuentes principales de información para el proceso de Minería Web de Uso y de algunas aplicaciones desarrolladas. Plantearemos nuestro modelo de datos, con el cuál trabajaremos a lo largo de este estudio y describiremos el proceso de caracterización de usuarios, donde estudiaremos diferentes métodos que se utilizan en la identificación de las sesiones de los usuarios.

En el **Cuarto Capítulo** nos referiremos brevemente a la problemática que existe en el manejo de la información imprecisa en la Web y cómo la lógica difusa puede ayudar a solucionarlas. Comentaremos algunas técnicas que se relacionan con los tipos de la Minería Web, especialmente en el área de la Minería Web de Uso. Explicaremos las reglas de asociación difusas y las aplicaremos para la obtención de patrones de navegación. Una vez interpretadas las reglas y contrastadas con la ayuda de medidas de interés objetivas y subjetivas, discutiremos los resultados obtenidos de los experimentos con un caso real,

concretamente en el sitio web de la E.T.S. Ingenierías Informática y de Telecomunicación de la Universidad de Granada (http://etsiit.ugr.es).

En el **Quinto Capítulo** analizaremos la técnica de minería llamada clustering, revisando los modelos generales del clustering para luego profundizar en esta técnica desde un punto de vista difuso. Además, completaremos este modelo con un análisis previo sobre la obtención de las particiones iniciales de los datos, así como una validación posterior de la técnica con diversas medidas. Los agrupamientos a realizar irán principalmente enfocados a páginas similares y a sesiones de usuarios; los experimentos de dichos análisis se han llevado a cabo en el caso real anteriormente comentado.

En el **Sexto Capítulo** analizaremos el proceso de personalización, centrándonos en la construcción de perfiles de usuario. A partir de la formalización matemática de la definición de perfiles de usuario, plantearemos un modelo para la representación de perfiles en XML. En cuanto a la metodología de obtención de los perfiles, planteamos un modelo basado en grupos previos obtenidos de la fase de minería desarrollada en capítulos anteriores. La identificación de perfiles de usuarios con grupos demográficos y la caracterización de perfiles de usuarios a través de determinadas páginas de navegación queda demostrada con experimentos reales sobre el sitio web de la E.T.S.I.I.T.

Finalmente, en el **Séptimo Capítulo** se encuentran las conclusiones sobre el trabajo y los futuros estudios relacionados con la investigación.

Para dar un mejor entendimiento a este trabajo, hemos incluido cinco apéndices, siendo *A* y *B* los apéndices relacionados con conceptos preliminares necesarios para el entendimiento y para los no familiarizados con la lógica difusa y las reglas de asociación, respectivamente. En el *apéndice C* podremos ver los resultados obtenidos en el análisis de perfiles de usuarios para el caso real del sitio web de la E.T.S.I.I.T. recogidos en el capítulo sexto. En el *apéndice D*, podemos ver una pequeña explicación del sistema WEKA, el cual utilizamos para realizar algunos experimentos de clasificación de perfiles recogidos en el capítulo sexto. Por último, para una mejor lectura de este trabajo, hemos incluido un glosario en el apéndice E donde se recoge la terminología comúnmente utilizada en la Web.

Capítulo 2

Descubrimiento del Conocimiento (KDD) : El Proceso de Minería

En este capítulo realizaremos un análisis general del Proceso de Extracción del Conocimiento (Knowledge Discovery in Databases (KDD)). Nos centraremos en los principales procesos de Minería, tales como la Minería de Datos, Minería Stream, Minería de Texto y Minería Web, siendo esta última el área principal de nuestro estudio.

2.1. Introducción

Con la necesidad de poder manejar grandes cantidades de datos, surge un área de estudio, que se denomina *descubrimiento del conocimiento en grandes volúmenes de datos* (*KDD*).

El proceso KDD lo podemos definir como ”el proceso no trivial de identificar patrones válidos, novedosos y potencialmente útiles y en última instancia, comprensible a partir de los datos” [FPSSU96]. Este proceso también es conocido por diferentes nombres que podrían ser sinónimos del mismo, entre los cuales se encuentran Data Archeology, Dependency Function Analysis, Information Recollect, Pattern Data Analysis ó Knowledge Fishing.

KDD supone la convergencia de distintas disciplinas de investigación; podemos nombrar algunas tales como el aprendizaje automático, estadística, inteligencia artificial, sistemas de gestión de base de datos, técnicas de visualización de datos, los sistemas para el

apoyo a la toma de decisión (DSS) ó la recuperación de información, entre otras.

2.2. Etapas en el proceso de KDD

Dentro del proceso KDD, uno de los elementos más importantes a considerar es el usuario, ya que es él quien determina el dominio de la aplicación o sea, decide cómo y qué datos se utilizarán en el proceso; el usuario debe entender y participar activamente en el desarrollo del mismo.

Por lo tanto, los pasos en el proceso global del KDD no están claramente diferenciados por ser un proceso iterativo e interactivo con el usuario experto. Las interacciones entre las decisiones tomadas en diferentes pasos, así como los parámetros de los métodos utilizados y la forma de representar el problema suelen ser extremadamente complejos.

Generalmente, se consideran las siguientes etapas del Proceso de Extracción del Conocimiento [FPSSU96], [Dan02]:

a. **Selección de datos.** Consiste en buscar el objetivo y las herramientas del proceso de minería, identificando los datos que han ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar la tarea. En otras palabras, lo primero que se tiene que tener en cuenta antes de comenzar con el proceso, es saber que es lo que se quiere obtener y cuales son los datos que nos facilitarán esa información para lograr la meta.

b. **Limpieza de datos.** En este paso se limpian los datos sucios, incluyendo los datos incompletos (donde hay atributos o valores de atributos perdidos), el ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes). Los datos sucios en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos. En resumen este proceso está formado por tres fases: definir y determinar los tipos de errores, buscar e identificar las instancias que contienen errores y corregir los errores descubiertos.

c. **Integración de datos.** Combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos. La inconsistencia en el formato puede llevar una redundancia e inconsistencia en los atributos y valores de los datos. Normalmente cuando se trabaja en un problema de proceso de descubrimiento es necesario primero formar un único conjunto con todos

los datos que provienen de distintas fuentes. Esta idea, en el ámbito de integración de datos de bases de datos empresariales, se conoce con el nombre de almacén de datos o ”datawarehouse” [Kim96] y proporcionan un punto único y consistente de acceso de datos corporativos sin importar la división departamental.

d. **Transformación de datos.** Las transformaciones consisten principalmente en modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para la técnica de minería aplicada. Las transformaciones discretas de los datos [HLT99] tienen la ventaja de que mejoran la comprensión de las reglas descubiertas al transformar los datos de bajo nivel en datos de alto nivel y también reduce significativamente el tiempo de ejecución del algoritmo de búsqueda. Su principal desventaja es que se puede reducir la exactitud del conocimiento descubierto, debido a que puede causar la perdida de alguna información. Existen diferentes métodos de transformación de variables continuas a discretas que se pueden agrupar según distintas aproximaciones [DKS95]: métodos locales (realizan la transformación discreta en una región del espacio de las instancias, por ejemplo, utilizando un subconjunto de las instancias), métodos globales (utilizan el espacio de las instancias), métodos supervisados (utilizan la información de la clave (valor del atributo objetivo) de la instancia de los datos cuando discretizan un atributo) y métodos no supervisados (no requiere la información de la clave para poder pasar a valores discretos. Sólo utilizan la distribución de los valores del atributo continuo como fuente de información).

e. **Reducción de datos.** Reducir el tamaño de los datos, encontrando las características más significativas para representar los datos dependiendo del objetivo del proceso. Se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas, o para encontrar otras representaciones de los datos. Las estrategias incluyen agregación del cubo de datos (Ej. sum( ) y min( )), reducción de dimensiones (la extracción irrelevante y débil de atributo), compresión de datos (reemplazando valores de datos con datos alternativos codificados), reducción de tamaño (reemplazando valores de datos con representación alternativa más pequeña), una generalización de datos (reemplazando valores de datos de niveles conceptuales bajos con niveles conceptuales más altos), etc.

f. **Minería de Datos.** Consiste en la búsqueda de los patrones de interés que pueden expresarse como un modelo o simplemente que expresen dependencia de los datos. El modelo encontrado depende de su función (por ej. Clasificación) y de su forma de representarlo (por ej. árboles de decisiones, reglas, entre otras). Se tiene que

especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos. También se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está determinado en el algoritmo de minería).

g. **Evaluación de los patrones.** Se identifican verdaderamente patrones interesantes que representan conocimiento usando diferentes técnicas incluyendo análisis estadísticos y lenguajes de consultas.

h. **Interpretación de resultados.** Consiste en entender los resultados del análisis y sus implicaciones y puede llevar a regresar a algunos de los pasos anteriores. Hay técnicas de visualización que pueden ser útiles en este paso para facilitar el entendimiento de los patrones descubiertos.

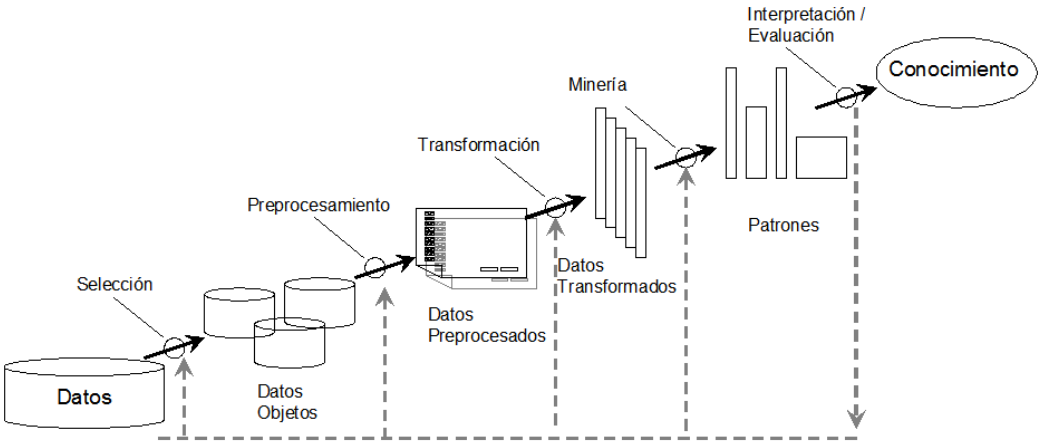

Figura 2.1: *Etapas del Proceso KDD*

Podemos ver en la figura 2.1 las etapas del proceso KDD. Para ver con más detalle éste proceso podemos encontrarlas en [KLR+98], [Pyl99], [CMS99], [FPSSU96], [Mit97].

2.2.1. KDD y minería

Es importante discutir sobre la diferencia entre KDD y Minería de Datos, ya que muchos estudios e investigaciones dan por hecho que ambos son sinónimos. Hay muchos casos que no es posible identificar o distinguir claramente la etapa de Minería de Datos

dentro del Proceso de Descubrimiento [VDS00], porque a veces no es necesario realizar todas y cada una de las etapas del mismo, como preprocesamiento, limpieza de datos, etc.

Entonces, la Minería de Datos la podemos definir como *una etapa particular en el proceso KDD, donde la Minería de Datos aplica algoritmos específicos o técnicas específicas para la extracción de patrones de los datos* [FPSSU96], diferenciándolo del proceso KDD que ya antes a sido definido.

Hasta el momento, sólo hemos mencionado dos términos importantes dentro del área de investigación de extracción de patrones útiles de conjuntos de datos, que son KDD y Minería de Datos, y hemos dejado de lado dos términos que actualmente se ocupan fuertemente en la extracción de patrones útiles desde documentos y desde la Web como son la Minería de Texto y la Minería Web.

Es importante analizar las diferencias entre los procesos de Minería de Datos, Minería de Texto y Minería Web, antes de comenzar a analizar a cada uno de ellos por separado. Hemos dejado de lado a la Minería Stream ya que es un área nueva en desarrollo y que está fuertemente relacionado con la Minería de Datos.

Si nos detenemos a pensar cuáles serían las diferencias entre estos procesos, deberíamos empezar diciendo que las fuentes de información sobre las que trabajan son diferentes. La Minería de Datos principalmente trabaja sobre grandes almacenes de datos y Bases de datos relacionales, por ejemplo; en cambio la Minería de Texto se centra en documentos y la Minería Web en todo lo que se relaciona con la World Wide Web. El tipo de información que procesan es muy distinta. La información que procesa la Minería de Datos es información estructurada, ya que los datos están almacenados en una base de datos generalmente relacional; sin embargo, la Minería de Texto, trata con complejas estructuras implícitas del texto, generalmente no estructuradas y por último, la Minería Web procesa tanto información estructurada (procesa información de bases de datos que están relacionadas con las páginas Web), semiestructurada (páginas HTML con texto y Hiperenlaces) y no estructurada (texto libre). Un aspecto a mencionar, es que en la literatura se mencione, a la Minería de Texto como KDT o a la Minería Web como KDW, para hablar sobre el proceso de descubrimiento en esas áreas, que es el mismo caso que ya hemos mencionado sobre el KDD y Minería de Datos.

Una vez comentadas brevemente las diferencias entre estos procesos vamos a entrar a especificar a cada uno de ellos, comenzaremos con la Minería de Datos; primero porque está mucho más extendido y segundo porque la gran mayoría de las técnicas de éste se aplican en los procesos de Minería Stream, Minería de Texto y Minería Web, y así tener una mayor referencia al momento de explicar estos últimos procesos.

2.3. Minería de datos

La Minería de Datos es la etapa más importante del KDD; es la que integra los procesos de aprendizaje y métodos estadísticos para la obtención de hipótesis de patrones y modelos. Es esencial que los algoritmos empleados en la Minería de Datos sean eficientes, escalables y robustos a la hora de manipular grandes cantidades de información con ruido.

Informalmente hablando, se puede decir que la Minería de Datos es el proceso de extracción de información o conocimiento de un conjunto grande de datos. Formalizando un poco más, lo podemos definir como *una etapa particular en el proceso KDD*, donde la Minería de Datos aplica algoritmos específicos o técnicas especificas para la extracción de patrones de los datos.

Debemos tener claro que el proceso de Minería de Datos genera muchos tipos de patrones, pero lo más importante es determinar qué patrones son útiles e interesantes y cuales no. Un patrón es interesante si cumple con ciertas condiciones, si es fácil comprenderlo; es válido con cierto grado de certeza, para otro conjunto de datos, ya sea nuevo o de prueba; tiene una utilidad potencial y expresa un conocimiento novedoso y no trivial.

Es importante destacar que la Minería de Datos, en forma muy general, puede procesar distintos tipos de datos, de diferentes fuentes como archivos planos (texto, binario,...), base de datos relacionales, base de datos heterogéneas, base de datos orientadas a objetos, datawarehouse, base de datos transaccionales, base de datos espaciales, base de datos multimedia y base de datos temporales, entre otras.

Un ejemplo de Minería de Datos sería el siguiente: hace más de dos temporadas que el club italiano de fútbol AC Milán usa redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta [Mth03]. Esto ayuda a seleccionar el fichaje de un posible jugador o alertar al médico del equipo de una posible lesión. El sistema, creado por Computer Associates International, es alimentado por datos de cada jugador, relacionados con su rendimiento, alimentación y respuesta a estímulos externos, que se obtienen y analizan cada quince días.

2.3.1. Técnicas de minería de datos

El objetivo de la Minería de Datos es la obtención de patrones útiles en un conjunto grande de datos, pero ¿qué tipo de patrones puede encontrar la Minería de Datos?. Dependiendo del conjunto de datos a analizar y del tipo de patrones que se quiera encontrar en

el proceso es la técnica de Minería de Datos a utilizar. Las técnicas de Minería de Datos pueden ser descriptivas o predictivas. Las descriptivas caracterizan las propiedades generales de los datos en una base de datos y por el contrario, la predictiva realiza inferencias en los datos para poder realizar predicciones. Hay que tener claro lo que se desea obtener para ver cual de los dos enfoques se utiliza.

A continuación clasificaremos las técnicas más usadas, según sean descriptivas o predictivas:[Dan02]

Técnicas Descriptivas.

- *Descripción de clases.* Hay tres formas de ver este punto, la primera se denomina caracterización de los datos (Data Caracterization), el cuál realiza un resumen de las características generales de una clase particular de datos; los resultados suelen representarse en términos de reglas de caracterización. La segunda es la discriminación de datos (Data Discrimination), que es una comparación entre las características generales de los objetos de una clase respecto a las de otro conjunto contrastante. Finalmente, también se puede aplicar una combinación de ambas.

- *Análisis de asociación.* Es el descubrimiento de reglas de asociación que muestran condiciones del tipo atributo-valor que ocurre con frecuencia dentro de un conjunto de datos. La minería mediante reglas de asociación es el proceso de búsqueda interesante de correlaciones entre un conjunto grande de datos [AGGR98], [AR94], [MTV94]. El descubrimiento de reglas de asociación en grandes volúmenes de transacciones de negocios, puede facilitar el proceso de toma de decisiones. Por ejemplo, una regla de asociación descubierta en un conjunto de transacciones de libros de computación puede ser como sigue:

  $$\text{Sistema Operativo} \longrightarrow \text{LINUX [soporte = 3 \%, confianza =45 \%]}$$

  Esta regla refleja un modelo de compra para libros de computación, donde el consumidor que compra libros de sistemas operativos, tiende a comprar libros de Linux al mismo tiempo. El soporte y la confianza son dos medidas que reflejan la utilidad y la certeza de la regla descubierta. En el ejemplo estos índices indican que el 45 % de las transacciones que contienen libros de sistemas operativos también contienen libros de Linux y que el 3 % de todas las transacciones contiene a ambos ítems.

- *Análisis de clusters.* Aquí se analizan objetos sin consultar clases conocidas. En general, las clases no se presentan en los datos de entrenamiento simplemente porque

no se conocen. El proceso trabaja agrupando objetos según el principio de ”maximizar la similitud dentro de una clase y minimizar la similitud entre clases”. Un cluster es una colección de objetos de datos mutuamente similares. Clustering es el proceso de agrupamiento de objetos [Har75], [JD98], [KR90]. El análisis de clustering, tiene una gran variedad de aplicaciones, incluyendo procesos de imágenes, análisis de transacciones comerciales y reconocimiento de patrones.

Técnicas Predictivas.

- *Clasificación y predicción.* Son dos tipos de análisis de datos, aquellos que pueden ser usados para clasificar datos y los que se usan para predecir tendencias. La clasificación de datos predice clases de etiquetas mientras la predicción de datos predice funciones de valores continuos. Aplicaciones típicas incluyen análisis de riesgo para préstamos y predicciones de crecimiento. Algunas técnicas para clasificación de datos incluyen: clasificación bayesianas. K-Nearest Neighbor, algoritmos genéticos, entre otros.

- *Árboles de decisión.* Definen un conjunto de clases, asignando a cada dato de entrada una clase y determina la probabilidad de que ese registro pertenezca a la clase. Podemos distinguir dos tipos de árboles, el primero es el árbol de decisión de clasificación, donde cada registro a clasificar fluye por una rama del árbol. La rama a seguir es determinada por una serie de preguntas definidas por los nodos de la rama. Cuando el registro llega a un nodo hoja, se le asigna a la clase del nodo hoja. El segundo es el árbol de decisión de regresión, cuando el registro llega a un nodo hoja, a la variable de salida de ese nodo, se le asigna el promedio de los valores de la variable de salida de los registros que cayeron en ese nodo hoja durante el proceso de entrenamiento.

- *Redes Neuronales.* Son modelos predictivos no lineales que aprenden a través del entrenamiento. Existen diferentes tipos de redes neuronales, las más conocidas son las simples y multicapas. Las tareas básicas de las redes neuronales son reconocer, clasificar, agrupar, asociar, almacenar patrones, aproximación de funciones, sistemas (predicción, control, entre otros) y optimización.

Aunque hemos detallado algunas de las principales técnicas que se utilizan en la Minería de Datos, existen muchas otras que se pueden ver en la tabla 2.1, agrupándolas en técnicas descriptivas y predictivas. [Jus04]

Técnicas de Minería de Datos	
Métodos Descriptivos	**Métodos Predictivos**
a. Visualización	**a. Regresión Estadísticas (interpolación y predicción)** *- Regresión Lineal* *- Regresión no lineal* *- Regresión* *-Regresión Adaptativa Lineal Ponderada Localmente*
b. Aprendizaje no supervisado *-Clustering* Métodos no jerárquicos (Partición) Métodos Jerárquicos (N-TREE) Métodos Paramétricos (Algoritmo EM) Métodos no Paramétricos (KNN, K-means Clustering, Centroides, Redes Kohonen, Algoritmo CobWeb, Algoritmo Autoclass)	**b. Aprendizaje Supervisado** *-Clasificación* Árboles de Decisión, ID3, C4.5, CART *-Inducción de Reglas* *- Redes Neuronales* (simple, multicapa) *- Aprendizaje Relacional y Recursivo* IFP (Inductive Functional Programming), IFLP (Inductive Functional Logic Programming), Aprendizaje de Orden Superior, Macro Average, Matrices de Coste y Confusión, Análisis ROC (Receiver Operating Characteristic)
c. Asociación	
d. Asociación Secuencial	
e. Análisis Estadístico	
f. Análisis Estadístico *-Estudio de la Distribución de los Datos* *-Detección de Datos Anómalos* *-Análisis de Dispersión*	
g.Correlaciones y Estudios Factoriales	

Tabla 2.1: *Técnicas de Minería de Datos*

2.4. Minería stream

En estos últimos años, las bases de datos y comunidades que se preocupan en minar datos enfocando la atención en un modelo nuevo de procesamiento de datos, donde los datos llegan en forma de streams o una secuencia continúa de datos.

Estos datos que llegan de forma continua y rápida presentan un gran desafío para la Minería de Datos tradicional, ya que es realmente desafiante realizar las operaciones que habitualmente se usan en el análisis de enormes cantidades de datos.

Este nuevo modelo de análisis es denominado Minería Stream y se puede definir como

un proceso de extracción del conocimiento de estructuras de registros rápidos y continuos de datos. Los ejemplos de datos streams incluyen tráfico de la red de computadoras, conversaciones telefónicas, transacciones ATM, búsquedas web y datos de sensores [BW01], [GMMO00], [Bar02], [CDH$^+$02].

La mayoría de los streams liberan datos en orden arbitrario, los cuales están intrínsecamente relacionados con un aspecto temporal, esto quiere decir que los patrones que son descubiertos en ellos siguen una tendencia dinámica, y por lo tanto son diferentes a los conjuntos de datos estáticos tradicionales que son muy grandes. Tales secuencias de datos se refieren a streams de datos de desarrollo y por esta razón, las técnicas que son dimensionables para conjuntos de datos enormes no pueden ser la respuesta para minar las secuencias de datos o streams de desarrollo, ya que estas técnicas siempre se esfuerzan en el trabajo de conjuntos de datos sin hacer distinción entre datos nuevos y datos viejos, y así esperar manipular la noción de patrones emergentes y obsoletos.

La investigación de Minería Stream ha sido activada en estos últimos años. Dentro de los estudios realizados en esta área podemos mencionar los trabajos realizados desde un punto de vista general, los cuales los podemos ver en [BW01], [BBD$^+$02], [Mol02],[GGR02]. Existen otros estudios relacionados con la administración de los streams y el procesamiento de búsqueda continua de los streams en [GM98].

Dos recientes progresos motivan la necesidad de los sistemas de procesamiento de streams [GO03], [Mth03]:

I. La generación automática de altas tasas de secuencias de datos en diferentes aplicaciones científicas y comerciales. Por ejemplo: El satélite, el radar, y aplicaciones científicas de las corrientes de datos astronómicas, la bolsa de valores y las transacciones web log de datos streams en las aplicaciones comerciales.

II. La necesidad para los análisis de estos datos de alta velocidad de los streams como clustering y la detección de valores atípicos, la clasificación y el cálculo de itemsets frecuentes.

Algunos algoritmos que se utilizan en el área de la Minería Stream están relacionados con proyectos de negocios y también en aplicaciones científicas. Estos algoritmos han sido desarrollados y debatidos en [BBD$^+$02], [GGR02], [Kar01]. Más adelante veremos algunos de los algoritmos y su aplicación en diferente áreas (Ver tabla 2.2).

Existen diferentes y recientes proyectos que estimulan la necesidad para las técnicas en vías de desarrollo que analizan datos streams en tiempo real. De los cuales podemos

mencionar a:

- JPL/NASA están desarrollando un proyecto llamó a Diamond Eye [BFR+99] que apunta a permitir que sistemas alejados puedan analizar objetos espaciales de imágenes streams en tiempo real. El proyecto enfoca la atención en facultar una nueva era de exploración espacial usando naves espaciales, exploradores y sensores altamente autónomos [BFR+99].

- En [Kar01] y [Kar03] podemos ver el proyecto llamado MobiMine, que es un sistema cliente/servidor sobre una PDA basada en la distribución de datos utilizando Minería de Datos para datos financieros.

- Kargupta [Kar03] ha desarrollado un Sistema Minador Stream Data Vehicle (VEDAS) que es un sistema de Minería de Datos ubicua que permite un monitoreo continuo y la extracción del patrones de datos stream generados por un vehículo de traslado.

- En [SS03] desarrollan un proyecto en la NASA para la detección abordo de procesos geofísicos como nieve, hielo y nubes usando métodos clustering para la compresión de datos conservando el ancho de banda limitado necesitado para enviar las imágenes streaming a los centros terrestres.

Estos proyectos y otros demuestran la necesidad para las técnicas de análisis de datos stream y las estrategias que pueden hacer frente a la alta tasa de datos y así dar los resultados de análisis en el tiempo real.

2.4.1. Desafíos en la minería stream

En esta sección, presentamos asuntos y desafíos que se originan en la Minería Stream y algunas soluciones que se ocupan de estos desafíos. En la figura 2.2 se muestra el modelo general de procesamiento de datos en la Minería Stream.

I. El requisito ilimitado de memoria debido al rasgo continúo de los elementos entrantes de datos.

II. Los algoritmos de Minería toman varios pasos por encima de datos stream y esto no es aplicable por el rasgo alto de tasa de datos de los stream.

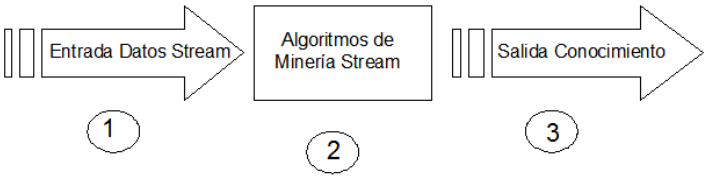

Figura 2.2: *Proceso de Minería Stream*

III. Datos stream generados de sensores y otras fuentes inalámbricas de datos crean un desafío real para transferir estas cantidades inmensas de elementos de datos para un servidor central para ser analizadas.

Hay varias estrategias que dirigen hacia estos desafíos. Éstas incluyen:

- **Los datos de entrada evalúan la adaptación:** Este acercamiento toma muestras, filtran, realiza agregación, y derramamiento de carga en los elementos entrantes de datos. El muestreo es el proceso de estadísticamente seleccionando los elementos del stream entrante que sería analizado. El filtrado es el muestreo de semántica en el cuál al elemento se le comprueba su importancia pues para ser analizado o no. La agregación es la representación de número de elementos en alguna medida estadística que usa elementos agregados, como el promedio. Mientras el derramamiento de carga, el cuál ha estado pensado en el contexto de datos stream que pone en duda [BBD+02], [TCZ+03], [VN02] en vez de extraer de la mina de datos stream, es el proceso de eliminar una cantidad de cosas de subsiguientes elementos de ser analizado comprobando cada elemento que es usado en la técnica de muestreo. En la Figura 2.3 ilustra la idea de adaptación de tasa de datos del lado de entrada usando muestreo.

- **Salida nivel de concepto:** Usando el nivel más alto de concepto de datos que se aplica minando así hacer frente a la tasa de datos, esto es para clasificar en categorías los elementos entrantes en un número limitado de categorías y reemplazando cada elemento entrante con la categoría que hace juego según una medida especificada o una tabla de búsqueda. Esto produciría menos resultados conservando la memoria limitada. Además, precisaría menos número de ciclos de la CPU.

- **Algoritmos:** Se utilizan algoritmos para aproximar los resultados minadores según algún margen de error satisfactorios.

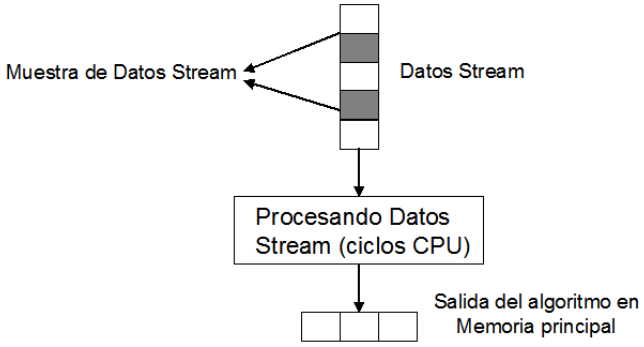

Figura 2.3: *Adaptación de tasa de datos usando muestreo*

- **Análisis:** Para evitar transferir cantidades enormes de datos, la Minería de Datos estaría hecha en la posición de la fuente de datos. Por ejemplo, VEDAS [Kar03] y Diamond Eye Proyect [BFR+99]. Esto sin embargo asume la disponibilidad de recursos computacionales significativos en el lugar de generación de datos stream.

- **Algoritmo de salida granulados:** Usa un parámetro de control como una parte del algoritmo lógico para controlar la tasa de producción del algoritmo según la memoria disponible, el tiempo restante para llenar la memoria disponible antes de la integración incremental de conocimiento tenga lugar y la tasa de datos del stream entrante. En la figura 2.4 se muestra la idea general del proceso que hemos analizado hasta ahora.

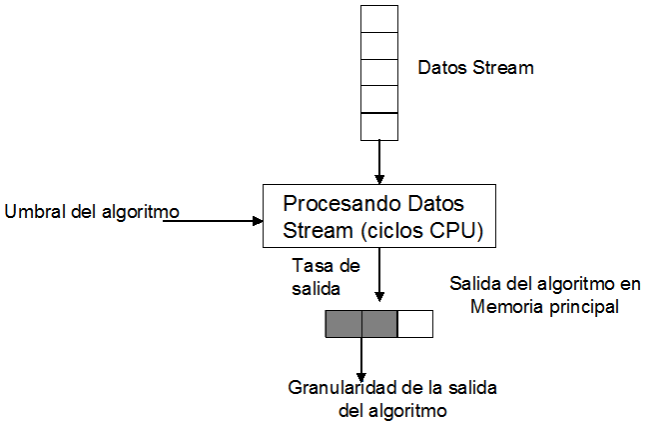

Figura 2.4: *Enfoque del algoritmo de salida granulado*

Algoritmo	Técnica	Enfoque	Estado
VFKM	K-means	Muestreo y reducción del número pasos en cada paso del algoritmo	Implementado y Testeado
VFDT	Árbol de Decisión	Muestreo y reducción del número pasos en cada paso del algoritmo	Implementado y Testeado
Conteo De Frecuencia Aproximada	Itemsets Frecuentes	Actualización y poda incremental de conjuntos de ítems por por cada bloques de transacciones	Implementado y Testeado
Clasificación conceptos sin dirección	Clasificación	Clasificación de Conjuntos	Implementado y Testeado
K-medias Aproximado	K-medias	Muestreo y reducción del número pasos en cada paso del algoritmo	Estudio Analítico
ClusStream	Clustering	Resumen online y clustering offline	Implementado y Testeado

Tabla 2.2: *Algoritmos, técnicas, enfoque y estado de implementación en la Minería Stream*

2.5. Minería de texto

Tradicionalmente la búsqueda de conocimiento se ha realizado sobre datos almacenados en bases de datos, pero la mayoría de la información dentro de una organización se encuentra en formato de texto, ya sea Intranet, páginas Web, informes de trabajo, publicaciones, correos electrónicos, entre otros. De esta manera, se abre un nuevo camino en la extracción de conocimiento de los documentos; de esta gran necesidad surge la técnica o proceso llamado Minería de Texto. La Minería de Texto difiere de Minería de Datos en el trato de la información, donde la información textual difiere de la estructurada principalmente en la ausencia de estructura o en la compleja estructura implícita del texto. De este modo, se hace necesario buscar alguna representación intermedia del texto que pueda ayudar a la aplicación de técnicas de descubrimiento, que nos permitan extraer patrones útiles.

La Minería de Texto implica a diversas áreas tales como la recuperación de información, extracción de información, tecnologías de bases de datos, aprendizaje de bases de datos, por nombrar algunas. Además, se presentan diversos problemas para los cuales se necesita aplicar el método más conveniente. Por ejemplo la naturaleza heterogénea y distribuida de los documentos, la diversidad de idiomas en los que se pueden presentar el texto, la ausencia de estructura de texto, que los hace difícil de tratar computacionalmente, la dependencia contextual del texto,...

Como en el KDD y la Minería de Datos, no existe una definición muy clara de este proceso, y existen algunas lagunas relacionadas con lo que debe o no hacer en la Minería de Texto, quizás por su corta vida; así, diversos autores de la Minería de Texto dan diferentes definiciones, por citar algunas: ”Proceso de extraer patrones interesantes a partir de grandes colecciones da textos para descubrir conocimiento”[Tan99], [DSVM03]; ”Descubrimiento de reglas de asociación importantes dentro del corpus del texto”[WCF+00]; ”Instancia del descubrimiento óptimo de patrones”[Fa00]; ”Descubrimiento de información útil y previamente desconocida a partir de texto sin estructurar”[XKPS02]; entre otras. Como se ve, algunas definiciones se basan meramente en la técnica que ocupan o simplemente en el descubrimiento de patrones, pero la Minería de Texto va más allá, ya que el conocimiento puede ser representado de muchas formas.

Podemos definir más claramente la Minería de Texto como: *”El proceso que descubre información útil que no está presente explícitamente en ninguno de los documentos objeto de análisis y que surge cuando se estudian adecuadamente y se relacionan dichos documentos”*[XKPS02].

Al contrario del acceso de información o recuperación, que ayuda a los usuarios a encontrar documentos que satisfacen sus necesidades de información, la meta de la Minería de Texto es el descubrimiento, reconocimiento o la derivación de información nueva de grandes colecciones de texto [BR99], [Cro95].

Un ejemplo de aplicación lo podemos ver en un trabajo publicado sobre la permeabilidad de las disciplinas científicas, [Swa94]. Donde Swanson ha demostrado cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica. Y de esta manera logró inferir que la deficiencia de magnecio podría representar un papel en algunos tipos de migraña, y que en estudios posteriores han probado experimentalmente esta hipótesis obtenida por la Minería de Texto.

2.5.1. Etapas de la minería de texto

Para poder descubrir conocimiento en texto, se debe pasar por algunas etapas importantes en este proceso, como es la etapa del *preprocesamiento* que le da al texto una Forma Intermedia (FI) que permita ser tratada computacionalmente, luego aplicar alguna técnica de Minería de Texto y finalmente la visualización de los resultados. (Como se ilustra en

la figura 2.5).

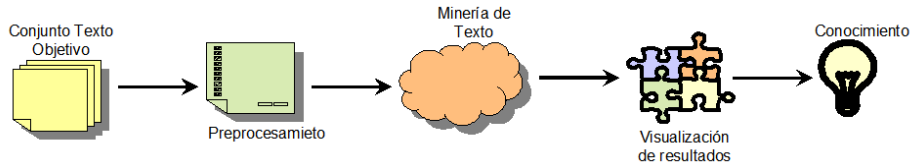

Figura 2.5: *Etapas de la Minería de Texto*

Por lo tanto, podemos considerar las siguientes etapas del proceso de Minería de Texto:

- **Preprocesamiento.** Ya sabemos que el texto no presenta una estructura fácil para aplicar técnicas de Minería de Texto directamente, así que de alguna manera se realizarán operaciones o transformaciones sobre el texto, en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis [Tan99]. Esta fase se realiza sobre un conjunto de documentos objetos de estudio, algunos autores como Tan en [Tan99] la llaman Text Refining. Este es un paso muy importante ya que, dependiendo del tipo de método usado en esta etapa de preprocesamiento, es el tipo de representación del contenido de los textos que ha sido construida y dependiendo de esta representación, es el tipo de patrones que se descubren. Las estructuras descubiertas con el procesamiento, unida a la información semántica obtenida de las bases de conocimiento, proporcionan la base para la aplicación de las técnicas de Minería, es decir, nos permiten obtener las diferentes representaciones o Formas Intermedias de documentos.

  Los tipos de Formas Intermedias en los que podemos representar una colección de documentos pueden ir desde la simplicidad de la palabra hasta la complejidad del documento completo. Si tomamos la definición de Ah-Hwee Tan en [Tan99], la Forma Intermedia se puede clasificar, en general, como:

  - *Estructurada.* Donde los datos se representan de forma relacional
  - *Semiestructurada.* Representación de un grafo conceptual
    - *Basada en conceptos.* Donde cada entidad representa un objeto o concepto de interés de un dominio especifico. Deriva patrones y relaciones a través de objetos de conceptos. Se pueden aplicar operaciones de Minería de Datos como el modelado predictivo y el descubrimiento asociativo.

- *Basados en documentos.* Cada entidad representa un documento. Deduce patrones y relaciones de interés en un dominio específico. La FI basada en documentos se puede transformar en una FI basada en conceptos, extrayendo información relevante de acuerdo a los objetos de interés de un dominio especifico.

Hay que destacar la importancia de encontrar dicha Forma Intermedia, ya que en función de la representación elegida puede cambiar el descubrimiento obtenido (Ver tabla 2.3). La elección de dicha Forma Intermedia, en principio, es independiente de las técnicas de Minería de Datos que se vayan a emplear.

Preprocesamiento	**Tipo de Representación**	**Tipo de Descubrimiento**
Categorización	Vector de Temas	Nivel Temático o relación entre temas
Análisis del Texto completo	Secuencia de palabras	Patrones de lenguaje
Extracción de Información	Tabla de Base de datos	Relaciones entre entidades

Tabla 2.3: *Relación entre Preprocesamiento, Tipo de Representación y Tipo de Descubrimiento*

Algunas de las técnicas utilizadas para la transformación de documentos en una forma intermedia pueden ser: análisis de texto, categorización, técnicas de procesamiento de lenguaje natural (etiquetado de parte del discurso, tokenización, lematización), técnicas de extracción de información (categorización, adquisición de patrones léxico sintáctico, extracción automática de términos, localización de trozos específicos de texto), técnicas de recuperación de información (indexación).

- **Minería de texto.** Fase de descubrimiento donde las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevo conocimiento. Aquí se emplean técnicas de Minería de Texto como la categorización y clasificación de textos, descubrimiento de asociaciones, detección de desviaciones, análisis de tendencias, entre otras.

- **Visualización de los resultados.** En esta fase se proporciona un ambiente de exploración de los datos guiado para el usuario que sea lo más amigable posible. Las últimas tendencias presentan los resultados mediante gráficas o páginas Web. Una vez obtenidos los conceptos, los términos o las tendencias, se pueden utilizar métodos automáticos de visualización o bien pueden interpretarse los resultados directamente.

Una vez que hemos visto el proceso de Minería de Texto, nos centraremos ahora de forma general en algunas técnicas que se utilizan en este área.

2.5.2. Técnicas de la minería de texto

Podemos clasificar las técnicas de la Minería de Texto de la misma forma que se hace en la Minería de Datos en descriptivas y predictivas, ya que cumple los mismos objetivos de encontrar patrones útiles para descubrir conocimiento, y sólo se diferencian en la fuente de información donde trabajan. La tabla 2.4 muestra una esquematización de las técnicas que se suelen aplicar para realizar Minería de Texto.

Técnicas de la Minería de Texto	
Métodos Descriptivos	**Métodos Predictivos**
a.Visualización de Documentos	**b.Aprendizaje Supervisado** *1.Clasifcicación* - Árboles de Decisión - Inducción de Reglas - Redes Neuronales - Clasificación Naive Bayes - Modelado Predictivo - Aprendizaje Relacional Recursivo *2.Categorización*
b.Aprendizaje No Supervisado *1.Clustering* *2.Clustering Conceptual*	
c.Reglas de asociación	
d.Asociación Secuencial	
e.Análisis Estadístico	
f.Aprendizaje de Patrones	
g.Soft Matching	

Tabla 2.4: *Técnicas de Minería de Texto*

A continuación explicaremos algunas técnicas de la más relevante de Minería de Texto.

Técnicas descriptivas.

- **Clustering de documentos.** Puede definirse como la tarea de separar documentos en grupos. El criterio de agrupamiento se basa en las similitudes existentes entre

ellos [KR90]. Sus aplicaciones más importantes son mejorar el rendimiento de los motores de búsqueda de información mediante la categorización previa de todos los documentos disponibles, facilitar la revisión de resultados por parte del usuario final, agrupando los resultados, tras realizar la búsqueda [Van79], [Cro95].

- **Clustering conceptual.** Consiste en encontrar todas las regularidades de un conjunto de grafos conceptuales en una jerarquía para facilitar la navegación a través del grafo [GMLL02]. Algunas formas más habituales de usarlos incluyen métodos tales como c-means y métodos no tradicionales basados en redes neuronales del tipo Kohonen (que es un modelo de Red Neuronal que simplemente utiliza la información derivada del termino anterior y posterior o para la clasificación de palabras claves de una base de datos o permite hacer Clustering como una clasificación topológica).

- **Detección de desviaciones.** Se detectan desviaciones en un conjunto de grafos conceptuales que tienen que ver con una característica del conjunto C de grafos conceptuales, el cual es una generalización g de más de m grafos conceptuales, donde m es un valor definido por el usuario; un grafo conceptual raro, es aquel que no tiene características representativas; una desviación d es un patrón que describe una o más de los grafos raros. Luego, una desviación conceptual es una expresión de la forma *g:d(r,s)*, donde g es el contexto, d es la descripción de los grafos raros del contexto, r es la rareza de la desviación en el contexto y s es el soporte del contexto con respecto al conjunto completo. [GMLL02]

Técnicas predictivas.

- **Clasificación de términos.** Técnicas que detectan clases en los datos de acuerdo con observaciones. Algoritmo de inducción de reglas, en donde aprender un concepto significa inferir su definición general a partir de un número de ejemplos específicos. Estos algoritmos se caracterizan por representar la categoría de prueba como un conjunto de reglas ”if-then”[PB02].

- **Reglas de asociación.** Se encuentran asociaciones entre conceptos que se expresan de la forma $A \longrightarrow B$ [soporte, confianza], donde A y B pueden ser uno o varios conceptos. El algoritmo Apriori [AR94] es una aproximación usada frecuentemente para encontrar reglas de asociación. Tiene dos pasos: encontrar todos los conjuntos de ítems frecuentes y generar reglas de asociación fuertes [PB02].

Hemos descrito el Proceso de descubrimiento de conocimiento en textos, y las técnicas más utilizadas, ahora analizaremos proceso de descubrimiento de conocimiento en la Web o Minería Web.

2.6. Minería web

Hoy en día, Internet juega un papel muy importante en la difusión de la información. Herramientas tecnológicas como el correo electrónico o el protocolo de transferencia de archivo (FTP), el comercio electrónico o simplemente leer el periódico en la Web han significado un cambio social muy importante, ya que pasan a ser indispensables en nuestro vivir diario.

El estudio de la World Wide Web se ha convertido en uno de los campos de investigación más interesantes y como comenta Kleinberg [KKR99], pocos eventos de la historia de la computación ha tenido tanta inferencia en la sociedad como la llegada y crecimiento de la World Wide Web

Respecto al tamaño de la Web es necesario además dar algunos datos referentes a su crecimiento, para tener una visión más amplia del futuro y de la razón que hace urgente el organizar esta información y optimizar su acceso y tratamiento. La Web consta de más de 72 millones de sitios actualmente, lo que sería aproximadamente 1000 millones de páginas; además, crece exponencialmente ya que se crean cerca de 1.5 millones de páginas diariamente.

Debido a esta situación, existen muchos desafíos con respecto a la obtención de información de la Web, ya sea por tener una inmensa cantidad de datos, por la diversidad de lenguaje que se presenta, datos redundantes o no estructurados, la calidad de la información, datos distribuidos en diferentes plataformas, por referirnos a algunos.

Al igual que en la Minería de Texto existe una gran heterogeneidad y falta de estructura en la información a explotar, en la World Wide Web existe una diversidad de fuentes de información muy variada, las cuales podemos clasificar de la siguiente manera:

- *Fuentes no estructuradas.* Son aquellas que no representan ningún esquema para la información que contiene. Un ejemplo típico son documentos en texto libre de cualquier tipo (documentos word, ficheros pdf, la gran mayoría de las páginas Web estáticas, etc.). Las herramientas que soportan este tipo de información sólo permiten la realización de búsquedas por palabra clave o por concepto, que ordenados

según algún indicador de relevancia, son directamente utilizados para presentar sus resultados de forma más o menos completa y relacionada al usuario final.

- *Fuentes estructuradas.* Son aquellas que presentan un esquema rígido bien definido, diferenciado de los datos. Un ejemplo típico es una base de datos relacional que presenta un esquema, almacenado en el diccionario de datos, que define la organización interna de la información. Como es conocido, el acceso a este tipo de fuente se realiza mediante potentes lenguajes de consulta (Ej. SQL).
- *Fuentes semiestructuradas.* Son aquellas fuentes que no presentan un esquema rígido [Abi97]. Esto quiere decir que el esquema es implícito y está contenido en (o puede deducirse de) los propios datos (self-describing). O que aun habiendo un esquema, éste es muy vago y permite cierta flexibilidad (sobre los tipos de datos, las restricciones,...). Por ejemplo son fuentes semiestructuradas ficheros XML, Excel, los logs emitidos por muchos sistemas o infinidad de documentos con información tabulada.

Además en la Web los usuarios se encuentran con diferentes situaciones o problemas, los cuales podemos mencionar:

I. *La información relevante encontrada.* Cuando los usuarios usan el servicio de búsqueda, usualmente ingresan palabras claves de búsqueda y obtienen su respuesta basándose en una lista de preguntas o consultas similares ya hechas. Hoy en día las herramientas presentan los siguientes problemas.

   La precisión: debido a la irregularidad de muchos resultados de búsqueda, que dificulta encontrar la información más relevante que necesita el usuario, descartando la irrelevante.

   Llamada lenta: es la incapacidad de mostrar o encontrar toda la información relevante indexada en la Web. [CDF+98], [Coh99]

II. *Aprender de los consumidores o usuarios individualmente.* Está relacionado con el problema anterior, desde el punto de vista del conocimiento que pueden obtener. El problema se presenta cuando se requiere personalizar las preferencias de los usuarios o del consumidor. Suelen ser problemas relacionados con el marketing, diseño y administración de la Web, etc.

III. *Personalización de la información.* El problema está a menudo asociado con el tipo y la presentación de la información, donde los usuarios no quedan satisfechos del contenido y la organización, mientras navegan por la Web.

IV. *Descubrir nuevo conocimiento de la información disponible en la Web.* Esto se podría decir que es un subproblema del problema en (I), mientras el anterior es pregunta-activación del proceso (Recuperación orientada), éste problema es un dato-activación del proceso, eso presume que ya tenemos una colección de datos en la Web y lo que se quiere es extraer el conocimiento útil. (Minería de datos orientada).

Para dar frente a algunos de estos desafíos en la Web, lo podemos hacer mediante áreas tales como la recuperación de información (RI) ó la extracción de información (EI). Además, dentro de la inteligencia artificial se encuentran las subáreas de aprendizaje automático procesamiento del lenguaje natural [Mae94], que son áreas de investigación relacionadas fuertemente con el proceso de Minería Web, que de alguna manera u otra ayudan a solucionar estos desafíos.

Otro aspecto importante de Internet es el que se le ha dado a nivel empresarial, donde podemos destacar la nueva filosofía de atención al cliente (Customer Relationship Management), la integración de funciones en las empresas u organizaciones (Enterprise Resource Planning), la coordinación con los proveedores (Supply Chain Management) y vendedores (Selling Chain Management), etc. Todos estos modelos tienen un potencial de desarrollo mucho mayor debido a la explosiva conectividad y riqueza de la comunicación que se puede establecer por medio de Internet.

Por lo tanto, la Web pasa a ser una enorme colección de datos e información muy heterogéneos que posee un aumento en problemas de escalabilidad y dinamismo. Por consiguiente, la Web es un área fértil para la investigación de Minería, con esa enorme cantidad de información en línea.

El proceso de la Minería Web, lo podemos definir formalmente como *" el proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web"* [Etz96].

La explotación de la información que se encuentra en la Web se puede realizar de diferentes puntos de vista. Se puede analizar a través del contenido que podemos extraer o encontrar en la Web; este punto de vista es enfocado principalmente en la extracción de conocimiento sobre el contenido de documentos y es llamado Minería Web de Contenido. Otra manera de inferir conocimiento es a través de la organización de la Web o de las relaciones entre los enlaces; esta forma de descubrir conocimiento de la estructura de la Web es llamado Minería Web de Estructura. Y por último, el proceso de extracción de patrones interesantes de la información de navegación o del tráfico del usuario en la Web es llamado Minería Web de Uso.

2.7. Etapas de la minería web

Para poder procesar los datos y transformarlos en información útil, podemos distinguir una serie de etapas dentro del proceso global de la Minería Web. Las etapas son ilustradas en la figura 2.6 y las comentaremos a continuación: [KB00].

- **Selección y recopilación de datos.** Lo primero es determinar que es lo que se quiere obtener y cuales son los datos que nos facilitarán esa información para lograr la meta. Posteriormente se localizan los documentos o archivos a adquirir, se capturan y se almacenan los datos pertinentes. El objetivo de esta etapa es recuperar automáticamente los documentos más importantes, indexándolos para optimizar la búsqueda. El proceso de indexación es complejo debido a la gran cantidad de páginas Web, además que éstas cambian continuamente, por lo cual existen cuatro enfoques de indexación, los cuales son: indexación manual, indexación automática, indexación inteligente o basada en agentes e indexación basada en Metadatos.
- **Extracción y preprocesamiento de información.** Se trata de filtrar y limpiar los datos recogidos. Una vez extraída una determinada información a partir de un documento, ya sea HTML, XML, TEXTO, PS, PDF, LateX, FAQs,... se eliminarán los datos erróneos o incompletos, presentando las restantes de manera ordenada y con los mismos criterios formales hasta conseguir una homogeneidad formal y demás labores enfocadas a la obtención de unos datos originales listos para su transformación por medios automáticos. El objetivo es identificar y etiquetar el contenido esencial del documento para mapear a algún modelo de datos. La extracción de la información entrega nueva información a partir de la estructura del documento y su representación.
- **Minería**. En esta etapa, se descubren automáticamente los modelos o patrones generales sobre un sitio Web, así como por múltiples sitios, utilizando recursos estadísticos, técnicas de Minería de Datos, etc.
- **Análisis.** Una vez teniendo los patrones identificados, es necesario interpretarlos; para esto existe diversas herramientas que permiten entender, ya sea visualmente o algún otro método que facilita la interpretación de dichos patrones.

2.7.1. Técnicas de minería web

Comentaremos brevemente algunas de las técnicas más utilizadas en la Minería Web.

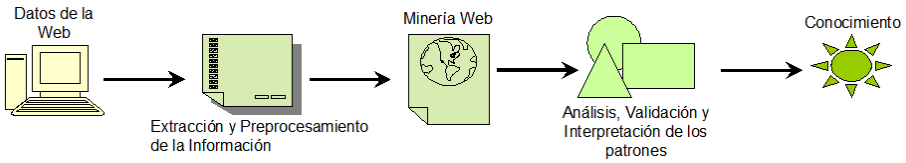

Figura 2.6: *Etapas de la Minería Web*

- **Reglas de asociación**. Por lo general, esta técnica es utilizada para descubrir la correlación entre los accesos de los clientes a varios archivos disponibles en el servidor. Cada transacción está compuesta por un conjunto de URL accedidas por el cliente en una visita al servidor. Por ejemplo, se pueden descubrir que 50 % de los clientes que acceden a una página A también acceden a la página B.

- **Path analysis**. Este análisis es una extensión de un modelo de regresión, usada para probar las correlaciones entre dos o más modelos causales que están siendo comparados. La regresión está hecha para cada variable en el modelo, como un dependiente de los otros donde el modelo indica causas. Los pesos de regresión predichos por el modelo son comparados en una matriz de correlación para las variables, y así se calcula el índice de bondad del ajuste. El mejor ajuste de dos o más modelos es seleccionado por el investigador como el mejor modelo. Esta técnica principalmente se utiliza para el análisis de caminos de navegación.

- **Secuencias de patrones**. Esta técnica se basa en descubrir patrones en los cuales la presencia de un conjunto de ítems es seguido por otro ítem en orden temporal. Analizando estos datos, se puede determinar el comportamiento de los usuarios con respecto al tiempo.

- **Clustering**. La agrupación automática de clientes o datos con características similares sin tener una clasificación predefinida. Puede ser utilizado para estrategias de marketing dirigido según las clases obtenidas. Por ejemplo si se reconoce un grupo de potenciales clientes se les podría enviar las ofertas por correo sólo a ellos.

En la tabla 2.5, podemos ver una clasificación de las técnicas de la Minería Web, agrupándolas en técnicas descriptivas y predictivas. En la siguiente sección veremos con más profundidad la Minería Web y también comentaremos con más detalle algunas de las técnicas más usadas en cada tipo de Minería Web.

Técnicas de Minería Web	
Métodos Descriptivos	**Métodos Predictivos**
a.Visualización	**a.Path Analysis**
b.Aprendizaje No Supervisado *1.Clustering* - Método no jerárquico(Partición) - Métodos jerárquicos(N-TREE) - Métodos no Paramétricos (K-NN K-means clustering,centroides SOM o Redes de Kohonen)	**b.Aprendizaje Supervisado** *1.Clasificación* - Árboles de Decisión - Inducción de Reglas - Redes Neuronales(Simples, Multicapas) - Clasificación Naive Bayes - Aprendizaje Relacional y Recursivo - ILP (Inductive Logic Programming) *2. Categorización*
c.Análisis Estadístico	**c.Secuencias de Patrones**

Tabla 2.5: *Técnicas de Minería Web*

2.8. Tipos de minería web

El término de Minería Web es generalmente usado en tres caminos, Minería Web de Contenido, Minería Web de Estructura y Minería Web de Uso, los cuales son ilustrados por la figura 2.7. A continuación describiremos brevemente cada tipo de Minería Web y posteriormente los analizaremos con mayor profundidad.

La **Minería Web de Contenido** es un proceso automático que va más allá de la extracción de palabras claves, ya que los datos se analizan para poder generar información de los documentos que se encuentran en la Web, ya sea, artículos, material audiovisual, documentos HTML, entre otros. La extensa Web puede revelar más información que la que se encuentra contenida en los documentos, por ejemplo, los enlaces apuntando hacia un documento indican la popularidad del documento, mientras algunos vínculos salen de un documento indican la riqueza o quizás la variedad de temas cubiertos en el documento. Esto puede ser utilizado para comparar las citaciones bibliográficas. En esta área se encuentra la **Minería Web de Estructura**, el cuál consiste en estudiar las estructuras enlaces de los entre o intra documentos, para descubrir patrones útiles de las estructuras de los enlaces. Y por último, la **Minería Web de Uso** es un proceso de descubrimiento automático de patrones de accesos o uso de servicios de la Web, centrándose en el comportamiento de los usuarios cuando interactúan en la Web. A continuación profundizaremos en cada tipo de Minería Web.

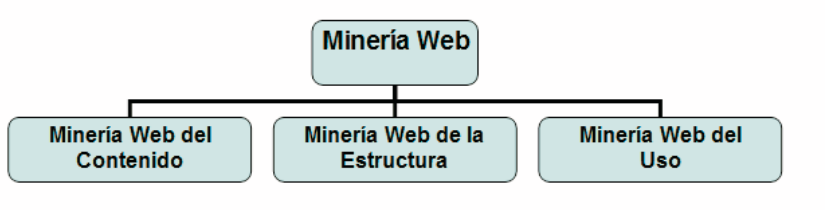

Figura 2.7: *Tipos de Minería Web*

2.8.1. Minería web de contenido

Como ya hemos mencionado anteriormente, en la Minería Web de Contenido los datos se analizan para poder generar información de los documentos que se encuentran en la Web, ya sea, artículos, material audiovisual, documentos HTML, entre otros. En este sentido, la Minería Web de Contenido esta relacionada con la Minería de Texto, donde los textos a analizar son documentos Web.

Las técnicas que se ocupan en esta rama de la Minería Web, varía dependiendo del contenido a tratar: técnicas de recuperación de información, fundamentalmente técnicas estadísticas y lingüísticas, hipertexto, minería de marcado (la información de las marcas contiene información como por ejemplo HTML: secciones, tablas, etc., minería multimedia (para imágenes, audio, videos,. . .) y técnicas de Minería de texto algunas de las cuales ya hemos mencionado en secciones anteriores.

Se pueden considerar dos enfoques relacionados con la Minería Web de Contenido [CMS97], [PTM02]. El primer enfoque se refiere al **basado en agentes** y el segundo enfoque esta **basado en las bases de datos**.

En el enfoque **basado en agentes** existen diferentes categorías, *agentes para la búsqueda*, *agentes para el filtrado* y *categorización*, y agentes para la *personalización de los documentos Web*.

Los *agentes inteligentes de búsqueda*, han sido desarrollados para la búsqueda de información relevante usando características del dominio y los perfiles del usuario para organizar e interpretar la información descubierta. Por ejemplo, ShopBot [DEW96] rescata la información del producto de una variedad de sitios usando sólo información general acerca del dominio del producto. ILA (Internet Learnig Agent) [PE95], el cuál aprende de fuentes diversas de información y traduce estos a su propia jerarquía de concepto. Otros agentes en esta misma área de aplicaciones son: Faq-Finder [HBML95], Harvest [BDH$^+$94], OCCAM [KW96], Information Manifold [KLSS95] y Parasite [Spe97].

El filtrado y categorización de información, donde un número de agentes Web usa varias técnicas de recuperación de información y características de documentos Web de hipertexto para recuperar información filtrando y clasificando por categorías. Por ejemplo, Hypursuit [WVS+96] usa información semántica incrustada en las estructuras del enlace y el documento contenido para crear jerarquías del grupo de documentos de hipertexto, y estructurar un ámbito de información. Otro como BO (Bookmark Organizer) [MS96], combina técnicas jerárquicas de Clustering e interacción del usuario para organizar una colección de documentos Web que se basan en la información conceptual.

Por ultimo, los agentes Web personalizados, aprenden de las preferencias de los usuarios y descubren información en la Web basados en las preferencias de estos y de otros usuarios con intereses similares (usando filtración colaborativa). Algunos ejemplos que podemos destacar son el WebWatcher [AFJM95], PAINT [OPW94], Syskill & Webert [PMB96], GroupLens [RIS+94], Firefly [SM95], entre otros. Para el ejemplo, Syskill & Webert, este utiliza perfiles de usuario y aprende a evaluar páginas Web de interés usando un clasificador Bayesiano.

El segundo enfoque que está **basado en el ámbito de las bases de datos**, donde las bases de datos en la Web están relacionadas con los problemas de administrar y consultar la información en la Web. Hay tres tareas relacionadas con estos problemas: modelado y consultas de la Web; extracción e integración de información; y la construcción y reestructuración de sitios [FLM98].

Este enfoque se centra en las técnicas para organizar los datos semiestructurados en la Web, en colecciones de información estructuradas y usando mecanismos estándar de consultas de base de datos y técnicas de Minería de Datos para analizarlos. La idea principal detrás de este acercamiento es que el nivel mínimo de la base de datos contiene información semiestructurada almacenada en lugares de depósito Web diversos, como documentos de hipertexto, y los metadatos de nivel superiores o las generalizaciones son extraídas y organizadas en colecciones estructuradas, o sea bases de datos de relaciones u orientadas a objetos; a esta estructura se le conoce con el nombre de bases de datos multinivel.

Por ejemplo, en [HCC93] proponen la integración incremental de una porción del esquema de cada fuente de información, en vez de confiar en un esquema global heterogéneo de bases de datos. Los sistemas ARANEUS [MAM97] extraen información relevante de documentos de hipertexto e integran estos en los hipertextos derivados de niveles Web superiores que son generalizados de la noción de vista de bases de datos. También podemos mencionar a los Web Query Systems, donde muchos sistemas de consulta basada en la

Web utilizan lenguajes de consulta estándar de la base de datos como SQL, información estructural acerca de documentos Web. Por último, el lenguaje de consulta W3QL [KS95] combina consultas de estructuras basadas en la organización de documentos de hipertexto, y las consultas basadas en las técnicas de recuperación de información.

2.8.2. Minería web de estructura

Mientras la recuperación de información convencional está enfocada primordialmente al texto de los documentos Web, la Web provee información adicional a través de la forma en la cual los diferentes documentos están conectados por hiperenlaces. La Web puede ser mirada como un grafo cuyos nodos son los documentos y las aristas son los hiperenlaces entre ellos.

La Minería Web de Estructura es el proceso que analiza la estructura de la información usada, que describe el contenido de la Web. La estructura de la información de la Web puede ser clasificado como: intra-página e inter-página.

La estructura de información inter-página, puede analizarse a través de los hiperenlaces y a menudo se llama Web asociado o enlazado a estructuras (Web Linking Structure) [CHMW01]. En este tipo de minería, el enlace de estructura puede representarse como un gráfico, en el cuál los documentos son los nodos y los hiperenlaces son las aristas del gráfico. Existe información útil que se puede descubrir por el procesamiento de las relaciones entre nodos y aristas.

La estructura de información intra-página refiere a las estructuras internas de los actuales tipos de documentos de la Web como HTML o XML, los cuales están usualmente representados por árboles.

La estructura contiene una gran cantidad de información sin descubrir, para ser analizada. Así, una comprensión de estructuras de alto nivel puede emerge sólo a través de un análisis posterior; a éste tipo de análisis también se le llama conectividad de análisis de topología de enlaces. Se pueden considerar dos tipos de descubrimiento de páginas o topologías llamadas Hubs y Autoridades.

Una autoridad puede verse como, páginas altamente referenciadas en un tema específico, como muestra la figura 2.8.

Un Hub puede definirse como el conjunto de páginas comparables para muchas relaciones de autoridad, como muestra la siguiente figura 2.9.

Los hub y autoridades exhiben una relación mutua fuertemente reforzada, ya que un

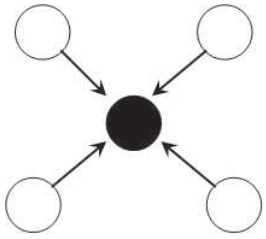

Figura 2.8: *Representación del concepto autoridad*

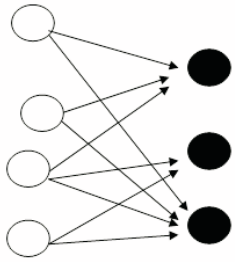

Figura 2.9: *Representación del concepto hub*

hub adquiere mayor peso cuando se acopla a una autoridad. Así mismo, la autoridad adquiere mayor peso cuando se asociado a muchos hubs. Este tipo de análisis es llamado *análisis de conectividad.*

El análisis de conectividad puede usarse para descubrir temas o consultas a una comunidad específica por la computación a través de los hub y autoridades para el tema. Encontrar a las comunidades, está relacionado con el problema de segmentación de la gráfica de NP-completo. Este problema se relaciona con la complejidad computacional en las tomas de decisión para la segmentación de la conectividad de las comunidades encontradas. [FLG00].

Comentaremos algunos estudios y algoritmos relacionados con la Minería Web de Estructura como por ejemplo HITS (Hiperlinks-Induced Tepic Search) [Kle99], el cual alinea las páginas en dos tipos distintos, que guardan una relación de mutua dependencia: autoridades y hubs. Esta idea asume que cuando alguien establece un enlace a una página es porque la considera interesante, y que personas con intereses comunes tienden a referirse a las autoridades sobre un tema dentro de una misma página. Otros algoritmos relacionados con la Minería Web de Estructura son mencionados en [Kle99], también podemos mencionar al Trawling [KRRT99], y el Hyperclass [CDI98].

Un algoritmo muy utilizado en este área de la Minería Web de Estructura es el *PagesRank* [PBMW99], este método asigna valores a las página Web para poder ordenarlas según su categoría o su relevancia, basándose en el grafo de la Web. La idea que subyace es que las páginas Web a las que más apuntan otras páginas serán más importantes que aquellas que tienen pocos enlaces entrantes. Se tiene en cuenta además la importancia de la página que te apunta. En resumen, una página tiene un PageRank alto si la suma del PageRank de sus enlaces entrantes es alta. De esta forma, una página será importante si tienen muchos enlaces entrantes o si tiene pocos enlaces, pero estos son de páginas importantes. PageRank tiene aplicación en la búsqueda, en la navegación y estimación del tráfico. Un ejemplo de la utilización de este algoritmo, es en el motor de búsqueda Google.

Un trabajo pionero en este campo ha sido realizado por [BKM$^+$00], el cuál utilizó datos obtenidos del motor de búsqueda de Internet Altavista a mayo de 1999, donde se obtuvo 203 millones de URL y 1466 millones de hipervínculos, los cuales fueron guardados en forma de grafo. Lo más importante de este trabajo fue que se pudo representar la estructura de la Web, el cual era parecido a la figura de un nudo hecho en medio una soga, o como fue llamado por los autores Connected Core Componet (CCC), el cuál tenía cerca de 56 millones de páginas, y a ambos lados de este nudo existen 44 millones de páginas aproximadamente. También en [KRRT99] hace referencia a la Web como un grafo, donde presentan diferentes algoritmos que basados en el grafo solucionan los problemas de búsqueda de tópicos, enumeración de tópicos, y clasificación.

En [CDI98] se basa en la clasificación de documentos como técnica de estudio, donde son evaluadas dos variantes, una que simplemente anexa el texto de las páginas vecinas (el predecesor y el sucesor) para el texto de la página objetivo. De este estudio [DL01] se concluyó que el texto de los vecinos es demasiado ruidoso para ayudar a la clasificación y se propuso una técnica diferente que incluye predicciones para las etiquetas de clases de las páginas vecinas en el modelo. Se implementó una técnica de relajación que demostró tener mejor desempeño sobre el acercamiento en textos estándar que ignora la estructura del hiperenlace. La utilidad de precisiones de clases para páginas se confirman con los resultados en [OML00], [YSG02]. Una línea diferente de investigación se concentra implícitamente en la estructura de la relación de la Web con Inductive Logic Programming. [DL01], [CDF$^+$00], [RM92], [Spa80], [CSN98].

2.8.3. Minería web de uso

Son muchos los sitios dedicados al comercio electrónico o a proveer información. Estos sitios necesitan aprender cada día sobre los clientes o usuarios que navegan en sus

sitios. Sólo de esta manera podrán dirigir adecuadamente los esfuerzos para mejorar los servicios de marketing y la personalización del sitio.

El descubrimiento de patrones de actividad y comportamiento relacionado con la navegación Web requiere el desarrollo de algoritmos de Minería de Datos capaces de descubrir patrones de accesos secuenciales de ficheros log (algunas técnicas de la Minería Web son mencionadas en la tabla 2.5).

En el siguiente capítulo analizaremos la Minería Web de Uso con mayor profundidad, ya que es uno de los objetivos principales de nuestro trabajo.

2.9. Conclusiones

El Proceso de Extracción de Conocimiento (KDD) de grandes volúmenes de datos ha adquirido una gran importancia tanto en el ámbito empresarial como en el mundo científico. Formando parte de dicho proceso, la Minería de Datos es la etapa central del análisis y explotación de los datos. La Minería de Datos se puede extender a otros campos de aplicación tales como la Web y las bases de datos documentales. De este modo, aparecen otros procesos interesantes de extracción o búsqueda de conocimiento tales como la Minería Stream, la Minería de Texto y la Minería Web. Tras analizar estos procesos, nos hemos centrado es el último de ellos, la Minería Web, por ser parte de los objetivos de este trabajo.

La Minería Web trabaja sobre fuentes de información muy diversas, ya que el principal origen de sus datos se encuentra en la Web y ésta contiene información heterogénea y sin estructurar que dificulta la extracción de conocimiento. La Web posee un gran potencial a la hora de aplicar procesos de minería, al poderse extraer conocimiento de datos relacionados con el análisis de la estructura, con el contenido y con el uso que hacen los usuarios de determinados sitios a través de su navegación. Estas diferencias dan lugar a los diferentes tipos de Minería Web, de los cuales nos centraremos en la Minería Web del uso. El objetivo de estos procesos no es solamente el descubrir patrones, sino también el poder aplicarlos para conocer a los usuarios y poder mejorar el contenido del sitio analizado.

En el siguiente capítulo analizaremos con mayor profundidad la Minería Web de Uso, identificando sus etapas. Tambíen hablaremos de la fuente principal de análisis en este tipo de minería, los llamados archivos log de servidores Web, y describiremos el modelo de datos a considerar para el análisis de estos archivos.

Capítulo 3

Minería Web de Uso: Modelo de datos

En este capítulo realizaremos un estudio más profundo en el área de Minería Web de Uso, analizando las diferentes etapas y las técnicas más utilizadas en esta área.

Al realizar una navegación por la Web, los usuarios dejan huellas digitales (direcciones de IP, navegador, cookies, etc) que los servidores almacenan automáticamente en una bitácora de accesos (log). El análisis de los ficheros log de los servidores Web puede proporcionar información valiosa sobre cómo mejorar la estructura de un sitio Web con objeto de crear una navegación más efectiva y un acceso más eficiente. Algunas herramientas de la Minería Web analizan y procesan estos logs para producir información significativa, como por ejemplo la navegación de un cliente al hacer una compra en línea.

Podemos definir la Minería Web de Uso como *el descubrimiento automático de patrones de acceso o uso de servicios de la Web.* La información de uso de la Web capta actividades del usuario en línea y descubre una gran variedad de patrones conductistas diferentes que recogen el comportamiento del usuario en línea.

3.1. Etapas de la minería web de uso

Hay cuatro etapas que describen la Minería Web de Uso: colección de datos, preprocesamiento de datos, descubrimiento de patrones y análisis de patrones de uso, los cuales comentaremos a continuación (Ver figura 3.1).

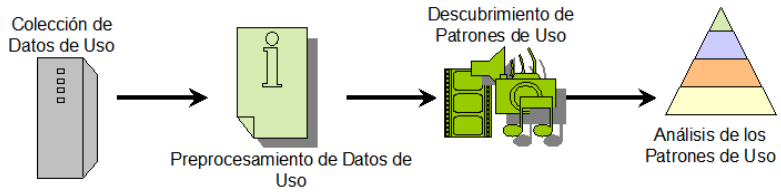

Figura 3.1: *Etapas de la Minería Web de Uso*

3.1.1. Colecciones de datos de uso

El uso de la Web, implica fuentes de información de muy diversa naturaleza tales como:

- *Datos de los registros de servicios del servidor Web.* Los servidores Web almacenan la información de acceso en los registros de servicios del acceso (Log). Estos registros son huellas digitales que el usuario registra cuando interactúa en la Web.

  Para cada sesión de navegación en la Web, se guardan los siguientes archivos: registro de servicios de accesos (Log), que guardan la información de acceso del cliente (estos registros almacenan información como IP cliente, petición de URL, transferencia de bytes, fecha de acceso, entre otros); log de error, que guardan información de los fallos de intento de petición de hechos por el cliente (dentro de los fallos incluyen perder enlaces, fallo de identificación, problemas de timeout, errores del servidor, errores de implementación, una mala petición, métodos no permitidos, entre otros); y log de cookies log, que guarda la información de acceso de sesión entre el cliente y el servidor [CMS99], [NFJK99].

- *Datos de los registros de servicios de servidores proxy.* Un servidor proxy es el que permite el acceso a la Web a varios equipos a través de una única dirección IP y posee un corta fuego, el cual restringe el acceso para una red protegida. Entonces, el log del servidor proxy puede ser usado como fuente de datos para caracterizar el comportamiento de navegación del usuario, compartiendo servidores proxy comunes; una función del servidor proxy es de gateway o también entrada de acceso de redes internas y externas. El gateway se ocupa de dos tipos de particiones: de los servidores externos a cuenta de clientes internos y los servidores internos a cuenta de clientes externos.

- *Datos de los registros de servicios de la máquina del cliente.* Los datos del log residen en el cliente; éste es individualmente el mejor recurso o fuente de datos para

revelar los patrones de navegación del cliente. Donde son las cookies una potente herramienta empleada por los servidores Web para almacenar y recuperar información acerca de sus visitantes.

- *Páginas web.* El modelo de datos del Web se basa en los paradigmas del enlace de hipertexto y la búsqueda de información de textos [BLCGP]. Este modelo se caracteriza porque:

  - La información sólo debe estar representada una vez y mediante una referencia se podrá replicar una copia de ella.
  - Los hiperenlaces permiten que la topología de la información evolucione, modelando el conocimiento humano en cualquier momento y sin restricción.
  - Los documentos en la Web no tienen por qué existir físicamente como archivo; pueden ser documentos ”virtuales” generados por un servidor en respuesta a una consulta.

  La información contenida en la Web se presenta y se relaciona principalmente por medio de documentos *html*, tanto estáticos como dinámicos. De este modo, la World Wide Web puede verse como un conjunto de documentos *html* relacionados por medio de hiperenlaces.

  Los documentos html también han sido representados en función del uso que se hace de ellos, es decir, en función del orden en el que se sigue los hiperenlaces de una página, la secuencia de páginas que se sigue dentro de un mismo dominio Web, entre otras.

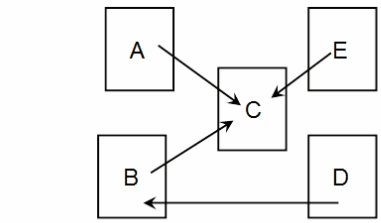

Figura 3.2: *Relaciones de páginas web*

Al final de este capítulo se encuentra un estudio detallado de las colecciones de datos, así como el modelo a utilizar, el cuál será la base para los diferentes experimentos y análisis que realizaremos a lo largo de este trabajo.

3.1.2. Preprocesamiento de datos de uso

El log de acceso del servidor Web contiene los registros del acceso del usuario. Cada entrada en el log, representa una petición de página de un cliente en Internet. Este Log debe ser obtenido del servidor Web para encontrar patrones en él. Normalmente, los ficheros logs son almacenados como archivos de textos bajo un directorio específico en el servidor de Web.

En esta fase del proceso de Minería Web de Uso se debe realizar la limpieza de los datos sucios de los archivos log, eliminar los datos irrelevantes, limpiar el ruido de los datos y datos inconsistentes de los archivos. También se deben identificar las entradas para aplicar alguna técnica de minería que permita agruparlas en unidades lógicas o sesiones del usuario.

Una tarea importante en cualquier aplicación de minería es encontrar datos objetivos adecuados para determinar cuál de las técnicas de minería es la más adecuada para utilizar. El proceso de preparación de datos es a menudo el paso más consumidor de tiempo e intensivo computacionalmente en el proceso de descubrimiento de conocimiento. La Minería Web de Uso no es la excepción: de hecho, el proceso de preparación de datos, a menudo precisa el uso de algoritmos especiales y heurísticos que son comúnmente empleados.

3.1.3. Descubrimiento de datos de uso

Diferentes tipos de datos usan diferentes técnicas de descubrimiento de patrones. El conocimiento descubierto se usa luego para realizar tareas cómo predicción, personalización e incrementar la infraestructura existente en el sitio Web.

Algunas aplicaciones que se realizan en el uso de datos de Web, a través de técnicas de minería, son: análisis estadísticos, clustering, clasificación, descubrimiento de secuencias de patrones, análisis de camino o path analysis y reglas de asociación.

3.1.4. Análisis de patrones de uso

Es la última fase del proceso de Minería Web de Uso. En esta fase del proceso se filtran reglas o patrones poco interesante de un grupo de patrones descubiertos en la fase anterior del proceso, lógicamente esto se realiza para obtener las reglas o patrones más interesantes del análisis.

A continuación comentaremos algunos trabajos previos que se han realizado dentro de esta área de investigación.

3.2. Técnicas asociadas a la minería web de uso

Analizaremos la Minería Web de Uso a través de las técnicas más utilizadas y comentaremos algunas aplicaciones. Las técnicas más utilizadas son **path analysis**, **reglas de asociación** y **clustering**.

Comenzaremos con la técnica **path analysis** o **análisis de camino de navegación**. Para realizar la tarea de análisis de patrones de navegación hace uso de grafos; el grafo representa algunas relaciones definidas en las páginas Web, donde las páginas son representadas por nodos y los hiperenlaces por aristas (R. Cooley B. Mobasher y J. Srivastava) [CMS97].

Dentro del artículo se menciona, el sistema WebMiner, el cuál aplica las técnicas de descubrimiento de regla de la asociación y path analysis como parte del motor de minería del sistema.

Se comenta que el path analysis puede ser usado para determinar la frecuencia de visitas de navegación en un sitio Web. Ya que una vez obtenido los patrones descubiertos, se necesitan apropiadas herramientas para entender, visualizar e interpretar esos patrones, por ejemplo son mencionados el OLAP y el sistema WebMiner.

Otra de las técnicas más utilizadas en la Minería Web es la técnica de **reglas de asociación**. Esta técnica es utilizada en el sistema WebMiner; este sistema está basado en el análisis de las entradas de los usuarios, para descubrir patrones o reglas relacionadas con su navegación. [CMS97]. Otra herramienta que aplica la técnica de reglas de asociación es WLSXP (Web Mining Log Sessionizator XPert) [AMW04], el cuál realiza una limpieza y preprocesamiento de los archivos log, para luego generar sesiones de usuarios y realizar un análisis sobre estas sesiones que permita la generación de reglas de asociación para comprender el comportamiento de los visitantes de su algún sitio Web. En [CM04] se tomó como caso de estudio el sitio de Ridier Internet (http://www.rieder.net.py), y los archivos log de accesos almacenados en su servidor Web. Se eligieron dos técnicas para cumplir con los objetivos del estudio, una de esas técnicas era las reglas de asociación. Se pretendía analizar las entradas de los usuarios para intentar describir patrones de comportamiento del usuario cuando navega por el sitio Web. Con las reglas de asociación se descubrieron las asociaciones y correlaciones entre las referencias o accesos en los archi-

vos log disponibles en el servidor Web y con la otra técnica que era el clustering, que es otra de las técnicas más utilizadas en la Minería Web, se intentó inferir los perfiles de usuarios, validar los patrones encontrados y descubrir nuevos. La técnica clustering fue aplicada sobre las sesiones de usuarios obtenidas en el preprocesamiento. Con la aplicación de esta técnica se pretendió agrupar las sesiones similares, para intentar describir el comportamiento general de los usuarios a través de los subsitios de Rieder.

Otro estudio interesante relacionado con el **clustering** se presenta en [NCRG03], que propone una metodología dimensionable de clustering, inspirada en el sistema inmunológico natural con el poder de aprender continuamente y adaptarse a patrones entrantes nuevos. Los mecanismos inteligentes de búsqueda son cruciales en la Minería Web por la naturaleza combinatoria grande de optimización de muchos problemas. Un sistema inmunológico artificial puede actuar como un monitoreo continuo y el sistema de aprendizaje hace frente a una corriente de datos entrantes con un número desconocido de grupos. Donde el servidor Web juega el papel del cuerpo humano, y las demandas múltiples entrantes desempeñan el papel de virus/antígeno/bacteria que necesitan ser detectados por la técnica basada en clustering. Por lo tanto, el algoritmo clustering desempeña el papel del agente cognitivo de un sistema inmunológico artificial, donde es aquel cuya meta debe continuamente realizar una organización inteligente de los datos ruidosos entrantes en los clusters. Se presenta una técnica de complejidad casi lineal, llamada *hierarchical unsupervised niche clustering* (H-UNC), para minar tanto grupos del perfil del usuario como asociaciones de url en un paso único.

Además algunas técnicas son explicadas muy brevemente en el artículo [SCDT00] como el análisis estadístico, reglas de asociaciones, clustering, clasificación, patrones secuenciales y modelos de dependencias. También podemos ver un análisis general de las diferentes técnicas que se utilizan en la Minería Web de Uso en [AM04]. En el artículo provee una taxonomía detallada de la Minería Web de Uso, algunas investigaciones en el área así como también comerciales. Se presenta una visión general del sistema WebSIFT, que es un ejemplo de un sistema prototípico de Minería Web de Uso, este sistema lo analizaremos en la siguiente sección.

A continuación comentaremos dos ejemplos de sistemas de Minería Web de Uso más relevantes llamados WebMiner y WebSift, donde daremos una visión general del proceso que realizan y las técnicas que utilizan.

3.2.1. Ejemplos de sistemas de minería web de uso relevantes: WebMiner y WebSift

El sistema WebMiner [CMS97], divide el proceso de Minería Web de Uso en: entrada de los datos, preprocesamiento, descubrimiento del conocimiento y análisis de patrones, como muestra la figura 3.3. La entrada de los datos proviene de servidores log (accesos, referencia y agentes), de los archivos html que forman el sitio y algunos datos opcionales como puede ser registros o log de agentes remotos.

La primera parte del proceso es el preprocesamiento que incluye las tareas de limpieza de datos, identificación de usuarios, identificación de sesiones y terminación de caminos. Todas estas tareas se realizan con el objetivo de crear un archivo de sesión de usuario, el cual es utilizado en la siguiente fase de descubrimiento del conocimiento.

Como muestra la figura 3.3, para poder realizar la técnica de minería de reglas de asociación, es necesario agregar un paso más, la identificación de transacciones. A diferencia de de los otros procesos, la identificación de las entradas es la tarea de identificar semánticamente los grupos significativos de referencias de las páginas. Por ejemplo en el caso de la canasta del supermercado, una transacción es definida como todos los ítems adquiridos por un consumidor o cliente en un tiempo.

La etapa del descubrimiento del conocimiento usa técnicas existentes en la Minería de Datos las cuales generan patrones y reglas. En esta fase se utilizan la generación de estadística de uso general, como es el número de ”hits” por páginas, páginas más comunes accedidas y promedio de tiempo transcurridos en cada página. Son generadas reglas de asociación y patrones secuenciales con algoritmos de minería que se encuentran implementados con el sistema WebMiner, pero la arquitectura abierta de este sistema permite fácilmente acomodar algunos algoritmos de minería como el clustering y el path analysis. Estas últimas técnicas fueron posteriormente agregadas en el proyecto WebSift, que es un sistema basado en el WebMiner [CTS99].

La información que es descubierta alimenta a las diversas herramientas de análisis de patrones. El filtro del sitio es usado para identificar reglas y patrones interesantes para comparar el conocimiento encontrado con el uso del sitio Web, o sea ver como el sitio podría ser usado. Como se muestra en la figura 3.3 el filtro del sitio pueden aplicarse a los algoritmos de minería con el objetivo de reducir el tiempo de análisis, o el de descubrimiento de reglas o patrones [CMS99].

El WebSift [CTS99] es diseñado para ejecutarse en el área de Minería Web de Uso, para archivos log con formato NSCA (que incluye los campos de referencia y agente). Co-

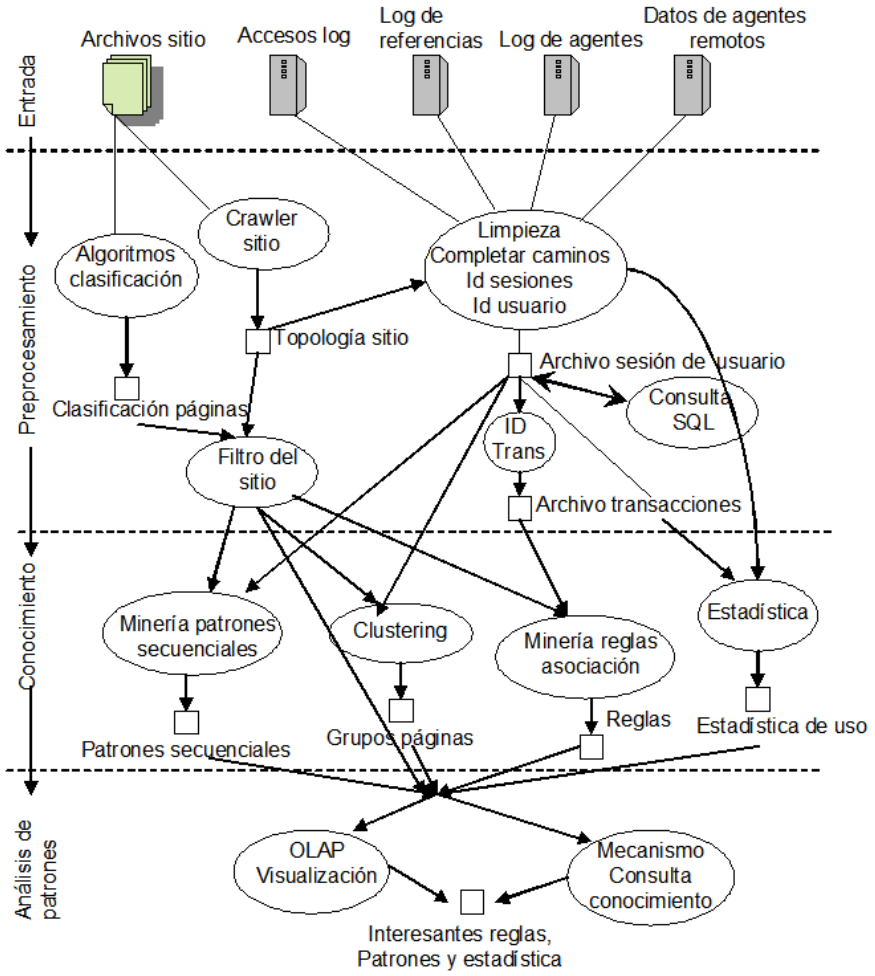

Figura 3.3: *Arquitectura del sistema WebMiner*

mo este sistema esta basado en el sistema WebMiner, que acabamos de comentar, posee una arquitectura muy similar, o sea que el primer proceso que se realiza es el preprocesamiento orientado principalmente al contenido y estructura; donde se limpia los datos, se identifican sesiones y usuarios y por último, se completan los caminos. Permite convertir las sesiones del usuario en episodios. Cada episodio es cualquier subconjunto de todas las páginas contenidas en las sesiones del servidor o todas las páginas de navegación. Existen varios algoritmos para poder identificar los episodios, los cuales son descritos en [CMS99].

3.3. Descripción de ficheros logs de servidores web

La principal fuente de información usada en la Minería Web de Uso es la de los ficheros log del servidor Web. Las fuentes adicionales de datos que son también esenciales para la preparación de datos y descubrimiento del patrón incluyen los archivos del sitio y metadatos, bases de datos operacionales, las plantillas de aplicación, y conocimiento de dominio. Los datos obtenidos a través de estas fuentes pueden ser clasificados en categorías en tres grupos: los datos contenidos (son los datos reales en las páginas de Web, o sea los datos de la página Web que fueron diseñadas para dar información a los usuarios, este usualmente consiste en texto y gráficos), los datos de la estructura (los datos que describe la organización del contenido, la principal clase de estructura de información es la intra-página, que es una página que conecta hiper-enlaces para otra) y los datos de uso (los datos que describen el patrón de uso de páginas Web, algo semejante como las direcciones IPs, las referencias de la página, la fecha, el tiempo de accesos, entre otras) [CMS99].

3.3.1. Archivos logs de web

Un archivo de Web log, es un grupo de datos de un servidor de Web relacionado con la conexión, como host, identidad y autentificación de los usuarios. Cada petición o cada click generan una entrada de un usuario en este archivo, esta información puede ser completada con el comportamiento de navegación. Las acciones que realiza el servidor en relación con el registro de su actividad son las siguientes: para cada fichero enviado al cliente (esto es, cada página html y cada elemento no textual que contiene, como botones, separadores, iconos, etc.), el servidor escribe una línea en un fichero de registro de accesos (access log) Si la transacción falla, algunos servidores escriben la línea en otro fichero: el

registro de errores (error log). Algunos servidores, además, registran el tipo de aplicación que efectúa cada petición (agent log) y el URL desde el que los usuarios llegan a la página en cuestión (referrer log). A continuación analizaremos más detenidamente cada uno de estos ficheros de registro (Ver figura 3.3).

3.3.1.1. Registro de acceso (access log)

Aunque puede cambiar el nombre, casi todos los servidores mantienen un fichero en el que escriben una línea por cada *hit* (es una transacción entre un cliente y un servidor) o transacción que se realiza, es decir, cada petición de un usuario y el resultado de ésta. Se registra el nombre o número IP de la máquina solicitante, la fecha y la hora, el comando, el código de estatus y la cantidad de bytes transferidos. El formato de este fichero se denomina *Common Log File Format* y ha sido consensuado entre los desarrolladores de servidores (lo cual quiere decir que, naturalmente, no todos los servidores siguen este formato). En Common Log File Format, cada línea de texto en el fichero contiene los siguientes campos: (Ver tabla 3.1) nombre del host remoto o número IP si no puede resolverse en el DNS, identificación del usuario (rfc931), a menudo no implementado y sustituido por '-'), autentificación del usuario (sustituido por '-'si no es una página que requiera autentificación), fecha y hora, petición del cliente, código de estado HTTP retornado al cliente, número de bytes enviados. Las siguientes líneas son un ejemplo (falso) del registro de acceso en "Common Log File Format":

Remotehost	**rfc931**	**Authuser**	**Fecha y Hora**
maquina.uji.es	-	-	9/Feb/1996:00:56:56 +0100
Petición	**Estado**	**Bytes**	
GET/documento.html	302	64	

Tabla 3.1: *Formato Common Log File Format*

Aparte de este formato, también existen otros dos como son el *Extended Common Log File Format* (ECLFF) que es una variación de Common Log File Format (Ver tabla 3.2). Este formato agrega dos campos adicionales al final de cada línea, los campos son *refer* (indica referencia de página) y *agent* (indica el tipo de agente de navegación). Un ejemplo de este formato sería:

El otro formato se denomina *Performance Log File Format* (PLFF) que es una variación de Extended Common Log File Format, donde se agrega el campo *time* después del campo de bytes. Cada línea de entrada de un archivo con este formato tiene la siguiente estructura (Ver tabla 3.3):

Remotehost	rfc931	Authuser	Fecha y Hora
maquina.uji.es	-	-	9/Feb/1996:00:56:56 +0100
Petición	**Estado**	**Bytes**	**Refer**
GET/documento.html	200	1240	http://www.skyweb.com/
Agent			
”Mozilla/4.0 (Win95; I)”			

Tabla 3.2: *Formato Extended Common Log File Format*

Remotehost	rfc931	Authuser	Fecha y Hora
maquina.uji.es	-	-	9/Feb/1996:00:56:56 +0100
Petición	**Estado**	**Bytes**	**Time**
GET/documento.html	200	1240	75
Refer	**Agent**		
http://www.skyweb.com/	”Mozilla/4.0 (Win95; I)”		

Tabla 3.3: *Performance Log File Format*

Estos 3 formatos son los más comunes y son usados por la NCSA, Apache, Samba, entre otros servidores http.

A la vez en los servidores Web también se almacenan otros tipo de archivos log que están relacionados con otros aspectos de la navegación del los usuarios. A continuación analizaremos algunos de estos archivos.

3.3.1.2. Registro de errores (error log)

Algunos servidores filtran los mensajes de error a un segundo fichero a fin de facilitar su análisis. Las siguientes líneas son un ejemplo de entradas en el registro de error:

- *[Thu Feb 29 00:17:43 1996] httpd: send aborted for 255.255.255.255*
- *[Thu Feb 29 00:21:25 1996] httpd: send aborted for maquina.dominio.es*
- *[Thu Feb 29 00:38:49 1996] httpd: access to /Web/noserver.html failed for 204.19.31.129,*
- *[Thu Feb 29 00:38:52 1996] httpd: send aborted for otro.dominio.edu*

Así, la primera línea indica que el jueves 29 de febrero a las cero horas, diecisiete minutos y cuarenta y tres segundos se abortó un comando ”send” desde el número IP

255.255.255.255 (ficticio). En la tercera línea, en cambio, se informa que un usuario ha pedido un fichero que no existe. Es necesario corregir el error (cambiando de sitio el fichero ’/Web/noserver.htmló el ’link’). Este es el tipo de información útil para los gestores del contenido del servidor.

3.3.1.3. Registro de referencias (referrer log)

También existe en ciertos servidores un registro de las URLs, que indican cuál fue la siguiente página visitada por el usuario. Sin embargo, no todos los servidores proporcionan esta información. El formato de dicho registro (en el servidor httpd NCSA,) es el: URL origen→ URL destino , por ejemplo:(Ver tabla 3.4)

URL Origen	→	**URL Destino**
http://www.w3.org/Servers.html	→	/spain-www.html
http://guide-p.infoseek.com/NS/tables/DB?C923&db=78	→	/bbedit-html-extensions.html
http://www.yahoo.com/Regional/Countries/Spain/	→	/spain-www.html

Tabla 3.4: *Ejemplo de Formato Referrer Log*

La primer línea indica que un cliente ha recuperado el fichero ”/spain-www.html” del servidor siguiendo el ”link”que existe. La segunda línea indica que procede de una búsqueda en InfoSeek y la tercera una entrada en el catálogo de Yahoo. Saber desde donde llegan los usuarios proporciona información sobre qué referencias a nuestras páginas existen y cuales son las más utilizadas.

3.3.1.4. Registro de agentes de usuario

Finalmente, algunos servidores registran qué agente de usuario se ha utilizado en cada transacción. Este fichero nos permite averiguar qué agentes usan predominantemente los usuarios, por ejemplo guardaría *Mozilla 4.0 (Win95,1)*, si estuviera el usuario utilizando ese tipo de agente. En la figura 3.4 podemos ver un resumen de los distintos tipos de archivos log que podemos encontrar en algunos servidores web.

3.3.2. Herramientas de análisis de logs

Existen múltiples herramientas para analizar los ficheros de registro de acceso, para extraer estadísticas, realizar informes (en html y texto) e incluso hacer gráficos. Entre ellas

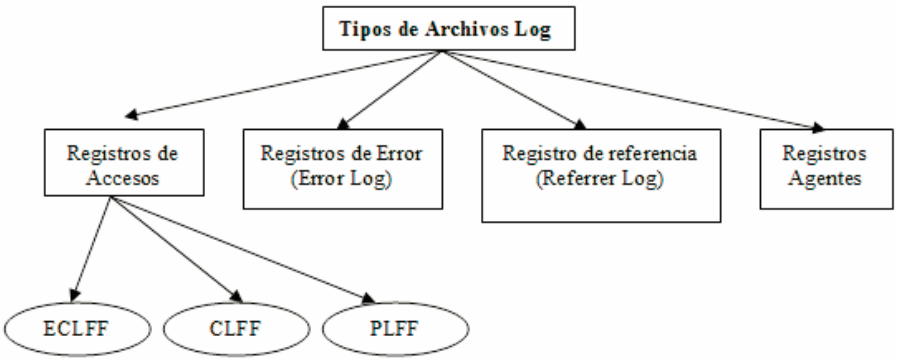

Figura 3.4: *Tipos de Archivos Log*

podemos mencionar algunos productos conocidos como *Webtrends, Getstats, Analog, Microsoft Intersé Market Focus, entre otros* y a continuación mencionaremos algunos, explicando cual es su utilidad.

Podemos mencionar el *3Dstats, M5 Analyzer, Web Log Explore, eWebLog Analyzer* que son generadores de estadísticas de acceso, con las cargas medias por día y hora, el usuario puede caminar a través de un escenario estadístico y examinar el gráfico de barras desde diferentes puntos de vista para informes diarios, semanales o mensuales. *WebTrends*, trabaja principalmente sobre los ficheros log del servidor, permitiendo capturar información sobre las cookies, sitios de referencias, identidad de los navegadores y detalles del usuario. Por último, mencionar el Web Mining Log Sessionizator XPert [AMW04], es una herramienta de procesamiento y análisis, que permite la generación de reglas para comprender el comportamiento de los visitantes de un sitio Web.

En resumen, la gran mayoría de estos analizadores sólo entregan resultados del punto de vista estadístico, y no un proceso de minería, con el cual se podría inferir de una manera más claro sobre lo que está sucediendo en el sitio Web.

3.4. Modelo de datos

Para un análisis de esta información es necesario un modelo de datos. Este modelo nos permitirá tener claro los elementos con los cuales trabajaremos a lo largo de los diferentes experimentos que realizaremos en los próximos capítulos.

Para conocer las acciones que realizan el usuario en la Web, es necesario analizar los archivos de entradas a la Web o archivos logs, los cuales están almacenados en los servidores Web.

Esta información que se encuentra en los ficheros es a veces incompleta, ya que no recogen todas las acciones de un usuario debido a las copias caché, tanto locales como proxies. Otra de las dificultades es la identificación de los usuarios y de las sesiones, ya que si no se tiene la identificaciones explícita de los usuarios, es complicado identificar las sesiones porque un usuario puede utilizar diferentes máquinas y diferentes navegadores en cada momento.

3.4.1. Preprocesamiento

Como la información que se encuentra en los ficheros log carece de cierta estructura para poder ser analizadas, es necesario realizar un preprocesamiento de la información para su análisis.

Esta etapa consiste en eliminar en una primera fase los elementos ruidosos que podemos encontrar en los ficheros log. Dentro de estos elementos ruidosos podemos mencionar a las imágenes, javascript, a las entradas que son idénticas a otras que se pueden presentar y a todo lo que sea irrelevante para el análisis.

Simultáneamente a la fase de limpieza, se puede realizar una identificación de entradas y así saber cuales son las entradas reales que el usuario va registrando durante su navegación y así lograr dar una estructura al archivo para poder ser analizado.

Un aspecto importante dentro de las entradas son las sesiones de usuarios. Por consiguiente, es importante la identificación de las sesiones de usuario, para poder aplicar alguna técnica de minería que nos permita por ejemplo, la obtención de grupos de usuarios que posean las misma características. A continuación definiremos que es una sesión de usuario para poder aplicar algún método que nos permita identificarlas.

3.4.2. Definición de sesión de usuario

Una sesión de usuario está definida por una secuencia de accesos temporales a un sitio particular de la Web por un usuario [MAFM99], [Arl]. Cada sesión de usuario es una representación lógica de una visita de un usuario a algún sitio Web (usualmente dentro de cierto intervalo de tiempo).

Luego de la preparación de los datos, se obtiene un conjunto de sesiones de usuarios **S**, definida como:

$$S = \{s_1, s_2, \ldots, s_m\} \tag{3.1}$$

Donde cada sesión de usuario *S* está definida por un conjunto de páginas *P*, que se define como:

$$P = \{p_1, p_2, \ldots, p_n\} \tag{3.2}$$

Las sesiones de usuarios pueden ser vistas conceptualmente como una matriz sesión-página *mxn* donde:

$$UP = [w(s_i, p_j)], 1 \leq i \leq m, 1 \leq j \leq n \tag{3.3}$$

Donde $w(s_i, p_j)$ representa el peso de la página p_j en la sesión de usuario u_i. El peso puede tomar valores binarios, indicando la existencia o no existencia de la página en la sesión o puede ser una función de ocurrencia o duración de la página en la sesión [XMZ05].

Dentro de las etapas de la Minería Web de Uso, la etapa de preprocesamiento es una de las más importantes, en la cual es necesario limpiar los datos e identificar las sesiones de los usuarios. La idea de la limpieza de los datos es eliminar la información que no sea importante para el análisis del comportamiento de los usuarios, como por ejemplos gráficos, imágenes, javascripts, etc. y también eliminar las entradas redundantes dentro del conjunto de datos, para impedir resultados imprecisos y poco representativos. Una extensa descripción del preprocesamiento de datos y métodos de preprocesamiento la podemos encontrar en [MCS99].

3.4.3. Identificación de sesiones de usuarios

Hay varias formas para identificar a las visitas individuales a un sitio Web. La solución más obvia es dar por supuesto que cada dirección IP identifica a una sola visita. No obstante, esto no es muy preciso ya que por ejemplo, una visita puede ganar acceso a la Web de diferentes computadoras, o muchos usuarios pueden usar la misma dirección IP (si es usado un proxy). Otra suposición que se podría hacer es que cada acceso realizado del mismo Host durante un cierto intervalo de tiempo podría provenir de un mismo usuario.

Un enfoque más preciso en la identificación del usuario, son las conocidas como Cookies o simplemente la inscripción que realiza el usuario en el sitio Web.

3.4.3.1. Método de timeout

Una vez que se tiene el usuario identificado, el siguiente paso es realizar la identificación de las sesiones, dividiendo cada transacción o clicks de cada usuario en sesiones. La solución más usual en este caso es ajustar un intervalo de tiempo y dar por hecho que los accesos consecutivos dentro de ella forman parte de la misma sesión, o ajustar un intervalo de espera máximo, donde las entradas o accesos consecutivos que excedan forman parte de sesiones diferentes. Este método para identificar sesiones de usuario es conocido por el nombre timeout o intervalo de espera [CPY96], el cual definiremos formalmente a continuación.

Formalmente, podemos definir un clickstream como $R = r_i, \ldots, r_q$ que será la secuencia de click que realiza el usuario y que son guardados en los archivos Web log cuando el usuario navega en un sitio y realiza click sobre imágenes, un link, etc. Cada de esos click los notaremos como r_i, con $1 = k = q$. Asumiremos que todos los clicks provenientes del usuario vienen de una misma IP, caracterizaremos al clickstream como:

- Sea r_k el k_{th} click del clickstream *R* de una dirección IP en t_k segundos.
- Sea r_{k+1} el $(k+1)^{th}$ click del clickstream de la misma dirección IP en t_{k+1} segundos después del click r_k.
- T es el tiempo de espera calculada como diferencia entre ambos click en un sitio Web, esto es $T = (t_{k+1} - t_k)$.
- Entonces, si $T < \beta$, siendo β el tiempo de espera máximo, entonces el click r_k y r_{k+1} son consideradas partes de la sesión S_i. En otro caso, Si $T > \beta$, entonces el click r_k es estimada como el final de la sesión S_i, mientras el clicks r_{k+1} es el clickstream inicial de la sesión S_{i+1}.
- Por lo tanto, la duración de una sesión de usuario puede ser calculada como:

$$\sum_{k=1}^{q} r_k \tag{3.4}$$

donde *q* es el número de click en *k*-clickstream de una sesión.

De un punto de vista más práctico podemos decir que al preprocesar un archivo log del tipo *Extended Common Log File Format (ECLLF)*, lo primero que se debe realizar es eliminar todas las entradas redundante del conjunto de datos, para poder organizar en sesiones de usuarios. Estas sesiones de usuarios serán organizadas asumiendo que cada IP o Host que se encuentran en los datos corresponde a un único usuario. De esta manera podemos organizar las sesiones de usuarios dentro de un umbral de tiempo relacionado con las IPs. Por ejemplo, si la IP A es igual a la IP B y a la vez la diferencia de accesos entre ellas están dentro del umbral de tiempo, podríamos decir que la IP A y IP B corresponden a la Sesión 1. Por lo contrario, si ambas IP son diferentes entonces diríamos que la IP A corresponde a la sesión 1 y la IP B corresponde a la sesión 2. Si posteriormente estos accesos se vuelven a repetir, solo tendríamos que determinar que si el la diferencia en el tiempo este dentro del umbral para poder determinar si corresponde o no a la sesión o estaríamos presente a una nueva sesión.

Para ver esto más ilustrativamente, lo veremos reflejado en la tabla 3.5, en la cual se muestran algunos de los campos existentes dentro del archivo log, como son la IP o host, la petición del usuario, y la fecha y hora de la conexión. Además a este archivo se ha agregado tres nuevos campos, los cuales representan la identificación de la sesión a la cual corresponda la transacción, la cantidad de click que se realizaron dentro de la sesión y la diferencia del tiempo que existe entre las entradas dentro de una sesión (en segundos).

El umbral de corte para comenzar una nueva sesión es determinado por el experto que esta realizando el análisis, por lo general el tiempo óptimo según estudios es de 30 minutos. Catledge y Pitkow en [CP95] propusieron 25 minutos como tiempo máximo, aunque generalmente se estima que media hora entre un acceso y otro es la medida adecuada. Cada vez que comienza una sesión nueva su tiempo será igual a 0.

IP o Host	**Id sesión**	**Fecha/Hora**	**Tiempo**
33.red-83-33-.dynamicip.rimade.net	1	[18/Jun/2006:07:41:14+0200]	0
lj2591.inktomisearch.com	2	[18/Jun/2006:07:41:20+0200]	0
70.42.51.20	3	[18/Jun/2006:07:41:35+0200]	0
33.red-83-33-8.dynamicip.rima-tde.net	1	[18/Jun/2006:07:41:39+0200]	25
clickstream	**Click**		
/alumnos/mlii/prolog	1		
/download/guia	1		
/proyectos/silviaacid/basd	1		
/alumnos/oscpi/fecha	2		

Tabla 3.5: *Ejemplo de identificación de sesiones*

3.4.3.2. Otros métodos de identificación de sesiones

Existen otras formas de identificar las sesiones de usuarios. Ya hemos mencionado y demostrado el método más común y más simple que es el timeout [HG00], [CP95]. Ahora analizaremos algunos trabajos relacionado con la identificación de sesiones de usuarios.

El primer método que comentaremos es el llamado *reference length*, el cual asume que la cantidad de tiempo que un usuario gasta en una página está correlacionada en el sentido que si la página es ”auxiliar” o ”de contenido” para ese usuario. Una vez que las páginas están clasificadas a través de un cálculo de estimación entre páginas auxiliares y de contenido basadas en el histograma, una sesión es detectada cuando una página de contenido es encontrada. El problema con este método radica en que sólo una página de contenido es incluida en cada sesión. Esto no es lo más óptimo, ya que un usuario miran más de una página de contenido para un sólo propósito de recuperación [CTS99] y en relación con el método timeout con el cual nos basamos para nuestros estudio, el método timeout nos entregaría sesiones de usuarios más reales que con este otro método.

Otro método al cual comentaremos brevemente, es el método de *detección de sesiones a través de un modelo de lenguaje estadístico*, donde la meta es predecir la probabilidad de la secuencia a través de tópicos relacionados con las sesiones. Estos tópicos se agrupan según la relación de las entradas de los ficheros log con algún tópico común o las entradas que son inconexas. Este modelo provee un acercamiento simple, natural para segmentar a los log. Pensando en un grupo de objetos de algún tópico común que es frecuentemente visitado uno tras otro. En este caso, la entropía de la secuencia está baja. Sin embargo, cuando un objeto nuevo es observado en la secuencia esto es relevante para el tópico original. La introducción de este objeto nuevo causa un incremento en la entropía de la secuencia porque es raramente visitado después de los objetos precedentes. Si el cambio en la entropía pasa un umbral, un límite de sesión podría ser antepuesto al objeto nuevo. En otras palabras la incertidumbre (la cual esta medida por la entropía) dentro de una sesión debería ser apenas constante, permitiendo un nivel fijo de variabilidad dentro de un tópico. Sin embargo, cuando la entropía aumenta más allá de un umbral, ésta presenta una señal clara donde la actividad del usuario ha cambiado para otro tópico. De esta manera, se ajusta un límite de sesión en el lugar donde la entropía se altera [HPS04].

Existen otros trabajos que presentan distintos métodos para la identificación de sesiones. Manila y Toivonen [MT96] usaron las páginas de acceso como medio para descubrir rutinas. [CPY96], aportaron la identificación de diferentes sesiones de un usuario a través de los referentes de máximo avance (maximal forward references), esto es, el máximo grado de profundidad en la navegación antes de salir de la Web o volver por el camino

inverso que se pueden ver en [HPS04] y por último podemos hace referencia al método de la reconstrucción de sesiones de usuarios a través de un método heurístico [AR03] y [SMBN03]. También podemos citar el trabajo realizado por R. Valenzuela [Val06] donde plantea una identificación de sesiones difusas.

3.5. Conclusiones

La Minería Web de Uso se centra en el análisis de los archivos logs (registros que se guardan en los servidores webs sobre los accesos de navegación). Mediante diversas técnicas provenientes de la minería clásica, la Minería Web de Uso extrae patrones de navegación y de comportamiento de los usuarios ante cierta organización de contenidos o estructuras de texto, preferencias del usuario, etc. Entre estas técnicas se encuentran las reglas de asociación, el clustering y la secuencia de patrones, aunque hay otras más específicas para los datos web de uso tales como el path analysis.

El uso de dichas técnicas aportan conocimiento no explícito que no se obtiene normalmente con la mayoría de las herramientas existentes para el análisis de este tipo de datos, que sólo dan resultados desde un punto de vista estadístico de los accesos al sitio Web.

Para la aplicación de estas herramientas, hemos descrito el modelo de datos a considerar, incluyendo la definición de sesión de usuario. También hemos estudiado diferentes métodos de agrupamiento de sesiones de usuario, siendo el método timeout el que hemos elegido para nuestros análisis.

En el siguiente capítulo, introduciremos la lógica difusa como posible técnica para la mejora de algunos procesos de la Minería Web de Uso. Concretamente, nos centraremos en la aplicación de las reglas de asociación difusas en este tipo de minería, que junto con el clustering difuso, son las dos técnicas principales a estudiar en este trabajo.

Capítulo 4

Minería web de uso y reglas de asociación difusas: Análisis de patrones de navegación

El objetivo principal de este capítulo es poder determinar patrones de navegación del usuario, y así conocer su comportamiento por la Web. Para ello, aplicaremos una de las técnicas más utilizadas en la Minería Web de Uso, las reglas de asociación, pero desde el punto de vista difuso, permitiendo además al usuario configurar la estructura de las reglas.

La información obtenida sobre a partir de los patrones podría ser utilizada para re-estructurar los sitios web o también desde el punto de vista del marketing para conocer cuáles son las preferencias de navegación de clientes potenciales.

4.1. Minería web y lógica difusa

Toda la información que se genera en este proceso se almacenada en bases de datos, en documentos o simplemente en la Web puede tener características difíciles de procesar, desde el punto de vista de la extracción de conocimiento, donde esta información puede ser incompleta, imprecisa, incierta o vaga.

Para representar y manejar este tipo de datos se puede utilizar la lógica difusa, uno de los grandes pilares del Soft Computing [Zad75], que nos ayuda a manipular los asuntos relacionados con comprensión de patrones, datos ruidosos e incompletos, información

de técnicas mixtas e interacción humana. El principal objetivo de la lógica difusa es el estudio de los principios de lo que se ha llamado razonamiento aproximado, es decir, aquel razonamiento que puede ser impreciso o poco fiable. Los conceptos de imprecisión se representan mediante el uso de variables lingüísticas [Zad75] y el grado de pertenencia de cada valor del universo de referencia sobre el que está definido el conjunto difuso que da significado a un término lingüístico, expresa el grado de compatibilidad de ese valor con dicho término.

La lógica difusa tiene como motivación aportar un marco más general que el de lógicas anteriores para el tratamiento de la imprecisión y de la incertidumbre en la información. Desde la aparición de esta teoría son incontables las aplicaciones que se han hecho de ella en el mundo de la investigación en general, y en particular en el área de las ciencias de la computación.

Hay un aumento en el rol actual de la lógica difusa en el área de Minería de Datos [MPM02]. El análisis de datos reales del mundo en la Minería de Datos a menudo trabaja sobre transacción con diferentes tipos de variable (datos simbólicos y datos numéricos) [Yag96]. Diversos navegadores de datos han sido implementados usando la teoría de los conjuntos difusos [Bal96].

En el área de la Minería de Datos, la Lógica Difusa esta principalmente preocupado de identificar patrones interesantes y describirlos de una manera concisa y significativa. En [MPM02] podemos ver una visión general de la relación de la Lógica Difusa y la Minería de Datos.

En la Minería de Texto, la lógica difusa se ha empleado para el análisis y la visualización de las incidencias de texto libre para la toma de decisión [SH01]. En [Jus04] se presenta una aproximación basada en reglas de asociación difusas para la extracción de conocimiento, como una herramienta concreta de Minería de Texto. La Minería de Texto y la lógica difusa juegan un papel progresivamente importante en descubrimiento biomédico [MH02], [CH05].

Podemos decir de forma general, que en la Minería Web y la lógica difusa se relaciona con áreas como la recuperación de información [Yag00], [GK00]. También las técnicas más utilizadas en el proceso son el clustering [MKH97] y las reglas de asociación [Gye00], [AM04]. En [EJMBSV06a], [EJMBSV06b], donde podemos ver aplicaciones de Minería Web con lógica difusa.

Según Zadeh, a parte de la medicina, Internet es otro de los ámbitos donde es posible que se registre la mayor proliferación en las aplicaciones de la lógica difusa. Con vista al

futuro se vaticina un amplio campo de expansión de su teoría, de modo que los ”buscadores” serán capaces de dar con la respuesta exacta, aun con preguntas formuladas ”a partir de conceptos bastante difusos”, como por ejemplo: ”como que distancia existe entre la ciudad chilena con mayor habitante y la mayor de Argentina”.

Otro de los ámbitos con gran proliferación de la lógica difusa en los próximos años seguirá siendo los relacionados con el control de productos de consumo, del control de aparatos técnicos como los ascensores, el control de la contaminación, los procesos de manufacturación y especialmente la Medicina.

Analizaremos algunas técnicas asociadas a los distintos tipos de Minería Web y la lógica difusa que podemos encontrar en la literatura. Podemos ver un análisis del Soft Computing, la Minería Web, lógica difusa y técnicas que se presentan en esta área en [AM04],[PTM02].

En el artículo [KC01] se presenta un sistema que provee a usuarios resultados personalizados derivados de un motor de búsqueda basado en los usos de los enlaces de la estructura. Este artículo está relacionado con el área de Minería Web de la Estructura, donde en esta área se utilizan principalmente técnicas de recuperación de la información, Clustering, por mencionar algunas de las más habituales.

Luego de realizar un análisis bibliográfico en el área de la Minería Web, podemos decir con mayor seguridad que las técnicas más utilizadas son las reglas de asociación y el clustering. Es por esta razón que serán estas técnicas las cuales estudiaremos a lo largo de nuestro trabajo. A continuación analizaremos las reglas de asociación difusas para continuar en el próximo capítulo con el análisis del clustering difuso.

4.2. Asociación en la minería web: Reglas de asociación difusas

En esta sección comentaremos algunos trabajos realizados en el área de la Minería Web relacionados con las reglas de asociación difusas, para luego realizar un análisis más detallados con las reglas de asociación y posteriormente las reglas de asociación difusas.

Uno de los problemas que se presentan en la Web está relacionado con la clasificación de documentos Web. En [HSCL01] se comenta sobre un método automático de clasificación o categorización de documentos Web, *basado en el concepto de asociación difusa*. Se muestran los resultados de los experimentos realizados, usando conjuntos de datos de dos portales Web: Yahoo y Open Directory Proyect.

Podemos hacer referencia al artículo [MBKV+02], el cuál habla de una aplicación de *reglas de asociación difusas* para el refinamiento de búsqueda, a partir de un conjunto inicial de documentos recuperados del Web, donde las entradas del texto son construidas y las reglas extraídas. Estos trabajos antes mencionados están dentro del ámbito de la Minería Web de Contenido.

En [WSP01] se utilizan la técnica de *reglas de asociación difusas* en conjunto con un árbol de índice difuso, para mejorar la exactitud y la eficiencia de predicciones de caminos de accesos Web, usando la metodología basada en casos de razonamiento (CBR) y además se propone una estructura para la personalización (a través de perfiles de usuarios y ficheros Web log).

La obtención de reglas de asociación clásica se realiza en bases de datos transaccionales. En ellas, cada transacción contiene o no contiene un ítem determinado. Por lo tanto, los atributos de las transacciones pueden considerarse booleanos [BARP: Boolean Association Rules Problem]. Un algoritmo típico para resolver este problema es Apriori. Sin embargo, cuando el proceso de Minería de Datos se aplica a otros tipos de bases de datos, los atributos pueden ser categóricos o numéricos [QARP: Quantitative Association Rules Problem]. De la misma manera sucede en la Web y resulta muy atractiva para los estudios de marketing de las organizaciones comerciales, que se dedican o tienen alguna relación con el comercio electrónico.

Para nuestros análisis utilizaremos las reglas de asociación difusas, ya que con estas podemos obtener mejores representación y saber realmente lo que esta sucediendo con el comportamiento del usuario mientras interactúa en el sitio Web.

4.3. Reglas de asociación

Siendo I el conjunto completo de ítems $T \subseteq I$, una transacción es un conjunto de ítems al que se le asocia un identificador único TID. Una transacción contiene un conjunto de ítems X si $T \subseteq I$. A partir de aquí denominaremos ítemset a un conjunto de ítems (para evitar confusiones cuando hablemos de conjuntos de ítemsets). Luego, una regla de asociación es una implicación de la forma $X \rightarrow Y$, donde X e Y son conjuntos de ítems de intersección vacía (ver Apéndice B).

Dado un conjunto de transacciones D, se trata de obtener todas las reglas de asociación que tengan una fiabilidad y una relevancia superiores a unos umbrales especificados por el usuario (Mínimo Confianza y Mínimo Soporte).

Como caso particular, los algoritmos de extracción de reglas de asociación se pueden aplicar tanto a bases de datos relacionales como a cualquier conjunto de datos con algún tipo de estructura que permita su análisis. Entonces un ítem será un par (atributo, valor) y podemos imponer la restricción adicional de que todos los ítems de un itemset han de corresponder a atributos.

Del punto de vista de la extracción de las reglas de asociación el problema se puede dividir en dos subproblemas:

I. Generar todas las combinaciones de ítems, entendiendo éstas como itemsets, con un soporte por encima de cierto umbral, previamente definido, el llamado soporte mínimo (minsupp). Esas combinaciones suelen encontrarse en la literatura con el nombre de ítemsets frecuentes.

II. Dado un itemset frecuente $Y = i_1, i_2, \ldots, i_k \geq 2$, generar todas las reglas que contengan todos los ítems de ese ítemset. Para ello, se toman todos los subconjuntos no vacíos X de Y y se generan las reglas $X \rightarrow Y . X$ que cumplan que su confianza es mayor que cierto umbral al llamado confianza mínima (minconf). El valor de la confianza viene dado por *soporte (Y)* / *soporte (X)* (ver Apéndice B).

Uno de los principales inconvenientes que se presentan en la extracción de las reglas de asociación, en bases de datos o conjuntos de datos que sean lo suficientemente voluminosos, son los costos de tiempo como el espacio necesario que en muchos de los casos pueden resultar inviables.

Para lograr los objetivos en la extracción de las reglas de asociación es necesario trabajar con todos los itemsets posibles. Por lo tanto, si tenemos *m* ítems, quiere decir que hay que considerar 2^m posibles itemsets. Afortunadamente los algoritmos existentes nos permiten aplicar técnicas heurísticas para deducir en la medida que sea necesario el número de itemsets que se consideren, de acuerdo a la estimación de si podrán o no ser frecuentes.

Aún contando con esa capacidad de los algoritmos para disminuir los requerimientos en tiempo de proceso y en espacio de memoria, mejorando en definitiva la eficiencia del procedimiento, todavía se puede encontrar con otros problemas, ésta vez asociados a la aplicación que les pueda dar el usuario final. El conocimiento obtenido es muy dependiente del contexto al que pertenece la información contenida en la base de datos original. Por este motivo, suele ser conveniente y necesaria la intervención de un experto humano que pueda dar una interpretación de las reglas obtenidas, indicando cuáles de ellas son

potencialmente útiles y cuáles no, debido por ejemplo a su trivialidad.

Pero la labor del experto humano puede verse entorpecida si el conjunto de reglas obtenido es demasiado amplio. Es por eso que, de cara a optimizar la obtención y posterior interpretación de reglas de asociación, dentro de una base de datos se pueden establecer ciertas restricciones [CVS04]:

- *Restricciones sintácticas.* Estas restricciones limitan los ítems que pueden aparecer en una regla. Por ejemplo, podemos estar interesados sólo en las reglas que tengan un ítem específico en el consecuente o en el antecedente, o en una combinación de restricciones.
- *Restricciones de soporte.* Podemos estar interesados sólo en las reglas cuyos ítems aparezcan en un porcentaje de las tuplas de T por encima de un soporte mínimo. Esto quiere decir que para que la información que nos da la regla tenga cierto peso es necesario que aparezca con cierta frecuencia en la base de datos.
- *Restricciones de cumplimiento.* La confianza nos da una medida de la fuerza de una regla. Nos informa sobre la dependencia entre la aparición del consecuente si aparece el antecedente. En general, interesa que la confianza supere un mínimo.

4.4. Medidas de reglas de asociación

Las medidas de soporte y confianza evalúan el interés y el grado de cumplimiento de las reglas de asociación con un enfoque meramente estadístico, si bien es cierto, el uso del soporte es bastante generalizado y se acepta como la mejor opción para medir la importancia ya que según [Ser03], reúne las siguientes ventajas:

- Es adecuado para la tarea de medir la relevancia estadística de una regla.
- Es una medida con significado intuitivo, y por tanto la interpretación de los valores de soporte es relativamente sencilla para el usuario.
- El uso del soporte contribuye al diseño de algoritmos eficientes para la búsqueda de reglas de asociación.

No ocurre lo mismo con la confianza, como medida del grado de asociación, implicación o dependencia entre antecedente y consecuente, ya que ha recibido diversas críticas. Los argumentos en contra de esta medida son los siguientes:

- La confianza no mide adecuadamente el grado de independencia estadística entre el antecedente y el consecuente.
- La confianza no refleja la dependencia negativa entre antecedente y consecuente.
- Por último, la confianza es una medida de probabilidad confeccionada. La probabilidad condicionada no es intuitiva, y por esta razón resulta difícil para un usuario no experto establecer umbrales mínimos de confianza semánticamente significativos a la hora de obtener reglas de asociación.

Para intentar resolver el problema se plantea en [San99] el uso del *factor de certeza*, que es una representación de la incertidumbre asociada al conocimiento, su principal objetivo es proporcionar unas medidas de incertidumbre más intuitivas que las medidas de probabilidad condicionada y sin algunos problemas de la teoría de la probabilidad.

Tal como se ha planteado, el primer problema a resolver en la extracción conocimiento, es la cantidad y calidad de las reglas de asociación que surgen al aplicar los modelos existentes a los conjuntos de datos con presencia de atributos con dominios de un alto nivel de granularidad.

El segundo problema, que plantea en [San99], es que la semántica de las reglas en sí y los parámetros que permiten al usuario interactuar con el proceso de extracción, sabiendo que el ser humano trabaja con descripciones del conocimiento con un menor nivel de granularidad, genera una dificultad de asignar un concepto claro al usuario, esto aumenta cuando se usa como criterio obtener reglas con el mayor soporte y confianza posible, aumentando además la complejidad del mecanismo de extracción.

Al revisar el concepto de reglas de asociación cuantitativas, al intentar resolver este problema de incompatibilidad entre el alto nivel de granularidad y nuestra forma de razonar, genera otros problemas como el del mínimo soporte (minsupp), mínima confianza (minconf), tiempo de ejecución (ExecTime) y aumento del número de reglas (ManyRules), para resolver esta problemática de una forma eficiente e intuitiva se sugiere el uso de las reglas de asociación difusas.

Como conclusiones de la investigación en [San99], se plantea que para resolver las problemáticas que presenta la extracción de conocimiento se debería hacer lo siguiente:

- Usar un nuevo método para definir en conjunto con el experto las etiquetas lingüísticas, obteniendo como resultado una reducción de la granularidad de los dominios de los atributos.

- Utilizar para la generación de reglas de asociación el nuevo concepto de asociación difusa usando en este proceso el soporte y los factores de certeza en lugar de la confianza.

- El uso de factores de certeza que sustituye a la confianza, evita los problemas que representa la confianza sin aumentar la complejidad del proceso de extracción.

- Aplicar el nuevo concepto de regla de asociación muy fuerte, evitando que las reglas de asociación no reflejen la realidad, resultado que se obtiene al tener en cuenta no solamente el soporte de la asociación entre la presencia del antecedente y el consecuente, sino también considerando el soporte de la asociación entre la ausencia del consecuente y la ausencia del antecedente.

Haciendo uso de la metodología recomendada en [San99], se elimina la gran cantidad de reglas falsas, obteniendo reglas de mayor calidad, basadas en casos positivos y negativos, que permitan discriminar si la asociación es correcta o no, y el uso de etiquetas lingüísticas que ayuden a reducir la segmentación del dominio conduce a otorgar reglas más intuitivas por su proximidad a la forma de razonar del experto humano.

Ahora nos centraremos en las reglas de asociación difusas. En estas secciones realizaremos algunos experimentos, enfocados principalmente en la interacción que realizan los usuarios que navegan por el sitio Web, para en próximos capítulos ampliar nuestra investigación enfocados a la identificación de sesiones de usuarios para creación de los perfiles de usuarios.

4.5. Reglas de asociación difusas

Dado un conjunto de ítems *I*, definimos una transacción difusa $\tilde{\tau}$ como un subconjunto difuso no vacío de *I*, donde $\tilde{\tau} \subseteq I$. Para toda $i \in I$, notaremos a $\tilde{\tau}(i)$ el grado de pertenencia de *i* en una transacción difusa $\tilde{\tau}$. Notaremos a $\tilde{\tau}(I_0)$ el grado de inclusión de un itemset en una transacción difusa , definida como [DSVM03]:

$$\tilde{\tau}(I_0) = min_{i \in I_0} \tilde{\tau}(i) \tag{4.1}$$

De acuerdo a la definición, una transacción es un caso especial de transacciones difusas. Podemos ver un ejemplo de un conjunto de transacciones difusas en la tabla 4.1.

	i_1	i_2	i_3	i_4
$\tilde{\tau}_1$	0	0.6	0.7	0.9
$\tilde{\tau}_2$	0	1.0	0	1.0
$\tilde{\tau}_3$	1.0	0.5	0.75	1.0
$\tilde{\tau}_4$	1.0	0	0.1	1.0
$\tilde{\tau}_5$	0.5	1.0	0	1.0
$\tilde{\tau}_6$	1.0	0	0.75	1.0

Tabla 4.1: *Transacciones Difusas*

Las columnas y las filas son descritas por identificadores de ítems y de transacciones, respectivamente. La celda para el ítem i_k y la transacción $\tilde{\tau}_j$, contienen un valor entre [0,1], que es el grado de pertenencia de i_k en $\tilde{\tau}_j$, o también $\tilde{\tau}_j$.

Con un ejemplo podremos ver con mayor atención este punto. Sea $I = \{i_1, i_2, i_3, i_4\}$ un conjunto de ítems, donde en la tabla 4.1 se muestra 6 transacciones definidas en *I*.

Luego,$\tilde{\tau}_1 = (0{,}6 \div i_2 + 0{,}7 \div i_3 + 0{,}9 \div i_2)$; $\tilde{\tau}_2 = (1{,}0 \div i_2 + 1{,}0 \div i_4)$ y así en todas; en particular $\tilde{\tau}_2$ es una transacción crisp,$\tilde{\tau}_2 = (i_2, i_4)$. Algunos grados de inclusión son $\tilde{\tau}_1(\{i_3, i_4\}) = 0{,}7$; $\tilde{\tau}_1(\{i_2, i_3, i_4\}) = 0{,}6$; $\tilde{\tau}_4(\{i_1, i_4\}) = 1{,}0$.

Llamaremos a *T-set* un conjunto de transacciones ordinarias y *FT-set* un conjunto de transacciones difusas, donde el ejemplo anterior muestra el conjunto de transacciones difusas $FT - set = \{\tilde{\tau}_1, \tilde{\tau}_2, ..., \tilde{\tau}_6\}$, el cual contiene a seis transacciones difusas.

Podemos definir entonces a *I* como un conjunto de ítems, *T* como *FT-set* y $A, C \subseteq I$ como dos subconjuntos crisp, con $A, C \neq 0$ y $A \cap C = 0$ Una regla de asociación difusa $A \rightarrow C$ es soportada en *T* si y solo si:

$$\tilde{\tau}(A) \leq \tilde{\tau}(C) \forall \tilde{\tau} \in T \tag{4.2}$$

donde, el grado de inclusión de *C* es más grande que *A* para toda transacción difusa $\tilde{\tau}$.

La definición anterior, preserva el significa de las reglas de asociación, porque si asumimos $A \subseteq \tilde{\tau}$ en algún sentido, deberíamos asumir que $C \subseteq \tilde{\tau}$. Entonces, una transacción es un caso especial de una transacción difusa, entonces una regla de asociación es un caso especial de una regla de asociación difusa.

4.5.1. Soporte, confianza y factor de certeza de reglas de asociación difusas

Utilizaremos un enfoque semántico basado en la evaluación de sentencias cuantificadoras. Una sentencia cuantificadora es una expresión de la forma *Q de F son G*, donde *F y G* son dos subconjuntos difusos de un conjunto finito *X y Q* es un cuantificador relativo difuso. Los cuantificadores relativos son etiquetas lingüísticas para porcentajes difusos que pueden ser representados en un conjunto difuso en [0,1], así como *la mayoría, casi todos o muchos*.

El soporte de I_0 en *T* , donde $I \subseteq I_0$, es la evaluación de la sentencia cuantificadora:

$$Q \quad de \quad F \quad son \quad \tilde{\Gamma}_{I_i} \tag{4.3}$$

donde $\tilde{\Gamma}$ es un conjunto difuso definido como $\tilde{\Gamma}_{I_0}(\tilde{\tau}) = \tilde{\tau}(I_0)$.

El *soporte* de una regla de asociación $A \rightarrow C$ en el conjunto de transacciones difusas *T* es $supp(A \cup C)$, y la evaluación de la sentencia cuantificadora:

$$Q \quad de \quad T \quad son \quad \tilde{\Gamma}_{A \cup C} = Q \quad de \quad T \quad son \quad (\tilde{\Gamma}_A \cap \tilde{\Gamma}_C) \tag{4.4}$$

La *confianza* de una regla de asociación difusa $A \rightarrow C$ en un conjunto de transacciones difusas es la evaluación de la sentencia cuantificadora:

$$Q \quad de \quad \tilde{\Gamma}_A \quad son \quad \tilde{\Gamma}_C \tag{4.5}$$

Comentada estas definiciones que establecen la familia de las medidas de soporte y confianza, dependiendo del método de evaluación y el cuantificador de nuestra elección, evaluaremos las sentencias por medio del método GD [DSV00], que ha sido demostrado que posee buenas características y mejor desempeño que los otros.

La evaluación de $Q \ de \ F \ son \ G$ por medio de *GD* está definido como:

$$GD_Q(\frac{G}{F}) = \sum_{\alpha_i \in \Delta(G/F)} (\alpha_i - \alpha_{i+1}) Q\left(\frac{|(G \cap F)_{\alpha_i}|}{F_{\alpha_i}}\right) \tag{4.6}$$

donde $\Delta(G/F) = \Delta(G \cap F) \cup (F)$,$\wedge(F)$ siendo el conjunto del nivel de *F*, y $\Delta(G/F) = \{\alpha_1, ...\alpha_p\}$ con $\alpha_1 \geq \alpha_{i+1}$ para todo $i \in \{1, ..., p\}$.

La evaluación de una sentencia cuantificada "$Q\ \ de\ \ F\ \ son\ \ G$", por medio del método GD puede ser interpretada como:

- Lo evidente, que el porcentaje de objetos en F que está también en G (cardinal relativo de G con respecto a F) es Q;
- Un cuantificador de agregación dirigido, del cardinal relativo de G con respecto a F por cada corte del mismo nivel de ambos conjuntos.

Por lo tanto, $Supp(A \rightarrow C)$ puede ser interpretado como el porcentaje de transacciones en $\tilde{\Gamma}_{A \cup C}$ es Q, y la $Conf(A \rightarrow C)$ puede verse como el porcentaje de transacciones en $\tilde{\Gamma}_A$ que es también en $\tilde{\Gamma}_C$ es Q. En ambos casos, el cuantificador es parámetro lingüístico que determina el final semántico de las medidas.

Muchos métodos de evaluación y cuantificadores pueden ser elegidos para caracterizar y evaluar el soporte y confianza de reglas de asociación difusas, con tal que las siguientes cuatro características intuitivas de las medidas para las reglas asociaciones ordinarias tenga aplicación:

I. Si $\tilde{\Gamma}_A \subseteq \tilde{\Gamma}_C$, entonces $Conf(A \rightarrow C) = 1$.

II. Si $\tilde{\Gamma}_A \cup \tilde{\Gamma}_C = 0$, entonces $Supp(A \rightarrow C) = 0$ y $Conf(A \rightarrow C) = 0$.

III. Si $\tilde{\Gamma}_A \subseteq \tilde{\Gamma}_{A'}$, (particularmente cuando $A' \subseteq A$), entonces $Conf(A' \rightarrow C) \leq Conf(A \rightarrow C)$.

IV. Si $\tilde{\Gamma}_C \subseteq \tilde{\Gamma}_C$ (particularmente cuando $C' \subseteq C$), entonces $Conf(A \rightarrow C) \leq Conf(A \rightarrow C')$.

Seleccionamos el cuantificador Q_M definido por $Q_M(x) = x$,luego siendo $I_0 \subseteq I$ cuando cada $\tilde{\Gamma}_{I_0}$ es crisp, entonces el $supp(I_0)$ medido por GD con Q_M es el soporte ordinario de un itemset.

GD verifica esto si F y G son crisp, entonces la evaluación de "$Q\ \ de\ \ F\ \ son\ \ G$" es:

$$GD_Q \frac{G}{F} = Q\left(\frac{|F \cap G|}{|F|}\right) \tag{4.7}$$

Por lo tanto,

$$supp(I_0) = GD_{QM}\left(\frac{\tilde{\Gamma_{I_0}}}{T}\right) = \frac{\left|\tilde{\Gamma_{I_0}}\right|}{T} \tag{4.8}$$

Entonces, $A \rightarrow C$ es una regla de asociación ordinaria en T. Luego el $Supp(A \rightarrow C)$ medido por GD con Q_M, es el soporte ordinario de una regla. De las propiedades de GD, tenemos que:

$$Supp(A \rightarrow C) = GD_{Q_M}\left(\frac{\tilde{\Gamma}_{A\cup C}}{T}\right) = \frac{\tilde{\Gamma}_{A\cup C}}{T} = supp(A \cup C) \tag{4.9}$$

Con la confianza sucede algo similar, cuando $A \rightarrow C$ es una regla de asociación ordinaria en T, la $Conf(A \rightarrow C)$ medido por GD con Q_M , es la confianza ordinaria de una regla. De las propiedades de GD, tenemos que:

$$Conf(A \rightarrow C) = GD_{Q_M}\left(\frac{\tilde{\Gamma}_C}{\tilde{\Gamma}_A}\right) = \frac{\left|\tilde{\Gamma}_A \cap \tilde{\Gamma}_C\right|}{\left|\tilde{\Gamma}_A\right|} = \frac{\tilde{\Gamma}_{A\cup C}}{\tilde{\Gamma}_A} = \frac{Supp\,(A \cup C)}{supp(A)} \tag{4.10}$$

A menos que una referencia específica para el cuantificador sea dada, de ahora en adelante consideraremos soporte y confianza basados en Q_M y GD.

En la tabla siguiente se ilustra el soporte y la confianza obtenida de varias reglas de asociación en T_6 [DSVM03] (Ver tabla 4.2).

$$Conf(\{i_1, i_3\}) = GD_{Q_M}\left(\frac{\tilde{\Gamma}_{\{i_4\}}}{\tilde{\Gamma}_{\{i_3,i_4\}}}\right) = 1, \;\; pues \;\; \tilde{\Gamma}_{\{i_1,i_3\}} \subseteq \tilde{\Gamma}_{\{i_4\}} \tag{4.11}$$

Reglas	Soporte	Confianza
$\{i_2\} \rightarrow \{i_3\}$	0.183	0.283
$\{i_1, i_3\} \rightarrow \{i_4\}$	0.266	1.0
$\{i_1, i_4\} \rightarrow \{i_3\}$	0.266	0.441

Tabla 4.2: *Soporte y Confianza de tres Reglas Difusas*

Otra medida interesante y muy ligada a las medidas ya vistas es el *factor de certeza*. Llamaremos *factor de certeza (FC)* de una regla de asociación difusa $A \rightarrow C$ al valor:

$$FC(A \longrightarrow C) = \frac{Conf(A \longrightarrow C) - supp(C)}{1 - supp(C)}, \;\; Si \;\; Conf(A \longrightarrow C) > supp(C) \tag{4.12}$$

Y

$$FC(A \longrightarrow C) = \frac{Conf(A \longrightarrow C) - supp(C)}{supp(C)}, \;\; Si \;\; Conf(A \longrightarrow C) \leq supp(C) \tag{4.13}$$

Asumiendo que si $supp(C) = 1$ entonces $FC(A \longrightarrow C) = 1$ y si el $supp(C) = 0$ entonces $FC(A \longrightarrow C) = -1$.

El *factor de certeza* toma valores entre [1,-1]. Cuando el factor de certeza es positivo indica que la dependencia entre $A \longrightarrow C$ es positiva, si el factor de certeza es igual a 0 quiere decir que son independientes y cuando el factor de certeza es negativo indica que la dependencia entre $A \longrightarrow C$ es negativa. Con esto podemos decir que una regla de asociación es fuerte cuando el factor de certeza y el soporte son más grandes que los umbrales puestos por el usuario como son el *minFC* y el *minSup* respectivamente.

4.6. Medidas de interés

Los procesos de Minería se concentran en el descubrimiento de patrones precisos y comprensibles; mientras que las medidas de interés proporcionan al usuario un grado de confianza de los patrones descubiertos teniendo en cuenta además de la precisión y la comprensión, la novedad [Jim07]. Por ejemplo, podemos hacer referencia a un caso muy conocido en esta área, que es el caso del Wal-Mart; sus datos relacionados con transacciones de compras, encontraron la novedosa regla de asociación entre pañales y cerveza, bastante sorprendente en encontrar, pero aunque un patrón sea inesperado o novedoso, esto no quiere decir que sea válido. Incluso un patrón es inesperado y válido puede pasar que no sea útil.

En la figura 4.1 podemos ver una taxonomía sobre los distintos tipos de medidas de interés [McG05].

A continuación veremos algunas medidas de interés más relevantes propuestas para el caso de las reglas de asociación, basándonos en la Figura 4.1.

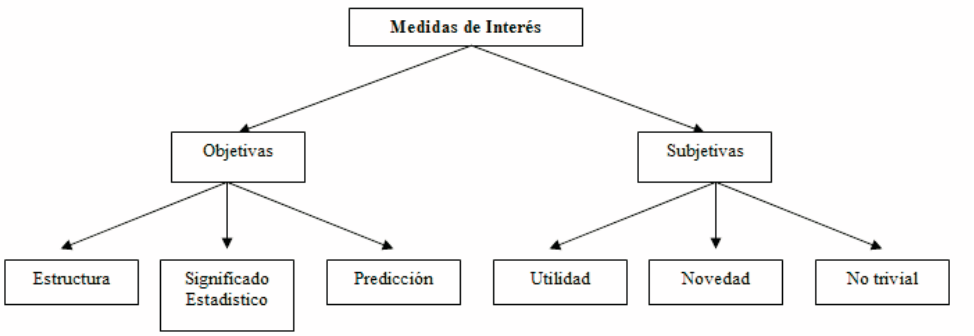

Figura 4.1: *Taxonomía de las Medidas de Interés.*

4.6.1. Medidas de interés objetivas

Muchas de las medidas objetivas están basadas en la estructura que posee la regla de asociación, como frecuencia de aparición de los ítems en las transacciones. También se han utilizado algunas adaptaciones de tests estadísticos como es el test de independencia X^2, el cual según el resultado obtenido nos da un significado estadístico a la regla. Otros de los enfoques seguidos es el de ver cuál es la precisión predictiva de la regla, es decir, si nos sirve para predecir otras reglas, o bien, si con ella se pueden predecir reglas que ocurran en un futuro [Jim07].

Estos aspectos como la estructura, el significado estadístico y la predicción son los aspectos más importantes teniendo en cuenta el cálculo de las medidas de interés objetivas. Además de estos aspectos, existen muchos otros para medir el interés de una regla de asociación de forma objetiva, como son: la fuerza de la regla, su cobertura, su precision, la confianza, etc., lo podemos ver con más detalle en [HH03], [TK00]. En la Tabla 4.3 podemos ver algunas medidas del interés objetivas [McG05].

4.6.2. Medidas de interés subjetivas

Las medidas objetivas no son suficientes para determinar sobre cuales son las reglas más interesantes producidas en el proceso de Minería. Sin embargo, las medidas subjetivas generalmente usan las creencias de los usuarios para ordenar las reglas o patrones descubiertos en el proceso de Minería. Esto es lo que los autores llaman medir lo *inesperado de una regla (unexpected rules).*

Una medida de interés según Frawley [FPSM92] será aquella que sea novedosa, útil

Medida	Descripción	Fórmula
Entropía de Shannon	Medida de la Entropía relativa de teoría calcula la información media contenida	$-\sum_{i=1}^{m} P_i log_2 P_i$
Lorenz	Curva de estadística, calcula las probabi_ lidades asociadas a los datos	$q\sum_{i=1}^{m}(m-i+1)P_i$
Gini	Medida de desigualdad basada en la Curva de Lorenz, usa el radio de la Curva de Lorenz y el área total.	$\frac{q\sum_{i=1}^{m} q\sum_{j=1}^{m}\lvert P_i - Pj\rvert}{2}$
Kullback-Leiber	Usa la medida de la entropía de Shannon y la distancia para calcular la diferencia entre la distribución actual y una distribución uniforme	$log_2 m - \sum_{i=1}^{m} P_i log_2 \frac{P_i}{q}$
Atkinson	Medida de desigualdad de Economía medidas de la distribución de la población.	$max(P_i)$
Lift	También denominado Interés o medida de independencia representa un test para medir la dependencia estadística. Si su resultado es mayor 1, la regla es buena, sino es peor que elegir un resultado aleatorio.	$\frac{P_{ij}}{P_i * P_j}$
Interestingness (P-S)	Es una de las primeras medidas propuestas para medir el interés de las reglas, fue introducida por Piatetsky-Shapiro. La independencia estadística ocurre cuando el valor de esta medida es 0.	$P_{ij} - (P_i * P_j)$

Tabla 4.3: *Medidas de interés objetivas*

y no trivial. La novedad dependerá de un sistema de referencia propuesto, que podemos convenir que sea el conjunto de creencias del usuario. La utilidad de la regla se medirá de acuerdo a un conjunto de objetivos que el usuario quiere conseguir. Y la no trivialidad la podríamos asociar a la parte no subjetiva del interés, es decir que dicha regla no puede obtenerse de otras reglas obtenidas con anterioridad.

Liu [LHCM00], [LHML99] realiza un análisis sobre las medidas subjetivas de interés y consideran que éstas deben tener en cuenta si la regla es *inesperada por el usuario (unexpectedness)* y también si es *útil para el usuario (usefullness)*. Este enfoque lo analizaremos con más profundidad más adelante y lo aplicaremos en algunos experimentos, no hay que olvidar que también son importante las objetivas.

Es difícil medir la utilidad y definir una medida para un patrón o una regla de aso-

ciación. Del punto de vista de la utilidad Sahar en [Sah99],[Sah01],[Sah02] propuso un enfoque para descubrir patrones interesantes eliminando aquellas reglas que parecían similares a otras utilizando el algoritmo Apriori. Este enfoque suprime la dificultad y el coste de tiempo que presenta el obtener el conocimiento del dominio experto y además define un tipo particular de regla interesante.

En la actualidad se utilizan diferentes técnicas para concebir los sistemas de creencias del usuario y esto normalmente implica un ejercicio de adquisición del conocimiento del dominio de los expertos. En los trabajos realizados se han usado principalmente las siguientes técnicas para definir el dominio del conocimiento subjetivo del usuario [Jim07]:

- *Medidas probabilísticas.* Se han utilizado aproximaciones Bayesianas para poder usar probabilidades condicionales.
- *Medidas para la distancia sintáctica.* Este enfoque se basa en la distancia entre las nuevas reglas y el conjunto de creencias. Por ejemplo, si los consecuentes de una regla son los mismos que los esperados por el usuario, pero los antecedentes son muy distintos, entonces esa regla se consideraría interesante.
- *Contradicción lógica.* Usa una medida estadística objetiva para medir si la regla es esperada o no (mediante soporte y confianza por ejemplo) y después se analiza si hay alguna diferencia con los grados esperados por el usuario de dichas medidas

A continuación comentaremos una enfoque basado en la creencia del usuario, que está relacionado con las medidas para la distancia sintáctica [LHCM00].

Este enfoque permite tres grados de precisión para el conocimiento, del cual nosotros nos enfocaremos en el primero para realizar nuestros experimentos. Estos tres grados de precisión del conocimiento los comentaremos a continuación:

- *Impresiones generales (GI)*, estará dado por un conjunto de ítems $g_i\,(\langle S_1 ... S_m \rangle)$ que el usuario cree que están asociados entre sí pero no sabe de qué forma.
- *Conceptos razonablemente precisos (RPC)*, vendrá dado por las reglas que el usuario cree que pondrán cumplirse, $rpc\,(\langle S_1 \wedge ... \wedge S_m \longrightarrow V_i \wedge ... \wedge V_g \rangle)$.
- *Conocimiento totalmente preciso (PK)*, serán reglas que el usuario especifique de forma precisa.

4.6.2.1. Medidas subjetivas mediante el enfoque de impresiones generales

Basándonos en estas premisas y para nuestro caso en particular utilizaremos las *impresiones generales* para así determinar ciertas medidas de interés para poder terminar si una regla es o no interesante.

Entonces, definiremos *U* como el conjunto de creencias del usuario y *A* el conjunto de reglas de asociación obtenidas en el proceso de Minería. Las técnicas que utilizan [LHCM00] en este enfoque para ordenar las reglas depende del concepto de regla interesante que se tenga. A continuación definiremos algunas formulas para determinar este ordenamiento:

- *Reglas conformes.* Una regla $A_i \in A$ es conforme a una regla $U_j \in U$ del conocimiento del usuario si el antecedente y el consecuente de ambas reglas no difieren mucho. Para medir esta conformidad se utiliza la medida $conf m_{ij}$. Si llamamos L_{ij} y R_{ij} (left and right) al grado de compatibilidad entre los antecedentes y consecuentes de las reglas A_i y U_j respectivamente, tendremos que:

$$conf m_{ij} = L_{ij} \cdot R_{ij} \tag{4.14}$$

- *Reglas con consecuente inesperado.* Una regla $A_i \in A$ tiene el consecuente inesperado con respecto a $U_j \in U$, si los antecedentes son similares pero no lo son los consecuentes. Para medirlo se utiliza la siguiente medida.

$$unexpConseq_{ij} = \begin{cases} 0 & \text{Si} \quad L_{ij} - R_{ij} \leq 0 \\ L_{ij} - R_{ij} & \text{Si} \quad L_{ij} - R_{ij} > 0 \end{cases}$$

- *Reglas con antecedentes inesperados.* Es similar al caso anterior pero siendo en este caso los antecedentes muy distintos. Este tipo de reglas son muy útiles para el usuario debido a que el resultado que éste esperaba se consigue con distintos antecedentes de lo esperado. En este caso usamos la medida dada a continuación.

$$unexpCond_{ij} = \begin{cases} 0 & \text{Si} \quad R_{ij} - L_{ij} \leq 0 \\ R_{ij} - L_{ij} & \text{Si} \quad R_{ij} - L_{ij} > 0 \end{cases}$$

- *Ambos lados de la regla inesperados.* Una regla $A_i \in A$ tiene ambos lados inesperados con respecto a $U_j \in U$ si los antecedentes y los consecuentes son inesperados. Estas reglas dirán al usuario que hay asociaciones que él no tuvo en cuenta o no conocía al especificar el conjunto de creencias inicial. Para medirlo usaremos la siguiente fórmula.

$$bsUnexp_{ij} = 1 - max(conf_{ij}, unexpConseq_{ij}, unexpCond_{ij}) \quad (4.15)$$

Todos estos valores se encuentran entre 0 y 1, donde 1 representa muy similar al conocimiento del usuario, y 0 indica que no hay ninguna coincidencia con las expectativas del usuario.

4.7. Uso de reglas de asociación difusas en la minería web de uso

Ya que hemos realizado un análisis de las reglas de asociación difusas, dentro de las cuales hemos discutido sobre las diferentes medidas que se utilizan. A continuación realizaremos una descripción general de los aspectos más importantes de la extracción de reglas de asociación difusas sobre los archivos Web log.

A continuación plantearemos el modelo con el cuál realizaremos los diferentes experimentos para obtener reglas que nos permitan saber el comportamiento o patrones de navegación del usuario sobre algún sitio Web.

4.7.1. Modelo de datos

Antes de comenzar a definir el modelo de datos que utilizaremos es importante comentar la etapa del preprocesamiento de los datos y así tener una idea más clara de los conjuntos de datos que analizaremos. Es de suponer que la preparación de los datos puede genera un conjunto más pequeño que el original, y de esta manera mejorar la eficiencia del proceso de Minería.

Dentro de la etapa del preprocesamiento de los datos se realizan diversos procesos de limpieza, como eliminar los datos irrelevantes, limpiar el ruido de los datos y datos inconsistentes de los archivos. Todos estos procesos se deberían realizar si tenemos un archivo en ”bruto”, o sea que el archivo se ha obtenido directamente del servidor Web y

en el caso contrario, se debería tener alguna certeza previa de que el archivo se encuentra listo para su análisis.

Al mismo tiempo que se realiza la etapa de preprocesamiento se puede realizar la identificación de las entradas o transacciones que va dejando el usuario durante su navegación. Es importante determinar las entradas de los usuarios para poder aplicar, en este caso, la técnica de reglas de asociación difusas para la búsqueda de patrones.

- **Ítems:** cuando nos referimos a los elementos estamos haciendo referencia a todos los campos que pueden componer un archivo Web log, estos campos dependerán del tipo de formato que tenga este.

  Dentro de los diferentes items podemos mencionar a la *IP*, *host*, *fecha/hora*, *páginas visitadas*, *páginas referenciada*, entre otras. En la figura 4.2, se muestra unas líneas del tipo de archivo log llamado CSV (Comman Serapated Value), el cuál está compuesto por 6 campos o elementos (identificador de compra, fecha, IP, Sesión, Página visitada y Página referenciada).

```
14, 23/Jun/2006:11:10:09+0200, 201.979.148.252, 6a1b42bg4234qj1323s, GET/ugr.es, www.utem.cl
12, 23/Jun/2006:22:13:10+0200, 203.349.118.762, 76a1b42bg4234qj1323r, GET/etsiit.ugr.es, www.utem.cl
14, 23/Jun/2006:07:12:44+0200, 201.659.128.532, 5ahytrwbcy65qh3qji9oiy, GET/css/estilo.css, www.profesores/jmaroza
13, 23/Jun/2006:08:56:23+0200, 203.129.258.112, 3ehndye64ye6qhbi999w, GET/shop2.com, www.canalplus.es
14, 23/Jun/2006:21:23:15+0200, 201.349.118.562, 2v4u66b3tw5hdi809yt, GET/shop3.com, www.ya.com
15, 23/Jun/2006:03:31:13+0200, 206.239.818.242, 6a1b42bg4234qj1323o, GET/shop1.es, www.shop4.cpm
```

Figura 4.2: *Líneas de archivo CSV*

  Para la extracción de las reglas de asociación difusas el usuario puede determinar cuales son los ítems de las transacciones que más le interesa y así poder encontrar reglas relacionadas solamente con esos ítems. Por ejemplo, el usuario podría elegir los ítems de las IPs y el ítem de las páginas visitadas, para así saber cuales serán las páginas que se conectan de determinadas IP o quizás podría seleccionar los ítems de la fecha y hora con el ítem de las páginas visitadas para saber cuál es el comportamiento de conexión del usuario a determinadas páginas.

  En general, podemos decir que la extracción de las reglas dependerán del objetivo del análisis, o sea de lo que se desea encontrar.

- **Transacciones:** explicaremos las posibles tablas transaccionales que utilizaremos para la obtención de diferentes tipos de reglas. Para explicar esto lo hemos separado en dos casos. El *caso A* está relacionado con los ítems de las páginas visitadas con las páginas referenciadas y el *caso B* esta relacionado con los elementos de la fecha/hora y las páginas visitadas.

- *Caso A:* en este caso veremos dos ítems las cuales son las páginas visitadas y referenciadas. Estos elementos están relacionadas con las IP's dentro de un conjunto de datos. Es importante destacar que se pueden utilizar diversos criterios para determinar el peso que se le asignará a las páginas durante su navegación identificada con ciertas IP's. Por ejemplo, se puede determinar a través de el periodo de tiempo de su conexión ó a través de la frecuencia de la página en el archivo log o mejor, la frecuencia de la página relacionada con alguna IP's en particular, entre otros criterios.

  Ahora explicaremos como obtener los diferentes pesos de las páginas Web y utilizaremos el criterio en donde las páginas están relacionadas con las IP's dentro de los archivos log. Por ejemplo algunas IP pueden ser: *74.6.68.215*, *217.216.61.116*, *217.216.61.116*, *117.134.41.234* y algunas páginas tanto visitadas como referenciadas pueden ser:
  GET/apps/foro/index.phpHTTP/1.0V,
  GET/apps/tablon/HTTP/1.1V,
  http://etsiit.ugr.es/apps/foro/index.php?action=hebra&idhebra=1939R,
  http://etsiit.ugr.es/apps/foro/index.phpR
  Es importante decir que hemos querido diferenciar las páginas visitadas de las referenciadas marcándolas con una letra al final de cada. La letra *V* nos representara una página visitada y la *R* una referenciada.

  En la tabla 4.4 podemos ver un ejemplo práctico para obtener los pesos de las páginas. En cada celta de la tabla podemos ver las veces que una página Web coincide en una determinada IP's.

Páginas / **IP**	Pag_1	Pag_2	Pag_3	Pag_4
IP_1	0	4	0	7
IP_2	7	0	8	0
IP_3	6	0	2	0
IP_4	0	3	0	10

Tabla 4.4: *Frecuencia de las páginas en una determinada IP*

A partir de esta tabla podremos obtener los diferentes pesos de las páginas en el archivo log por ejemplo: para la Pag_1 en la IP_3 sería 0.6 ya que dividimos el número de veces de la Pag_1 en la IP_3 por el valor máximo de la frecuencia de las páginas, que para este caso es 10, así podremos obtener valores entre [0,1].

En la tabla 4.5 veremos las transacciones difusas relacionadas con los elementos que hemos mencionado para este caso, y de esta manera poder determinar las reglas.

Páginas / **IP**	Pag_1V	Pag_2V	Pag_3R	Pag_4R
IP_1	0	0.4	0	0.7
IP_2	0.7	0	0.8	0
IP_3	0.6	0	0.2	0
IP_4	0	0.3	0	0.8

Tabla 4.5: *Transacciones Difusas Caso A*

Para distinguir de las páginas visitadas de las páginas referenciadas las hemos marcado con una *V* a las visitadas y con una *R* a las referencidas.

Veremos un pequeño ejemplo relacionado con la tabla de las transacciones difusas que hemos realizado.

En la tabla 4.5 podemos ver cuatro transacciones difusas para ejemplificar el proceso de búsqueda. Las columnas y filas son descritas por identificadores de ítem, para este caso las hemos descrito como Pag_n y las transacciones por IP_k, y el cruce entre ambas contienen valores entre [0,1], que es el grado de pertenencia de Pag_n en IP_k.

Sea $IP = \{Pag_1V, Pag_2V, Pag_3R, Pag_4R\}$ el conjunto de ítems. Luego, $IP_1 = \{0{,}4/\ Pag_2V + 0{,}7/Pag_4R\}$; $IP_2 = \{0{,}7/Pag_1V + 0{,}8\ \ /Pag_3R\}$. Algunos grados de inclusión son:

$IP_1\ (\{Pag_2V, Pag_4R\}) = 0{,}4$; $IP_2\ (\{Pag_1V,\ \ Pag_3R\}) = 0{,}2$.

Teniendo esta información podremos obtener las diferentes medidas como son el soporte, la confianza y el factor de certeza, y así determinar las reglas más interesante para este caso.

- *Caso B:* en este caso veremos los elementos fecha/hora y páginas visitadas. Estos elementos están relacionadas con las IP's dentro del archivo log.

  De la misma manera que el caso anterior el peso para el ítem de las páginas estará determinado por la frecuencia en una determinada IP y para el ítem de la fecha/hora la obtendremos con los periodos de tiempo de conexión determinado por las etiquetas lingüísticas que podemos ver figura 4.3. A continuación veremos un ejemplo práctico para determinar el peso del ítems fecha/hora (Ver tabla 4.6), y no lo haremos para el ítems de página ya que es el mismo que hemos explicado para el caso A (Ver tabla 4.4).

Fecha/Hora	Peso	Etiqueta
08:30	1.0	Mañana
12:45	0.5	Medio Día
15:25	0.4	Tarde
20:20	0.3	Noche

Tabla 4.6: *Periodos de tiempo y peso para el item fecha/hora*

Como el campo fecha/hora esa dentro de un rango de valores, y estos valores pueden ser representados de una forma difusa a través de las etiquetas, utilizaremos este enfoque para determinar el peso de este ítems en el archivo log (Ver anexo A).

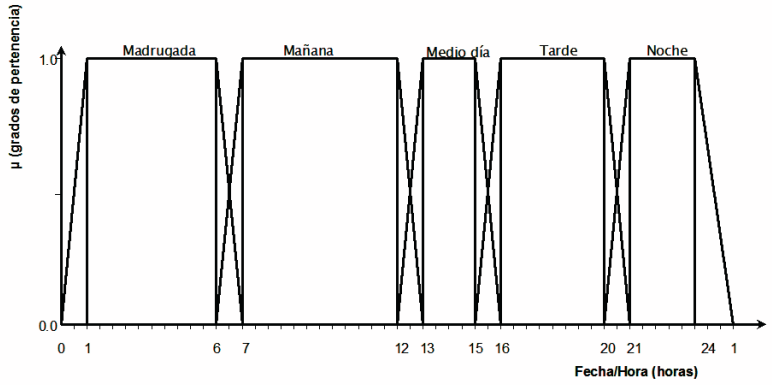

Figura 4.3: *Etiquetas lingüísticas del campo Fecha/Hora*

Páginas / IP	$hora/fecha$	Pag_1V	Pag_2V	Pag_3V
IP_1	1.0	0.4	0	0.8
IP_2	0.5	0	0	0.4
IP_3	0.4	0.7	0.3	0
IP_4	0.3	0.2	0	0

Tabla 4.7: *Transacciones Difusas Caso B*

En la tabla 4.7 podemos ver cuatro transacciones difusas para ejemplificar el proceso de búsqueda. Los identificadores de ítem los hemos descritos, para este caso, como Pag_n relacionadas a las páginas y fecha/hora para la hora de conexión, y las transacciones por IP_k, y el cruce entre ellas contienen valores entre [0,1], que es el

grado de pertenencia de Pag_n en IP_k.

Sea $IP = \{fecha/hora, Pag_1V, Pag_2V, Pag_3V\}$ el conjunto de ítems. Luego,

$IP_1 = \{1{,}0/\text{Mañana}+0{,}4/Pag_1V + 0{,}8/Pag_3V\}$;

$IP_2 = \{0{,}4/Tarde + 0{,}3/Pag_3V\}$.

Algunos grados de inclusión son:

$IP_1\left(\{hora/fecha, Pag_3V\}\right) = 0{,}8$;

$IP_2\left(\{hora/fecha, Pag_3V\}\right) = 0{,}3$.

A partir de esta información, al igual que en el anterior caso, se puede obtener las diferentes medidas para obtener las reglas.

A continuación mostraremos algunos diagramas que permitan visualizar de mejor manera los procesos que se realizan para la extracción de las reglas.

4.7.2. Modelo para la obtención de reglas de asociación difusas

En esta sección veremos las distintas etapas para la obtención de las reglas de asociación difusas. Las etapas las describiremos a través de diferentes diagramas para ver mejor este proceso.

La figura 4.4 vemos un diagrama, que nos muestra el proceso de inicio para la búsqueda de reglas, partiendo de la selección del archivo o el conjunto de archivos a analizar. El archivo puede ser visualizado en sus primeras líneas (Ver figura 4.2), y así determinar si necesita o no ser preparado para el análisis. Si el archivo está preparado para su análisis se selecciona el nivel de regla para posteriormente configurarla.

Una vez que tenemos el conjunto de datos para el análisis, se puede decidir qué tipo de información se puede obtener, dependiendo de que campos sean elegidos, para eso debemos configurar la reglas. Por ejemplo, si el usuario elige los campos de fecha y páginas visitadas, se puede hacer una idea de qué páginas han sido más visitadas a ciertos horarios, o también si el usuario elige los campos de IP y campo de página visitada, de alguna manera podría identificar a los usuarios que visitan ciertas páginas en especial y/o si el usuario elige los campos de página visitada con la página referenciada, podría saber cual es tipo de navegación más habitual y por donde navega el usuario con mayor frecuencia (Ver figura 4.5).

Una vez configurada la regla y comprobada que la configuración se haya hecho de forma correcta, pasaremos a la etapa del proceso de búsqueda de reglas de asociación

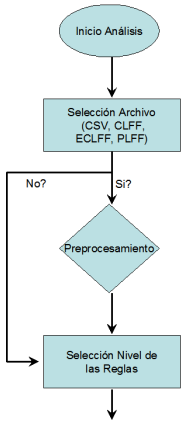

Figura 4.4: *Diagrama inicial para el proceso de búsqueda de reglas*

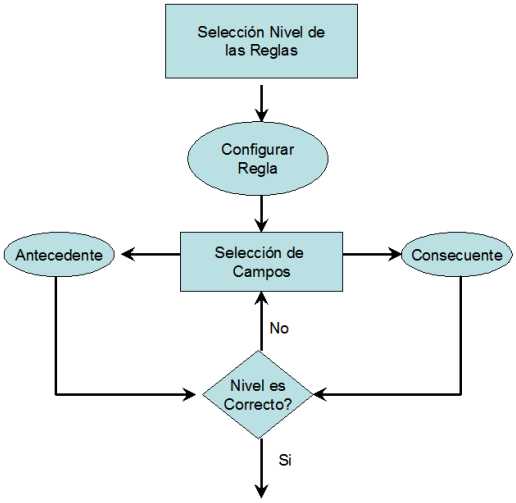

Figura 4.5: *Diagrama para la configuración de reglas*

difusas. En la figura 4.6 se muestra el diagrama, en donde el usuario debe determinar los valores de medidas de mínimo soporte, mínima confianza y mínimo factor de certeza para buscar las reglas.

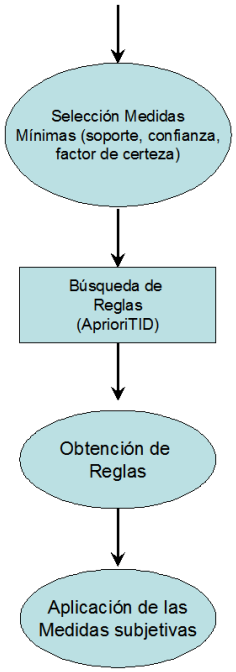

Figura 4.6: *Diagrama para la búsqueda de reglas*

Una vez obtenidas las transacciones y determinadas las medidas mínimas de corte para la búsqueda de las reglas, utilizamos el algoritmo APrioriTID [AR94], para encontrar las reglas de asociaciones, el cuál ataca el problema reduciendo el número de conjuntos considerados.

4.7.3. Ejemplo sencillo de extracción de reglas de asociación difusas de uso: Interpretación de las reglas y obtención de medidas subjetivas

Presentamos un par de ejemplos de posibles resultados que podemos obtener, en este caso, en la configuración de la regla se ha seleccionado en primer lugar los campos de Fecha/Hora con el campo de páginas visitadas por el usuario, esto para saber en que horario son visitadas estas páginas o simplemente para saber el hábito de conexión del usuario.

Ahora veremos diferentes casos de posibles reglas y sus interpretaciones:

- **fecha/hora** $\longrightarrow$ **Página visitada**
  - Mañana $\Longrightarrow$ http://www.shop2.cz/ls/index.php?&id=98&filtr=102
    - Soporte = 0.6
    - Confianza = 1.0
    - FC = 1.0

Interpretación: del conjunto analizado el 60 % presentaba esta regla, la cual nos indica que los usuarios se conectan por la mañana a esa página. Esta regla la podremos utilizar para realizar algún proceso de marketing que dependa exclusivamente de las cosas o temas que existen en esa página y también para entrega información de mejor calidad.

- **Página visitada** $\longrightarrow$ **fecha/hora**
  - http://www.shop2.cz/ls/index.php?&id=98&filtr=102 $\Longrightarrow$ Mañana
    - Soporte = 0.2
    - Confianza = 1.0
    - FC = 1.0

Interpretación: del conjunto analizado sólo el 20 % presenta este tipo de regla, la cual nos dice que ha esta página ha sido visitada por la mañana. Esto puede ser útil para entregar al usuario información más adecuada sobre esa páginas y también nos permitiría identificar quizás a los usuarios que se conectan a esas página por la mañana.

En segundo lugar, hemos seleccionado los campos de páginas visitadas por el usuario con el campo de páginas referenciadas por el usuario, y esto para saber la página que visitan los usuarios a partir de una página inicial visitada. Algunas reglas obtenidas fueron:

- **Página visitada⟶ Página referenciada**
  - /dt/?c=11670 $\Longrightarrow$ http://www.shop2.cz/ls/index.php?&id=98&filtr=102
    - Soporte = 0.6
    - Confianza = 1.0
    - FC = 1.0

Interpretación: esto indica que los usuarios vistan a la página /dt/?c=11670 y luego se van a la página http://www.shop2.cz/ls/index.php?&id=98&filtr=102, esta regla se encuentra en un 60 % dentro del conjunto analizado. Esta regla nos muestra ciertos patrones de navegación del usuario, con el cual podemos realizar mejoramientos en la estructura del sitio para que el usuario pueda obtener una información mucho más interesante en esas páginas. También las podemos utilizar para la identificación de ciertos grupos de usuarios que presenten las mismas características de navegación.

- **Página referenciada⟶ Página visitada**
  - http://www.shop7.cz/akce/?kat=239 $\Longrightarrow$ /dt/?c=123
    - Soporte = 0.2
    - Confianza = 1.0
    - FC = 1.0

Interpretación: esta regla nos dice que dentro del conjunto analizado el 20 % de los usuarios que se han ido a esta página http://www.shop7.cz/akce/?kat=239 antes han visitado la página /dt/?c=123. Esto nos hace pensar en que relación existe entre ambas páginas y así poder entregar una mejor información dependiendo de que es lo que más le interesa al usuario y también podemos realizar algún proceso de marketing que este relacionada con ambas páginas.

4.7.4. Aplicación de las medidas subjetivas

Comenzaremos explicando con un breve ejemplo los resultados obtenidos con las medidas subjetivas. Para las medidas subjetivas se necesita obtener un conocimiento apriori de las creencias de los usuarios. Este conjunto de creencias de las posibles páginas que están relacionados, según el usuario son las siguientes :{*foro, tablón de anuncio, actividades, asignaturas, programación, eventos, página principal*}.

Para ello hemos realizado una pequeña encuesta, con el fin de poder adquirir las creencias del usuario sobre la navegación que ocurre dentro del sitio de la Escuela de Informática y Telecomunicaciones de la Universidad de Granada. Esta encuesta fue realizada a 20 usuarios que habitualmente visitan el sitio de la Escuela.

Dentro de la encuesta hemos preguntado por las secciones del sitio que el usuario cree que son más utilizadas. Según la creencia del usuario hemos obtenido los siguientes resultados: *Foro 40 %*, *Tablón de anuncios 20 %*, *Asignaturas 10 %*, *Eventos 10 %*, *Programación 10 %*, *Servicios 5 %*, *Otras páginas 5 %*.

Y por último le hemos preguntado qué páginas del sitio cree que están relacionadas con más frecuencia. Esta última respuesta la utilizaremos para obtener los valores de las medidas subjetivas en las reglas de asociación difusa. Según el usuario estas serían las relaciones entre las páginas:

- foro $\longrightarrow$ asignatura
- tablón de anuncio $\longrightarrow$ actividades
- evento $\longrightarrow$ página principal
- programación $\longrightarrow$ página principal

Estos serán los conjuntos que utilizaremos a lo largo de todos los demás experimentos relacionados con las reglas de asociación al momento de obtener las medidas de interés subjetivas.

4.7.5. Obtención de las creencias

A través de la encuesta que hemos comentado anteriormente, podemos obtener las medidas subjetivas a través de las creencias del usuario. Por ejemplo, en el análisis del conjunto de datos de la Universidad de Granada hemos comparado para el caso de las reglas de asociación difusas con medidas subjetivas el conjunto de reglas que el usuario cree que se obtendrán en el análisis con las reglas obtenidas. O sea, si comparamos (Ver tablas 4.8 y tabla 4.9):

Analizando las tablas 4.8 y 4.9, se pueden ver los resultados obtenidos en la tabla 4.10 con las medidas subjetivas. Podemos decir que los consecuentes de las reglas son las más inesperados de lo que el usuario pensaba. Son estas creencias las que utilizaremos

	Reglas Obtenidas
regla 1	GET/web/tablon_anuncios/HTTP/1.1 → etsiit.ugr.es
regla 2	GET/wwwforum/HTTP/1.1 → etsiit.ugr.es

Tabla 4.8: *Reglas Obtenidas durante el proceso de análisis*

Creencias del usuario
GET/web/tablon_anuncios/HTTP/1.1 → www-etsi2.ugr.es/actividades.html
GET/evento/ HTTP/1.1 → www-etsi2.ugr.es
GET/programacion/ HTTP/1.1 → www-etsi2.ugr.es
GET/wwwforum/HTTP/1.1 → www-etsi2.ugr.es/asignaturas.html

Tabla 4.9: *Supuestas reglas según las creencias del usuario*

ReglasObtenidas	**confm (Cf)**	**unexpConseq (Ucq)**	**unexpCond (UCd)**	**bsUnexp (bs)**
regla 1	0.285	0.714	0.00	0.285
regla 2	0.285	0.714	0.00	0.285

Tabla 4.10: *Resultados con medidas de interés subjetivas*

para realizar los diferentes experimentos ligados a las reglas de asociación difusas para las medidas subjetivas.

4.8. Obtención de reglas de asociación difusas a partir de archivos log: Caso real

Realizaremos diferentes análisis sobre varios conjuntos de datos. El tipo de archivo log utilizado es el llamado *Extended Common Log File Format (ECLFF)* (ver sección 3.4.1) obtenido del servidor web de la E.T.S. de Ingenierías Informática y de Telecomunicación de la Universidad de Granada. En los experimentos que aquí se incluyen, los conjuntos de datos 1, 2, 3 y 4 se obtuvieron del sitio web antiguo con dirección http://www-etsi2.ugr.es, mientras que el conjunto de datos 5 se obtuvo con el sitio web nuevo con dirección http://etsiit.ugr.es. No obstante, cabe resaltar que la estructura en sí del sitio apenas se vio afectada, por lo que a nivel de interpretación, hemos considerado de igual manera las reglas extraídas de un sitio u otro, incluidas las obtenidas en la encuesta para la construcción de las medidas subjetivas.

4.8.1. Características generales del experimento

- **Objetivo general:** de este conjunto de datos queremos descubrir patrones de navegación del usuario y así tener una mejor descripción de su comportamiento, para ellos utilizaremos las reglas de asociación difusas para extraer reglas de diversos conjuntos de datos.

- **Conjunto de datos:** a continuación describiremos la experimentación realizada sobre diferentes conjuntos de datos para la obtención de los patrones de navegación del usuario. En la siguiente tabla veremos un resumen de los diferentes conjuntos de datos con los cuales trabajaremos y sus respectivas descripciones (Ver tabla 4.11).

Conjuntos De Datos	**N° Transacciones Originales**	**N° Transacciones Objetivas**	**Preprocesamiento**
Conjunto 1	100900	100810	Eliminación transacciones idénticas
Conjunto 2	100810	46950	Eliminación transacciones sin el campo de referencia
Conjunto 3	46950	16518	Eliminación de imágenes
Conjunto 4	16518	12910	Eliminación de javascript
Conjunto 5	98202	15676	Preprocesamiento Completo

Tabla 4.11: *Resumen de los Conjuntos de Datos*

- **Modelo:** el modelo que utilizaremos se basa en la obtención de transacciones difusas que hemos planteado para el caso de las páginas visitadas y páginas referenciadas (Ver sección 4.7.1). Para ello $IP = \{Pag_1, Pag_2, ..., Pag_n\}$ es el conjunto de ítems. Donde los identificadores de los ítems son Pag_n que son las páginas visitadas como referenciadas y las transacciones están identificadas por IP_k, que serán las IP's con la cual están relacionadas las páginas, y el cruce entre ambas contienen valores entre [0,1], que es el grado de pertenencia de Pag_n en IP_k (Ver tabla 4.4), este grado corresponde a la frecuencia de la página dentro de una determinada IP, que posteriormente se ha normalizado para obtener los valores entre [0,1].

- **Técnica utilizada:** para la obtención de las reglas de asociación difusas utilizaremos el algoritmo *AprioriTID*.

- **Medidas:** las medidas objetivas utilizadas son el Soporte (Sup), la Confianza (Con), Factor de Certeza (FC), el Interés o Lift (Lif), PiatesKy-Shapiro (P-S) y de las medidas subjetivas utilizaremos el enfoque de impresiones generales para obtener

las medias, las cuales son: confm (Cf), unexpConseq (UCq), unexpCond (UCd) y bsUnexp (bs).

4.8.2. Resultados con el conjunto 1

- **Conjunto 1:** El primer conjunto de datos que hemos analizado está formado por 100810 entradas o transacciones (Ver tabla 4.11). La información analizada corresponde a tres días dentro del mismo horario, para así poder encontrar algún comportamiento del usuario.

- **Resultados conjunto 1:** Luego de haber hecho un análisis se han obtenido las siguientes reglas de asociación difusas(Ver Tabla 4.12 y 4.13).

N° Regla	Reglas Obtenidas
Regla 1	GET/web/tablon_anuncios/ $\longrightarrow$ www-etsi2.ugr.es
Regla 2	GET/wwwforum/ $\longrightarrow$ www-etsi2.ugr.es

Tabla 4.12: *Resultados reglas obtenidas: Conjunto 1*

N° Regla	Sup	Con	FC	Lif	P-S	cf	Ucq	UCd	bs
Regla 1	0.01	0.66	0.60	3.76	0.007	0.285	0.714	0.00	0.285
Regla 2	0.017	0.66	0.50	4.012	0.005	0.285	0.714	0.00	0.285

Tabla 4.13: *Medidas para las reglas del Conjunto 1*

Los resultados obtenidos indican que el 1 % es representado por la regla 1 y sólo en un 1,7 % la regla 2. Estas reglas representan un el comportamiento de navegación del usuario, que va desde el tablón de anuncios o desde el foro hasta la página principal de la Web de la escuela. Un factor importante es que los valores de confianza están alrededor del 70 % y además el factor de certeza de ambas reglas está sobre el 50 %, lo cual nos indica que ambas reglas son fuertes, pero poco significativas en relación al valor de Soporte de las reglas. También podemos decir la regla más interesante para este caso sería la que posee mayor valor en la medida de lift y más cercano a cero para la medida P-S, o sea *GET/wwwforum* $\longrightarrow$ *www-etsi2.ugr.es* sería la más interesante entre las dos.

Ambas reglas poseen los mismos índices de conformidad, que es relativamente baja con respecto a lo que esperaba el usuario. Esto quiere decir que el usuario esperaba obtener mucho más reglas con respecto a lo que hemos obtenido, también estos

valores nos indica que el usuario ha coincidido en los antecedentes de las reglas que hemos obtenido, pero no es el caso de las consecuentes, que son muy diferentes a los esperados.

Al archivo antes analizado le hemos eliminado las transacciones que no tengan el campo de la página de referencia del archivo anterior y así tener un conjunto más objetivo para el análisis (Ver tabla 4.11).

4.8.3. Resultados con el conjunto 2

- **Conjunto 2:** conjunto de datos de 46950 entradas o transacciones (Ver tabla 4.11).
- **Resultados conjunto 2:** se han obtenido las siguientes reglas de asociación difusas, las cuales se pueden ver en la tabla 4.14 y 4.15.

N° Reglas	Reglas Obtenidas
Regla 1	GET/web/tablon_anuncios/ ⟶ www-etsi2.ugr.es
Regla 2	GET/wwwforum/⟶ www-etsi2.ugr.es

Tabla 4.14: *Resultados reglas obtenidas: Conjunto 2*

N° Reglas	Sup	Con	FC	Lif	P-S	cf	Ucq	UCd	bs
Regla 1	0.022	0.67	0.60	1.80	0.009	0.285	0.714	0.00	0.285
Regla 2	0.037	0.66	0.46	2.176	0.006	0.285	0.714	0.00	0.285

Tabla 4.15: *Medidas para las reglas del Conjunto 2*

En este análisis hemos obtenido las mismas reglas que en el experimento anterior pero con un mayor valor de soporte, en la primera regla hemos obtenido un 2,2 % del conjunto de datos y de la segunda regla un 3,7 %. También podemos decir la regla más interesante sigue siendo la misma que en el caso anterior, o sea *GET/wwwforum ⟶ www-etsi2.ugr.es*, ya que posee mayor valor de interés o lift y el menor en la medida P-S. Con respecto a las medidas subjetivas, sucede lo mismo que en el análisis anterior, ya que no se ha encontrado nuevas reglas en este conjunto de datos.

Para poder obtener mejores resultados en la búsqueda de reglas hemos eliminado todas las transacciones del conjunto de datos anterior, donde las páginas visitadas por el usuario estuviesen relacionada con alguna imagen (*.jpg, *.gif, *.png, *.bmp, entre otras).

4.8.4. Resultados con el conjunto 3

- **Conjunto 3:** conjunto de datos de 16518 transacciones (Ver tabla 4.11).

- **Resultados conjunto 3:** hemos obtenido las siguientes reglas de asociación difusas(Ver tabla 4.16 y 4.17)

N° Reglas	Reglas Obtenidas
Regla 1	GET/js/gamelib_core.js ⟶ www-etsi2.ugr.es
Regla 2	GET/css/estilo.css/ ⟶ www-etsi2.ugr.es
Regla 3	GET/js/gamelib_mouse.js ⟶ www-etsi2.ugr.es
Regla 4	GET/js/marquee.js ⟶ www-etsi2.ugr.es
Regla 5	GET/js/mo_popup2.js ⟶ www-etsi2.ugr.es
Regla 6	GET/web/tablon_anuncios ⟶ www-etsi2.ugr.es
Regla 7	GET/wwwforum ⟶ www-etsi2.ugr.es

Tabla 4.16: *Resultados reglas obtenidas: Conjunto 3*

N° Reglas	Sup	Con	FC	Lif	P-S	cf	Ucq	UCd	bs
Regla 1	0.016	0.50	0.24	1.48	0.005	0.00	0.00	0.142	0.857
Regla 2	0.015	0.48	0.23	1.47	0.005	0.00	0.00	0.142	0.857
Regla 3	0.016	0.49	0.4	1.48	0.005	0.00	0.00	0.142	0.857
Regla 4	0.017	0.86	0.80	2.60	0.01	0.00	0.00	0.142	0.857
Regla 5	0.016	0.50	0.23	1.47	0.005	0.00	0.00	0.142	0.857
Regla 6	0.063	0.69	0.54	2.08	0.032	0.142	0.857	0.00	0.142
Regla 7	0.011	0.67	0.50	2.70	0.023	0.142	0.857	0.00	0.142

Tabla 4.17: *Medidas para las reglas del Conjunto 3*

De estas 7 reglas que hemos obtenidos, siguen siendo las más significantes las reglas 6 y 7 que son las mismas que ya hemos obtenido anteriormente, con un 6,3 % y un 11 % respectivamente. De las otras reglas podemos decir que son poco significativos y son páginas relacionadas con javascript, los cuales eliminaremos para poder tener un conjunto mucho más objetivo y representativo para nuestro análisis. En este análisis la única regla que podemos decir que es diferente a los demás análisis y que podemos decir que es más interesante que las demás, ya que no está dentro del conjunto de páginas del usuario es *GET/css/estilo.css ⟶ www-etsi2.ugr.es*, esta regla quizás no nos entregue demasiada información con respecto al usuario, ya que el antecedente corresponde a una página de estilo simplemente.

En este nuevo análisis hemos eliminados las páginas que poseen extensiones de *.js (javascript) así obtener un conjunto más representativo.

4.8.5. Resultados con el conjunto 4

- **Conjunto 4:** el conjunto de análisis que hemos obtenido después del preprocesamiento es de 12910 transacciones (Ver tabla 4.11).
- **Resultados conjunto 4:** A continuación del análisis de este conjunto hemos obtenido las siguientes reglas de asociación difusas las cuales se pueden ver en la tabla 4.18 y 4.19.

N° Reglas	Reglas Obtenidas
Regla 1	GET/css/estilo.css/⟶ www-etsi2.ugr.es
Regla 2	GET/web/tablon_anuncios/⟶ www-etsi2.ugr.es
Regla 3	GET/wwwforum/⟶ www-etsi2.ugr.es

Tabla 4.18: *Resultados reglas obtenidas: Conjunto 4*

N° Reglas	Sup	Con	FC	Lif	P-S	cf	Ucq	UCd	bs
Regla 1	0.02	0.48	0.23	1.45	0.006	0.00	0.00	0.142	0.857
Regla 2	0.08	0.69	0.54	2.06	0.04	0.142	0.857	0.00	0.142
Regla 3	0.13	0.66	0.50	0.24	0.03	0.142	0.857	0.00	0.142

Tabla 4.19: *Medidas para las reglas del Conjunto 4*

Analizando los resultados de este conjunto de datos, que es mucho más objetivo que los conjuntos anteriores de datos, podemos decir que las reglas 2 y 3 son las más representativa con un 8,1 % y un 13 % respectivamente. Con lo cual podemos decir, que el usuario después de visitar las páginas del foro o el tablón de anuncio regresa a la página principal o ”Home”de la escuela, siendo este el comportamiento más común dentro de este conjunto de datos. Siguen siendo estas dos reglas las más interesantes a lo largo de nuestros análisis, tanto del punto de las medidas objetivas como de las subjetivas.

4.8.6. Resultados con el conjunto 5

- **Conjunto 5:** El conjunto inicial era de 98202 transacciones, luego de realizar el proceso de preprocesamiento hemos obtenido un conjunto objetivo para el análisis

de 15676 transacciones (Ver tabla 4.11).

- **Resultados conjunto 5:** De este conjunto hemos obtenido las siguientes reglas de asociación difusas, las cuales las podemos ver en la tabla 4.20 y 4.21.

N° Regla	Reglas Obtenidas
Regla 1	GET/apps/tablon/ ⟶ http://etsiit.ugr.es/
Regla 2	GET/apps/foro/index.php ⟶ http://etsiit.ugr.es/
Regla 3	GET/apps/foro/index.php ⟶ http://etsiit.ugr.es/apps/foro/index.php?idforo=general
Regla 4	GET/apps/foro/index.php?idforo=asignaturas ⟶ http://etsiit.ugr.es/apps/foro/index.php
Regla 5	GET/apps/foro/index.php?action=foro&idforo=escuela ⟶ http://etsiit.ugr.es/apps/foro/index.php
Regla 6	GET/apps/foro/index.php?idforo=general ⟶ http://etsiit.ugr.es/apps/foro/index.php

Tabla 4.20: *Resultados reglas obtenidas: Conjunto 5*

N° Regla	**Sup**	**Con**	**FC**	**Lif**	**P-S**	**cf**	**Ucq**	**UCd**	**bs**
Regla 1	0.052	0.47	0.39	6.98	0.04	0.0	1.0	0.0	0.00
Regla 2	0.076	0.50	0.32	2.98	0.05	0.0	1.0	0.0	0.00
Regla 3	0.31	0.12	0.04	1.51	0.01	0.0	0.0	0.0	1.0
Regla 4	0.02	0.85	0.83	6.35	0.02	0.0	0.0	0.0	1.0
Regla 5	0.01	0.83	0.80	6.17	0.01	0.0	0.0	0.0	1.0
Regla 6	0.03	0.65	0.60	4.88	0.02	0.0	0.0	0.0	1.0

Tabla 4.21: *Medidas para las reglas del Conjunto 5*

Analizando los las reglas obtenidas de este conjunto de datos, podemos decir que las dos primeras reglas de la tabla son las más representativa del conjunto de datos con un 5,2 % y un 7,6 % respectivamente de soporte. Podemos ver que en los análisis anteriores se ven las mismas tendencias de los usuarios con respecto a estas dos reglas. Eso si que en este caso los valores del factor de certeza son mucho mas bajos que el anterior caso.

Con respecto a las demás reglas podemos decir que estas para el usuario son inesperadas en todo el sentido de la regla, según los valores subjetivos que nos entregan. Estas reglas nos indican que el usuario una vez que haya ingresado en diferentes secciones del foro de la escuela regresa a la página principal del foro. Estas reglas son bastante interesante desde un punto subjetivo, pero desde un punto objetivo los valores de soporte

varían entre 1 % y 3 % o sea no son las más representativas dentro del conjunto de datos analizados pero si poseen valores bastante interesantes en las demás medidas.

4.9. Discusión sobre las reglas obtenidas e interpretación

Lo primero, es importante resaltar que al momento de realizar el análisis a los conjuntos que tenían aun elementos ruidosos, no se podía tener certeza de que fueran las reglas más representativas del del conjunto de datos, por esta razón podemos decir que es de suma importancia realizar un buen preprocesamiento de los datos, para poder obtener resultados mucho más validos y representativos.

También es importante destacar el tipo de regla que se puede obtener, como lo hemos señalado en la sección 4.7.3. Para este caso en especial sólo hemos obtenido reglas que tuvieran relaciones con las páginas visitadas con las referenciadas, principalmente de esta forma *página visitada* $\longrightarrow$ *página referenciada*. Con este tipo de reglas nos permite saber el comportamiento de navegación del usuario sobre determinadas páginas. Esto nos permite realizar procesos de marketing, de re-estructuración del sitio, entre otras, y así poder entregar una mejor información mientras el usuario realiza su navegación.

Con respecto a las reglas obtenidas, podemos decir que en el análisis del conjunto 4 hemos obtenidos tres reglas, siendo las más importantes del punto de vista de las medidas obtenidas, la **regla 2** GET/web/tablon_anuncios/$\longrightarrow$ www-etsi2.ugr.es y la **regla 3** GET/wwwforum/$\longrightarrow$ www-etsi2.ugr.es las cuales nos indican que los usuarios navegaban en las secciones del sitio como son el foro y el tablón de anuncio para volver a la página principal del sitio. Estas reglas fueron las que se han repetido durante los otros tres análisis previos a éste.

En el último análisis hemos encontrados muchas más reglas interesantes dentro del conjunto de datos que entregaba el usuario durante su navegación. Podemos decir en general que se mantiene lo obtenido en los análisis anteriores, donde el usuario visitaba el foro y el tablón de anuncio y volvía a la página principal del sitio. Esto quiere decir que es un comportamiento habitual en la navegación del usuario.

En relación a las otras reglas podemos decir que también el usuario navegaba por diversas secciones del foro como secciones relacionadas con temas como asignaturas, la escuela, aspectos generales para luego volver a la página principal del foro.

Las reglas encontradas nos indican un cierto comportamiento relacionado principalmente a la sección foro del sitio Web de la escuela, esto quiere decir que estas secciones

pueden ser las más visitadas.

4.10. Conclusiones

La lógica difusa nos permite manejar datos ruidosos, imprecisos, vagos e incompletos. Este tipo de datos, precisamente, es muy común en el ámbito de la Web, donde un exceso de información y la falta de estructura en los datos dificulta generalmente su manejo. La aplicación de la lógica difusa en la minería nos permite, además, mejorar la comprensión de los patrones obtenidos.

En este capítulo hemos revisado diferentes trabajos que aplican técnicas difusas en todos los tipos de Minería Web. Concretamente, nosotros nos hemos centrado en la aplicación de reglas de asociación difusas en la Minería Web de Uso.

Para ello, hemos planteado un modelo de obtención de reglas partiendo de un preprocesamiento de los datos en los ficheros log. A continuación, se han obtenido transacciones con valores difusos. Esto nos ha permitido extraer, mediante el algoritmo AprioriTID, reglas de asociación difusas de las que el usuario puede definir su composición. Hemos completando el modelo con la interpretación semántica de las reglas, aplicando medidas de evaluación de las reglas, tanto subjetivas como objetivas.

Además, hemos podido validar el funcionamiento del modelo en un caso real, extrayendo reglas de asociación difusas del análisis de los ficheros log del sitio web de la E.T.S. Ingenierías Informática y de Telecomunicación del Universidad de Granada (http://etsiit.ugr.es).

En el próximo capítulo, analizaremos la otra técnica que más se utiliza en la Minería Web, que es el Clustering. Éste lo veremos desde un punto de vista difuso, principalmente utilizando el algoritmo difuso de C-medias para la agrupación de sesiones de usarios.

Capítulo 5

Minería web de uso y clustering difuso: Análisis demográfico

En esta capítulo vamos a aplicar la técnica de clustering [1] para obtener diferentes grupos demográficos a partir de la información contenida en los ficheros log de servidores web. Concretamente, vamos a agrupar páginas Web para determinar cuáles son las páginas más representativas que el usuario utiliza cuando interactúa en algún sitio Web. Por otro lado, vamos a agrupar sesiones de usuarios, ya que queremos saber cuáles son los grupos de usuarios que se conectan, caracterizados por sus preferencias o intereses al navegar.

5.1. Clustering en la minería Web de uso

En el ámbito de la Web podemos decir que se han hecho diversos estudios orientados principalmente a realizar agrupamientos por contenido. Por ejemplo, cuando hacemos búsquedas sobre algún tema lo hacemos con algún buscador de Internet. Estos sistemas de búsqueda por temas son denominados motores de búsqueda los cuales indexan archivos almacenados en los servidores Web de los cuales podemos citar al sistema *Grokker*. Grokker es un sistema de búsqueda que permite realizar búsquedas en la base de datos de Yahoo!, en la tienda de libros Amazon y en Librería Digital ACM. Los resultados se agrupan por similitud de contenidos y también se pueden presentar de forma gráfica, en forma de esferas (clusters) agrupando temáticas.

[1] Ver nota a pie de página 1 en el capítulo de Introducción, página 4.

Podemos comentar el artículo [KC03] que presenta un enfoque relacionado con la creación de perfiles de usuarios a partir de los documentos de html y el registro de navegación del usuario. A través de un mapa estructurado (SOM) que es una red neuronal muy útil para visualizar datos de grandes dimensiones y una herramienta eficiente para aglomerar datos. La evaluación del contenido de la Web es usado para predecir las preferencias de los usuarios en la Web y la construcción automática de los perfiles, aquí podemos ver lo que hemos comentados anteriormente sobre el agrupamiento por contenido.

También, con el algoritmo de clustering jerárquico llamado COBWEB [Fis87], se caracteriza porque utiliza aprendizaje incremental para realizar las agrupaciones instancias a instancias. El algoritmo utiliza un árbol de clasificación donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Las instancias se van agregando una a una y el árbol se va actualizando en cada paso. La clave para saber cómo y dónde se debe actualizar el árbol, la proporciona una medida denominada utilidad de categoría, que mide la calidad general de una partición de instancias en un segmento.

Además, COBWEB pertenece a los métodos de aprendizaje conceptual o basado en modelos. Esto significa que cada cluster se considera como un modelo que puede describirse intrínsecamente, más que un ente formado por una colección de puntos [GCS+05]. Dentro de los algoritmos jerárquicos podemos mencionar a Slink [Har75], Cure [GRS97] y Chameleon [KHK99], entre otros.

En el caso de la Minería Web de Uso, los elementos a agrupar pueden ser las páginas Web y las sesiones de usuarios, con el objetivo de poder realizar un estudio o análisis demográfico.

Hay trabajos en la literatura relacionado con el clustering, por ejemplo en [NFJK99] que describe una técnica de *clustering difuso* para el caso de las sesiones de usuarios. El algoritmo llamado CARD (Competitive Agglomeration of Relational Data), que es de clustering de sesiones de usuarios, el cual considera la estructura del sitio y las URLs para calcular la similaridad entre dos sesiones de usuarios. Este enfoque requiere del cálculo de la distancia entre todas las sesiones pares, formando similaridades o una matriz de relaciones difusas. El objetivo de esta aplicación es definir las sesiones de usuarios a partir de los accesos de los usuarios a la Web y de la estructura que posee el sitio.

En este capítulo, nuestro estudio se basará principalmente en hacer agrupamientos o clustering de *páginas Web* y *sesiones de usuarios*. La primera para la búsqueda de las páginas Web más representativas para el usuario durante su navegación y la segunda para determinar grupos de usuarios con ciertas características o preferencias que nos permita,

por un lado realizar un análisis demográfico y por otro lado entregar una mejor información al usuario durante una futura navegación.

También es importante mencionar que a partir del análisis de agrupamiento de sesiones de usuarios podremos obtener diferentes perfiles de usuarios. En [MBKV+02], se presenta un enfoque en el estudio de perfiles de los usuarios, como la creación, modificación, almacenamiento, aglomeración e interpretación de los perfiles. Donde se describen los perfiles extendidos, los cuales contienen información adicional relacionada con el perfil del usuario, que puede usarse para personalizar y desarrollar a la medida el proceso de recuperación, así como también el sitio Web. Se consideró un modelo general de *clustering difuso* para poder manejar las imprecisiones o ambigüedades que se pueden presentan dentro de los perfiles extendidos.

Tomaremos este enfoque para continuar el estudio acerca de los perfiles de usuarios, el cuál lo realizaremos en el siguiente capítulo, y para mejorar la representación de los posibles grupos de usuarios que navegan por algún sitio Web que podamos obtener en nuestros análisis.

Para ello en este capítulo nos centraremos en otra de las técnicas más utilizadas en la Minería Web de Uso que es el Clustering. Comentaremos el Clustering de dos puntos de vista, del punto de vista *crisp* y del punto de vista *difuso*, es en este último enfoque donde nos detendremos para profundizar un poco más. Realizaremos varios experimentos, en los cuales agruparemos páginas similares y sesiones de usuario, utilizando diferentes medidas de similitudes y de validación.

5.2. Introducción al clustering

Las técnicas de clustering son técnicas de clasificación no supervisadas de patrones (observaciones, datos o vectores de característicos) en grupos o clusters. Estas técnicas han sido utilizadas en diversas disciplinas y aplicadas en diferentes contextos, lo cual refleja una gran utilidad en el análisis experimental de datos.

El *agrupamiento* se ha incluido dentro del ámbito de la inteligencia artificial encuadrándose dentro del *aprendizaje no supervisado*, por último la Minería de Datos recoge el agrupamiento como una de las clases de problemas a tratar dentro de su ámbito y recupera las técnicas y metodologías previamente desarrolladas extendiéndolas al volumen de datos que se procesan en este campo [And73], [DO74], [Har75], [Spa80], [JD98], [Bac95], [MJF99]. De forma más general, podemos definir el clustering como *el proceso*

de clasificación no supervisada de objetos.

5.2.1. Clustering vs clasificación

Primero, es importante distinguir entre *agrupamientos* o *clasificaciones no supervisadas* y *análisis discriminante* o *clasificación supervisada*. En el primer caso no se tiene ninguna información relacionada con la organización de los ítems en los grupos o clases y el objetivo es encontrar dicha organización en base a la proximidad entre ítems. Casi no existe información previa acerca de la estructura y la interpretación de las clases o grupos obtenidos es realizada posteriormente por el analista. En el segundo se posee información de qué clase pertenece cada ítem y lo que se desea es determinar cuales son los factores que intervienen en la definición de las clases y que valores de los mismos determinan estas. Se puede clasificar el agrupamiento y la clasificación en general según distintos criterios. Podemos ver esta clasificación con mayor detalle en [LW67] la cual posteriormente fue aplicada en [MJF99].

Un ejemplo claro de agrupamiento sería la búsqueda de grupos de clientes de una entidad bancaria utilizando para ellos los datos de la cuenta corriente: edad, dirección, nivel de renta,...etc. Y un ejemplo de clasificación sería encontrar los elementos que determinan la aparición de cáncer de pulmón analizando datos de, edad, calidad de vida, nivel económico,... etc. tanto de personas enfermas como sanas.

Segundo, podemos decir que la tarea de clasificar o clasificar objetos en categorías es una de las actividades más comunes y primitivas del Hombre y viene siendo identificada en función de grandes volúmenes de información en diversas areas [Bac95].

Intuitivamente, dos ítems o variables pertenecientes a un grupo válido deben ser más parecidos entre sí que aquellos que estén en grupos distintos y partiendo de esta idea se desarrollan las técnicas de agrupamientos. Estas técnicas dependen claramente del tipo de dato que se este analizando, de que medidas de semejanzas se estén utilizando y de que clase de problema se este resolviendo [MJF99].

En un sentido más concreto, el objetivo es reunir un conjunto de objetos en clases tales que el grado de asociación natural para cada individuo es alto con los miembros de su misma clase y bajo con los miembros de las otras clases. Lo esencial del análisis de agrupar se enfoca entonces a cómo asignar un significado a los términos, grupos naturales y asociación natural, donde natural usualmente se refiere a estructuras homogéneas y bien separadas [Kan].

5.2.2. Algoritmos de clustering

Existe una gran variedad de algoritmos de clustering que han surgido en los últimos años. En esta sección analizaremos brevemente algunos de los algoritmos más utilizados. En [GCS07] podemos ver una clasificación de estos algoritmos de clustering, según al método que corresponda.

Uno de los algoritmos más utilizados en aplicaciones de clustering es el algoritmo *c-medias*. Este algoritmo representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su centroide. El algoritmo se basa en la minimización de la distancia interna (la suma de las distancias de los patrones asignados a un agrupamiento al centroide de dicho agrupamiento). De hecho, este algoritmo minimiza la suma de las distancias al cuadrado de cada patrón al centroide de su agrupamiento [McQ67].

Dentro de la familia de los *c-medias*, podemos mencionar los algoritmos PAM [KR87], CLARA [KR90], CLARANS [RJ02] y sus extensiones.

Tradicionalmente los procesos del clustering generan particiones, en la cual cada patrón pertenece a uno y sólo un cluster o clase. Por lo tanto, los cluster o clases generados por un ”hard” clustering son disjuntos. La técnica del clustering difuso extiende este concepto para asociar cada patrón a todos los clusters usando una función de pertenencia [Zad75]. La salida de este algoritmo es una agrupación, y no una partición.

Otro algoritmo denominado EM [GCS$^+$05] pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjunto de datos. Este algoritmo en concreto es un clustering probabilístico, donde se trata de obtener la función de densidad de probabilidad desconocida a la que pertenece el conjunto completo de datos. Cada uno de los clusters determinado por el algoritmo, estará definido por los parámetros de una distribución normal. Otros algoritmos pertenecientes a esta familia son AUTOCLASS [CS96], SNOB [WD94] y MCLUST [FR99].

Dentro del área del clustering podemos mencionar el *clustering jerárquico*. Esta técnica de clustering puede ser utilizada para determinar las particiones iniciales de los datos, esto para poder aplicar alguna otra técnica de clustering que necesite apriori saber el número de las particiones. Profundizaremos más este tema en secciones posteriores.

Existe muchos más algoritmos que se utilizan en esta área del clustering, y muchos de estos algoritmos han sido extendido a una forma difusa.

5.3. Modelo general de clustering

El problema del clustering puede ser tratado de dos diferentes enfoques, entre ellas, esta el convencional **crisp**, donde cada objeto clave es clasificado única y totalmente en una determinada categoría o grupo, y el enfoque **difuso** que es más flexible, donde cada objeto puede pertenecer a varias categorías o grupos con diferentes grados de asociación.

Todos los enfoques que son mencionados en el artículo de [LW67] y también en [MJF99] parten de la hipótesis de no-solapamiento, es decir que se desea un agrupamiento donde cada punto pertenece sólo a un grupo. Cuando se relaja esta hipótesis aparecen otros enfoques de agrupamiento que admiten solapamiento o no-exclusivos. Los métodos de agrupamiento no-exclusivo con mayor éxito son los que suponen que los grupos son conjuntos difusos de forma que un ítem puede pertenecer a diversos grupos con cierto nivel de pertenencia a cada uno.

Ahora definiremos formalmente ambos casos, primero definiremos el enfoque [2] *crisp* y luego el *difuso*.

5.3.1. Modelo para el clustering crisp

Dado un conjunto $X = \{x_1, x_2, ..., x_n\}$ objetos, está definido como una partición crisp como una familia de clases (o clusters) $P = \{A_1, A_2, ...A_C\}$ tal que:

$$\bigcup_{i=1}^{C} A_i = X \qquad A_i \cap A_j = \oslash \qquad \forall i \neq j \qquad \oslash \subset A_i \subset X \quad \forall i \tag{5.1}$$

donde *C* puede tomar valores $2 < C < N$, cuando el valor de $C = 1$, esto significa que no existe una agrupación en *X*. Y cuando $C = N$ representa el caso en el cual cada objeto representa un grupo o clusters por si mismo.

Una partición *P* crisp se representa por medio de una matriz $U = [u_{ik}]$, $i = 1$, $k = 1$, mientras un elemento $u_{ik} \in \{0,1\}$ que representa la pertenencia o no pertenencia del k-ésimo objeto de X_k al i-ésimo cluster *i*.

[2] Desde aquí en adelante, cuando mencionemos al Clustering, estaremos haciendo referencia al Clustering *"tradicional"* o Crisp, en el caso contrario será el Clustering Difuso.

5.3.2. Modelo para el clustering difuso

Dado un conjunto de datos $X = \{x_1, x_2, \ldots, x_n\}$ donde x_k es generalmente un vector de características $x_k = \{x_1, x_2, \ldots, x_{kp}\}$ para todo $k \in \{1, 2, \ldots, n\}$ siendo un espacio p-dimensional. Un problema de la clasificación difusa es encontrar una pseudopartición difusa que represente la estructura de los datos de mejor forma posible.

Una pseudopartición difusa de *X* es una familia de *c* subconjuntos difuso de X, denotada por $P = \{A_1, A_2, \ldots, A_c\}$ que satisface la ecuación:

$$\sum_{i=1}^{c} A_i\left(x_k\right) = 1 \tag{5.2}$$

para todo $k = \{1, 2, 3, \ldots, n\}$ siendo n el número de los elementos del grupo *X*. Esto es, el grado de pertenencia de un elemento en todas las familias debe ser igual a 1 y también debe satisfacer la ecuación:

$$0 < \sum_{i=1}^{m} A_i\left(x_k\right) < 1 \tag{5.3}$$

para todo $i = \{1, 2, 3, \ldots, c\}$ donde c representa el número de clases. Esto es, la suma del grado de pertenencia de todos los elementos de una familia que debe ser menor que el número de elementos existentes en el conjunto X.

5.3.3. Medidas de semejanza

Cuando hablamos de medidas de semejanza, nos referimos a las relaciones que existen entre los ítems o variables que son analizadas en el proceso de Clustering, ya que es importante la semejanza de los ítems a la hora de definir un cluster, y es necesario establecer una forma de medir esta semejanza. Lo más común es calcular el concepto contrario, es decir, la diferencia o desemejanza entre los ítems usando la medida de distancia en un espacio de características.

La distancia Euclídea es una de las medidas más intuitiva de la distancia entre dos puntos en un espacio de dos o tres dimensiones. Esto puede ser útil cuando los clusters son compactos [MJ96]. En la tabla 5.1, veremos distintas funciones de distancia.

Otras clasificaciones de distancias pueden ser consultadas en [DOF03].

Nombres	Expresiones
Euclídea o *norma-l2*	$d_2(i,k) = \left[\sum_{j=1}^{m} (x_{ij} - x_{kj})^2\right]^{1/2}$
Manhatan o *norma- l1*	$d_1(i,k) = \sum_{j=1}^{m} \lvert(x_{ij} - x_{kj})\rvert$
Norma del supremo	$d_\infty(i,k) = sup_{i\in\{1,2,\ldots,m\}} \lvert x_{ij} - x_{kj}\rvert$
Minkowski o $normal_p$	$d_p(i,k) = \sum_{j=1}^{m} \left[\lvert x_{ij} - x_{kj}\rvert^p\right]^{1/p}$
Distancia de Mahalanobis	$d_M = \left[(x_i - x_k)^T \sum^{-1} (x_i - x_k)\right]$ $\sum$ es la covarianza muestral o una matriz de covarianza intra-grupos.

Tabla 5.1: *Distintas funciones de Distancia*

5.4. Obtención de la partición inicial de datos

Un tema importante al momento de realizar un análisis con el clustering es saber de alguna forma el número de clusters, ya que los algoritmos buscan optimizar una determinada estructura de clusters para ese número determinado de clusters. Dar por conocido el número de cluster en algunos casos puede ser algo razonable si se sabe con certeza cuales son las propiedades y la estructura de los datos que se están analizando, pero no tiene ninguna validez si se esta realizando un análisis sobre datos que se desconoce cuales son sus propiedades. Sin embargo, hay ciertos casos donde esta información no es necesaria, con lo que pasa a ser un valor meramente informativo.

Una forma para determinar el número de clusters es llevar a cabo un clustering aumentando el número de clusters (*c*), observando como la función cambia con dicho número. Por ejemplo, si esta función es la suma de los errores cuadráticos J_e, su valor decrecerá de forma monótona con el valor de *c*. Si el conjunto de datos está formado por *n* muestras, y éstos pueden agruparse correctamente en $\hat{m}$ clusters compactos y bien separados, lo que se esperaría que J_e disminuyera rápidamente hasta llegar al valor $m = \hat{m}$, decreciendo después mucho más lentamente hasta llegar a $m = n$, por lo que un sencillo análisis gráfico de la dependencia entre J_e y *m* podría ser suficiente para determinar el número óptimo de grupos [DHS00].

Otra forma de poder terminar el número de clúster es utilizar el clustering jerárquico, pudiendo ser lo suficientemente representativo como para determinar el número óptimo de grupos, ya que suele aceptarse que cuando la unión (o división) de los dos clusters da lugar a una situación muy diferente de la que comentamos anteriormente, ya que esto es indicativo de la presencia de un agrupamiento natural, de un Clustering correcto [DHS00].

Para nuestro caso es necesario conocer apriori el número de grupos, y así obtener durante el análisis los grupos más representativos del conjunto de datos. Para eso es necesario utilizar alguna técnica que nos permita obtener los números de los grupos.

Las técnicas más utilizadas son dos: el clustering particional y el clustering jerárquico para la obtención de los conjuntos previos.

- *El clustering particional* consiste en dividir el conjunto de datos en grupos, de manera que los datos que se encuentren en un grupo sean lo más parecidos entre sí, a la vez que lo más diferentes posible a los datos que se encuentren en los demás grupos. Es decir, se trata de dar una partición del conjunto de datos.
- *El clustering jerárquico* es otra técnica de clustering que en vez de crear una única partición, crea una sucesión encajada de particiones cuya estructura puede ser representada por medio de un árbol.

Para la obtención de la partición inicial de datos hemos elegido el clustering jerárquico, principalmente porque nos da una mejor representación de los posibles grupos a través de un dendograma o jerarquía de grupos. A partir de ese dendogramas podemos inferir cual sería la mejor partición inicial para realizar el análisis utilizando diferentes criterios.

5.4.1. Clustering jerárquico

Ahora definiremos formalmente lo que es un clustering jerárquico así, formalmente una clustering jerárquico o también conocido como dendograma (Ver figura 5.1) es un *par (T, h)* formado por árbol *T* junto con una aplicación *h* definida en los nodos de *T* de manera que se cumple:

- $h(A) = 0 \longleftrightarrow A = \{w\}$ para algún $w \in \Omega$
- Si $A \cap B = \varphi$ entonces $A \subset B \longleftrightarrow h(A) \leq h(B)$

A partir de un dendograma es posible construir una nueva distancia ultramétrica en los datos de la siguiente manera: $U(w_{A \in T}, w') = min\{h(A)/w, w' \in A\}$

Esta disimilaridad posee unas propiedades especiales:

I. $\forall w, w' \in \Omega \Rightarrow U(w, w') = U(w', w)$

II. $\forall w \in \Omega \Rightarrow U(w, w) = 0$

III. $\forall w, w', w'' \in \Omega \Rightarrow U(w, w'') \leq max\{U(w, w'), U(w', w'')\}$

La última propiedad se denomina propiedad ultramétrica y las distancia ultramétrica cumplen estas tres condiciones. Se puede probar que existe una biyección entre distancias ultramétricas y clasificaciones jerárquicas. Este hecho nos permite medir la bondad de una clasificación jerárquica (basta utilizar alguna medida que calcule la diferencia entre la distancia ultramétrica inicial de los datos y la distancia ultramétrica de la clasificación obtenida) y nos permite realizar la búsqueda en el conjunto de las distancia ultramétricas en lugar de hacerlo directamente en el de las clasificaciones jerárquicas. [Bez81], [DGSV96].

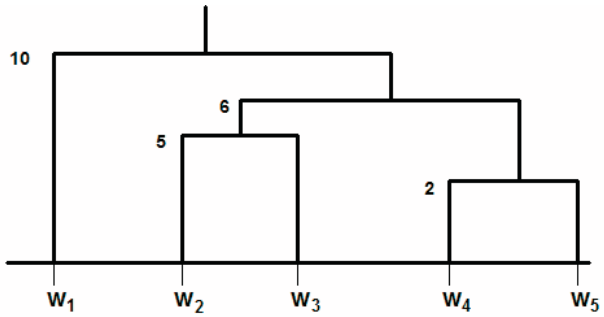

Figura 5.1: *Un ejemplo de un dendograma para el caso del Clustering Jerárquico.*

5.4.2. Algoritmos aglomerativos y divisivos

Existen principalmente dos tipos de algoritmos para llevar a cabo el clustering jerárquico: los algoritmos *aglomerativos* y los algoritmos *divisivos*. Ambos algoritmos tienen una característica en común: la forma de construir el árbol es local. Supondremos que la información de entrada al algoritmo va a ser una matriz cuadrada simétrica que representa las distancias entre los datos y que llamaremos matriz similitudes.

Explicaremos brevemente cada uno de estos algoritmos, y señalaremos algunos trabajos relacionados con ellos:

- **Algoritmos jerárquicos aglomerativos**: estos algoritmos producen una sucesión de conglomerados de tal manera que en cada paso el número de conglomerados va disminuyendo. Son algoritmos del tipo ”botton-up”(de abajo a arriba). Inicialmente

se empieza con conglomerados que consisten de un sólo elemento. Los conglomerados de un paso dado son obtenidos al combinar dos conglomerados del paso anterior. Los criterios más usados para juntar los conglomerados son ”single-link” (enlace individual), ”complete-link”, ”ward-link” o cualquier medida de distancia intergrupo. Todos estos criterios usan una medida de distancia entre los vectores. Los algoritmos jerárquicos aglomerativos son los más usados para construir conglomerados y están disponibles en la mayoría de los programados estadísticos. También son los que se computan más rápidamente.

- **Algoritmos jerárquicos divisivos**: también producen una sucesión de conglomerados pero en este caso el número de ellos crece en cada paso. Son algoritmos del tipo ”top-down”(de arriba hacia abajo). Inicialmente se empieza con un sólo conglomerado que contiene a todas las observaciones. Los conglomerados de un paso dado son obtenidos al dividir en dos un conglomerado del paso anterior. Loa algoritmos jerárquicos divisivos demandan más esfuerzo computacional que los algoritmos aglomerativos.

Algunos algoritmos que pertenecen a esta clasificación son:*Agnes* [KR90], *Chameleon* [KHK99], *Diana* [KR90], BIRCH (Balanced Iterative Reducing and Clustering using Hierchical) [TZL96], CURE (Clustering Using Representatives) [SRK98] y ROCK (RObust Clustering algorithm using linKs) [SRK00].

Nosotros utilizaremos el algoritmo jerárquico aglomerativo para el análisis, ya que como lo hemos mencionado anteriormente son los más rápidos en los análisis.

5.4.3. Criterios para el cálculo de particiones

En el clustering jerárquico los criterios usuales para obtener los grupos están relacionados principalmente con minimizar la distancia dentro del grupo y/o maximizar la distancia entre los grupos, definiendo esta distancia como la separación entre los centroides, medias, etc [JD98].

Analizaremos estos criterios, hay que tener en cuenta el conjunto de posible particiones obtenidos de la secuencia S_k, y seleccionar de ellos el que minimiza una cierta medida de agregación del grupo (la distancia intracluster) y maximiza una medida de separación del grupo (la distancia del intergrupo).

Sea $H = \{(P_1, l_1), (P_2, l_2), ..., (P_h, l_h)\}$ un agrupamiento jerárquico definido sobre $X = \{x_1, x_2, ..., x_n\}$, definimos: $\forall x_i, x_j \in X \; u_{ij} = l_t \; / \forall t' \geq t \; x_i, x_j \in P_{t'}$. Es decir

u_{ij} es el mínimo nivel para el cual x_i y x_j pertenecen al mismo conjunto.

Podemos definir como $d_2^{S_k}$ la distancia máxima dentro de los elementos de un grupo, y $D_2^{S_k}$ será la distancia mínima entre los grupos. Una vez definidas las funciones de distancias, ahora el problema se convierte en la búsqueda de un cierto $S_t \in S_k$ tal que [DGSV96]:

$$D^{s_t} = max_k D^{s_k}, \quad d^{s_t} = min_k d^{s_k} \tag{5.4}$$

Veremos diferentes definiciones, el primero está basado en una optimización de criterios sobre los valores medios, mientras el segundo se basará en el criterio de maximizar y minimizar, clásico en la teoría de decisión difusa [DGSV96].

- **Mínima distancia entre los elementos de un grupo, máxima distancia entre grupos**:
  - *Distancia media global entre todos los posibles clusters*: $\forall \alpha \in S$; sea $P \in S^\alpha$; sea $|P|$ el cardinal de *P*.

$$d_1^\alpha = \frac{\sum_{P \in S^\alpha} d^\alpha (P)}{|S^\alpha|} \tag{5.5}$$

Con

$$d_1(P) = \begin{cases} \frac{\sum_{x_i, x_j \in P (i \neq j)} 2u_{ij}}{|P|(|P|-1)} & \text{si} \quad |P| < 1 \\ 0 & en \quad otro \quad caso \end{cases}$$

  - *Distancia media global entre los clusters*: $\forall P, P' \in S^\alpha$; $D(P, P') = u_{ij} / x_i \in P x_j P'$.

$$D_1^\alpha = \frac{\sum_{P, P' \in S^\alpha} D^\alpha (P, P')}{|S^\alpha| (S^\alpha - 1)} \tag{5.6}$$

- **Mínima distancia entre elementos de un grupo, máxima distancia entre grupos**:
  - *Distancia media*: $\forall \alpha \in S$; sea $P \in S^\alpha$; sea $|P|$

$$d_2^\alpha = max_{P \in S^\alpha} \left(max_{x_i, x_j \in P} (u_{ij}) \right) \tag{5.7}$$

$$D_2^\alpha = min_{P, P' \in S^\alpha, P \neq P'} \left(D^\alpha \left(P, P' \right) \right) \tag{5.8}$$

- **Tambien se puede utilizar una combinación entre las medidas anteriores**:

$$D_1^\alpha - d_1^\alpha \quad o \quad bien \quad D_2^\alpha - d_2^\alpha \tag{5.9}$$

  el segundo caso proporciona la partición más ”estable”, el mayor intervalo de similitud.

- **crisp más cercano:** otro de los criterios que se puede utilizar es el relacionado con el crisp más cercano, donde el *intervalo que contiene al 0.5* pasaría hacer la partición óptima inicial.

Utilizaremos algunos de estos criterios para determinar el número de la partición inicial de datos.

5.5. Clustering c-medias

Visto como un problema de optimización, el objetivo del algoritmo C-medias es minimizar la disimiliridad de los elementos dentro de cada cluster al mismo tiempo que se maximiza la disimilaridad de los elementos que caen en diferentes clusters [McG67]. Este método, el *clustering c-medias* o *c-medias* es uno de los más utilizados para el análisis de agrupamientos debido a su sencillez y a sus bajos requerimientos computacionales.

El nombre del algoritmo hace referencia a que existen ”C” clases, siendo necesario, por tanto conocer apriori el número de grupos. El algoritmo C-medias selecciona un número de objetos ”C” para ser utilizados como centroides iniciales. Esta selección puede relizarse de diferentes forma, por ejemplo escogiendo la primera muestra del conjunto, de forma aleatoria o realizando una partición al azar en ”C” clusters y calculando su centroides. Luego asigna cada elemento a su centro más próximo en función de una medida de proximidad.

Calcula los nuevos centroides utilizando para ello todos los objetos de cada grupo, consiguiendo de esta forma la minimización a través de la siguiente función:

$$J(\Theta, U) = \sum_{i=1}^{n} \sum_{j=1}^{m} u \left\| X_i - u_j \right\|^2 \tag{5.10}$$

$$u_{ij} = \begin{cases} 1 \quad \text{si} \quad d(x_i, \Theta_j) = min_{k=1,\ldots,m} d(x_i, \Theta_k) \\ \\ 0 \qquad en \quad otros \quad casos \end{cases}$$

donde *U* es una matriz de dimensiones *nxm* cuyos elementos (*i* e *y*) se corresponden con $u_j(x_i)$, $d(x_i)$, Θ_j es la distancia entre el *i-ésimo* y el prototipo *j-ésimo*, *n* es el número de elementos y *m* el número de grupos.

5.6. Clustering difuso c- medias

El método de *clustering difuso c-medias* fue propuesto por Bezdek [Bez81]. Este método difuso c-medias puede ser desarrollado a través de un algoritmo iterativo, basado en la minimización de un índice de desempeño, que indica la adecuación de la pseudopartición generada. El desempeño del algoritmo está influenciado por la elección del número de cluster o clases *c*, de los centros del inicial cluster, de la medida de distancia que se aplique, del criterio de parada y de las propiedades geométricas de los datos.

Así, el objetivo de este algoritmo es encontrar la mejor partición matricial que puede tener valores entre [0,1] de acuerdo a las condiciones planteadas en anteriormente. Este objetivo es obtenido cuando se minimiza la siguiente función:

$$Jm = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|, 1 \leq m \leq \infty \tag{5.11}$$

donde *m* es cualquier número real mayor que 1, u_{ij} es el grado de pertenencia de en el cluster j, x_i es el i-ésimo valor d-dimensional del conjunto de datos, c_j es el centro d- dimensional del cluster y $\|*\|$ es cualquier norma que exprese la similaridad entre cualquier valor medido y el centro.

La partición difusa es realizada a través de una optimización iterativa de la función objetivo mostrada anteriormente, con la actualización de la pertenencia u_{ij} y los centros del cluster c_j por las siguientes ecuaciones:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{|x_i - c_j|}{|x_i - c_k|}\right)^{\frac{2}{m-1}}} \tag{5.12}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} \cdot x_i}{\sum_{i=0}^{n} u_{ij}^{m}} \tag{5.13}$$

La iteración del algoritmo para cuando $max_{ij}\left\{\left|u_{ij}^{k+1} - u_{ij}^{k}\right|\right\} < \epsilon$, donde ϵ es el criterio de terminación cuyo valor debe estar entre [0,1]. Donde k es el número de iteración

del algoritmo [Bez81]. Utilizaremos este algoritmo para encontrar diferentes grupos de usuarios enfocados principalmente en la interacción que estos realizan en el sitio Web.

5.7. Validación del clustering

Es importante en todo experimento corroborar de alguna manera que los resultados son óptimos o simplemente que sean correctos. De esa manera es importante utilizar medidas que nos entreguen información acerca de los resultados.

Para validar los resultados en las diferentes agrupaciones que realizaremos, tanto para la agrupación de páginas similares como para las sesiones de usuarios, utilizaremos dos medidas importantes para evaluar los resultados obtenidos. Estas medidas son el *Coeficiente de partición* y *la Entropía de la partición* [PB95].

El Coeficiente de partición lo definiremos de la siguiente manera:

$$CP = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^2}{N} \tag{5.14}$$

y la Entropía se define:

$$CE = -\frac{1}{N} \cdot \left[\sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \cdot log_a(u_{uk}) \right] \tag{5.15}$$

Estas medidas nos permitirán saber si los resultados obtenidos de los diferentes análisis serán óptimos o no. También nos permitirán saber si algo va mal dentro de los diferentes grupos, o si se está realizando algo erróneo.

5.8. Aplicaciones del clustering en la minería web de uso

Dentro del área de la Minería Web de Uso podemos encontrar diversos estudios relacionados principalmente en agrupamientos por contenido, siendo este uno de las principales área donde se utiliza el clustering en la Web. Por ejemplo podemos nombrar algunos buscadores que utilizan esta técnica para realizar agrupamiento o clustering por contenido como Vivisimo, Grokker, Clusty, iBoogie.

Con esto podemos decir que existen diferentes sistemas que se preocupan de saber cuales son las características del usuario relacionado principalmente en el contenido que el usuario visita o los temas que se relacionan con su navegación.

Por esta razón surge una necesidad, la necesidad de agrupar las páginas de los usuarios para saber cuales son las páginas más representativas, también un segundo enfoque relacionado con la agrupación de las sesiones de usuarios, ya que a partir de esta agrupación podemos identificar grupos de usuarios con ciertas características, preferencias y/o intereses en su navegación. Lo cual nos permitirá realizar un estudio demográfico y también obtener diferentes perfiles que representen a los conjuntos de las características de los usuarios. Realizando estas agrupaciones podemos de alguna manera entregar una mejor información al usuario durante su navegación.

La figura nos muestra un enfoque general de lo que hemos planteado hasta estos momentos. Esta representación que muestra la figura esta hecho en un pseudo-lenguaje que nos permitirá ver todo lo relacionado con la partición inicial de los datos, pasando por la técnica de agrupamiento tanto para las páginas como para las sesiones y finalmente la validación de los resultados que es un punto de suma importancia al momento de obtener los resultados.

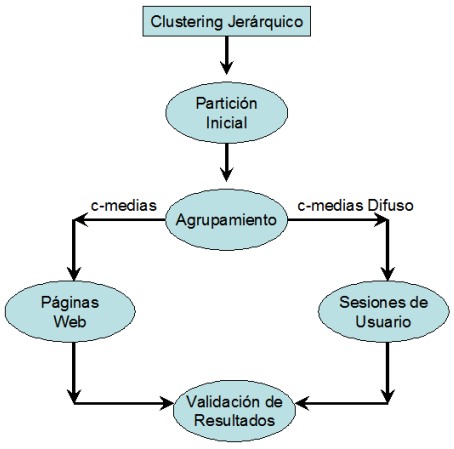

Figura 5.2: *Diagrama de los diferentes enfoques planteados para el clustering*

Para la obtención de la partición inicial hemos utilizado el clustering jerárquico, primero porque nos entrega una visualización de los grupos a través de un dendograma y a partir de este podemos utilizar diferentes criterios para determinar la partición óptima inicial.

En la obtención de páginas similares hemos utilizado el algoritmos c-medias tradicional, ya que buscamos sólo conjuntos representativos de páginas visitadas por el usuario. Y para el agrupamiento de sesiones de usuarios utilizamos el c-medias difuso, ya que este nos permite obtener una mejor representación de grupos de usuarios y nos permite obtener similaridades entre los diferentes conjunto de sesiones mucho más cercanas a la realidad que otro algoritmos tradicionales. Y a través de este último análisis poder obtener y representar diferentes perfiles de usuarios.

5.8.1. Modelo de datos

Es importante antes de realizar cualquier tipo de análisis saber cuales son los elementos con los cuales se cuentan para realizarlo. Para esto, nos basaremos en el modelo de datos que fue definido en capítulos anteriores (Ver sección 3.4). Así un conjunto de sesiones de usuarios **S**, se puede definir como:

$$S = \{s_1, s_2, \ldots, s_m\} \tag{5.16}$$

Donde cada sesión de usuario *S* está definida por un conjunto de páginas *P*, que se define como:

$$P = \{p_1, p_2, \ldots, p_n\} \tag{5.17}$$

Para este análisis veremos a las sesiones de usuarios conceptualmente como una matriz sesión-sesión *mxn* donde:

$$UU' = [sim(s_i, s'_j)], 1 \leq i \leq m, 1 \leq j \leq n \tag{5.18}$$

Donde $sim(s_i, s'_j)$ representa la semejanza de la sesión de usuario s_j en la sesión de usuario s'_i.

A continuación veremos diferentes experimentos relacionado con el clustering de páginas similares y luego con el agrupamiento de sesiones de usuarios.

5.9. Clustering de páginas similares: caso real

Ahora veremos algunos resultados relacionado con el análisis de los datos. El principal objetivo es encontrar grupos o clases diferentes en los archivos Web log, en este caso fue el tipo de archivo ECLLF, que nos permitan saber cuales son los principales grupos de interacción de los usuarios que navegan por el sitio Web.

Es importante mencionar que igual que ocurrió en la experimentación del capítulo anterior, el sitio de la Escuela de Informática y Telecomunicaciones de la Universidad de Granada se ha actualizado durante los diversos experimentos que hemos realizado. El conjunto 1 y 2 de datos se extrajo del sitio web antiguo mientras que el conjunto 3 es del sitio web nuevo. Aunque la raíz de las direcciones de páginas cambia de un sitio a otro, la estructura de ambos sitios es similar por lo que no hemos tenido problemas a la hora de agrupar por direcciones de páginas.

Por ejemplo, cual de las páginas el usuario visita habitualmente y así poder saber que información es la más demanda por el usuario. También podemos determinar la hora que el usuario se conecta habitualmente, podemos determinar cuan paciente es el usuario en su navegación, entre otras cosas. En la figura 5.3 podemos ver un diagrama que explica lo relacionado con la obtención de grupos de páginas similares. Para solucionar estas situaciones utilizaremos el algoritmo de clustering c-medias, porque es uno de los algoritmos más utilizados y a la vez uno de los más óptimos al momento de realizar agrupamientos.

De la misma manera que se realizó en la sección de las reglas de asociación difusas (ver sección 4.4.1), analizaremos diferentes conjuntos de datos para determinar los grupos o clusters más representativos. Para realizar la agrupación de páginas más similares utilizaremos la técnica de clustering.

Para los diversos análisis que realizaremos en esta sección plantearemos el objetivo, el modelo y la medida de similaridad que utilizaremos:

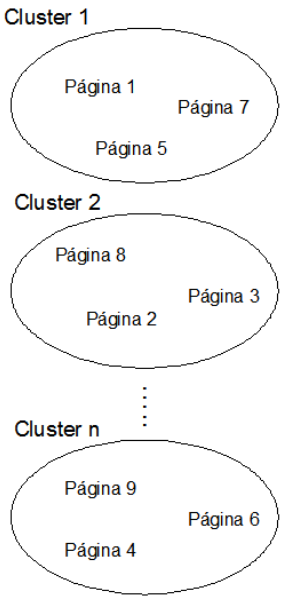

Figura 5.3: *Agrupación de páginas similares*

5.9.1. Características generales del experimento

- **Objetivo general:** es obtener conjuntos de páginas Web similares entre si para saber cuales son las grupos de páginas más utilizadas por el usuario mientras navega en el Web.

- **Conjunto de datos:** En tabla 5.2 podemos los diferentes conjuntos de datos que utilizaremos en los análisis.

Conjuntos De Datos	**N° Transacciones Originales**	**N° Transacciones Objetivas**	**Preprocesamiento**
Conjunto 1	100900	100810	Eliminación transacciones idénticas
Conjunto 2	100810	12910	Preprocesamiento Completo
Conjunto 3	98202	15676	Preprocesamiento Completo

Tabla 5.2: *Resumen de los Conjuntos de Datos para el análisis de páginas similares*

- **Modelo:** sea un conjunto de páginas Web $P = \{p_1, p_2, ..., p_n\}$, donde p_n corresponde a una página Web. Y tenemos otro conjunto de páginas Web $Z = \{z_1, z_2, ..., z_n\}$,

donde z_n corresponde a una página Web. Se quiere obtener la similaridad entre las distintas páginas Web $sim(p_n, z_n)$.

- **Medida:** la medida para determinar las semejanzas de cada página es la llamada *Levenshtein* [RB03]. Esta medida consiste en determinar cuán similares son dos cadenas de caracteres. Esta medida la utilizaremos para comparar diferentes páginas Web y así obtener conjuntos de páginas similares. Siendo *x* una palabra definida $x = \{x^1, ..., x^p\}$ e *y* otra palabra definida $y = \{y^1, ..., y^q\}$, y *z* está definida por la distancia de Hamming como $\sum_i^p z\left(x^i, y^i\right)$. Esta medida está definida de la siguiente manera:
  $L((x^1, ..., x^p), (y^1, ..., y^q)) =$

$$\begin{cases} p & q = 0 \\ q & p = 0 \\ min\{L\left(\left(x^1, ..., x^{p-1}\right), \left(y^1, ..., y^q\right)\right)\} + 1 & \\ L\left(\left(x^1, ..., x^p\right), \left(y^1, ..., y^{q-1}\right)\right)\} + 1 & otrocaso \\ L\left(\left(x^1, ..., x^{p-1}\right), \left(y^1, ..., y^{q-1}\right)\right) + z\left(x^p, y^q\right)\} & \end{cases}$$

- **Técnica utilizada:** utilizaremos la técnica del Clustering a través del algoritmo C-medias para poder obtener los diferentes grupos de páginas similares, principalmente el algoritmo c-medias.

- **Obtención del número de particiones iniciales:** para determinar el número de clusters para el análisis hemos utilizado la técnica del clustering jerárquico, basándonos en algunos criterios que hemos comentados en la sección 5.3.4. En la tabla 5.3 podemos ver el análisis para la obtención del número de particiones inicial óptima.

  Al analizar la tabla nos damos cuenta que la mejor opción para una partición óptima sería el α-corte **0.76**, por dos razones. Primero porque nos entrega el valor más alto en H que es 0.28 y eso nos indicaría que es la partición más "estable" y segundo porque si hacemos referencia a otro criterio de partición denominado el *más proxima a la relación*, que indica que será la mejor partición aquella que contenga el intervalo 0.5, que es el caso del α-corte **0.76**.

  Por lo tanto, según lo analizado anteriormente, el número óptimo para realizar el análisis de clustering es 9.

5.9.2. Resultados con el conjunto 1

- **Conjunto 1:** en este análisis utilizaremos el mismo conjunto de datos que fue ana-

Nivel (α)	**Particiones**	**d2** $[1-\alpha_i]$	**D2** $[D2-d2]$	**H**
0.04	{74,75};{38,39};{36,37};{56,57};{18,19};{66,67};{46,47};{28,29};{64,65};{10,11};{48,49};{26,27};{5,4};{42,43};{32,33};{60,61};{14,15};{52,53};{72,73};{2,3};{40,41};{34,35};{58,59};{16,17};{20,21};{68,69};{6,7};{44,45};{30,31};{62,63};{50,51};{24,25};{8,9};{70,71};{22,23};{54,55};{12,13}	0.96	1.0	0.04
0.28	{74,75,1}	0.72	0.96	0.24
0.48	{38,39, 36,37};{56,57,18,19};{66,67,8,9};{46,47,28,29};{64,65,10,11};{48,49,26,27};{70,71,5,4};{42,43,32,33};{60,61,14,15};{52,53,22,23};{72,73,2,3};{40,41,34,35};{58,59,16,17};{54,55,20,21};{68,69,6,7};{44,45,30,31};{62,63,12,13};{50,51,24,25}	0.52	0.76	0.24
0.76	{74,75,1,38,39,36,37};{66,67,8,9,46,47,28,29};{64,65,10,11,48,49,26,27};{70,71,5,4,42,43,32,33};{60,61,14,15,52,53,22,23};{72,73,2,3,40,41,34,35};{58,59,16,17,54,55,20,21};{68,69,6,7,44,45,30,31};{62,63,12,13,50,51,24,25}	0.24	0.52	**0.28**
0.84	{74,75,1,38,39,36,37,56,57,18,19}	0.16	0.24	0.08
0.96	{66,67,8,9,46,47,28,29,64,65,10,11,48,49,26,27};{70,71,5,4,42,43,32,33,60,61,14,15,52,53,22,23};{72,73,2,3,40,41,34,35,58,59,16,17,54,55,20,21};{68,69,6,7,44,45,30,31,62,63,12,13,50,51,24,25}	0.04	0.16	0.12

Tabla 5.3: *Cálculo para la obtención de una partición óptima: conjunto 1*

lizado para las reglas de asociación difusas (Ver sección 4.1.1). Este conjunto de datos consta de 100810 transacciones y corresponde al conjunto 1 (Ver tabla 5.2).

- **Resultados Conjunto 1:** de este grupo de datos hemos obtenidos los siguientes grupos o clusters de las páginas visitadas por el usuario (Ver tabla 5.4). Estos resultados han sido validados a través de dos medidas; la entropía y el coeficiente de partición (Ver ecuaciones 5.14 y 5.15).

Cluster	**[Centroides]**	
Cluster 0	[GET/wwwforum]	
Cluster 1	[GET/web/tablon_anuncios]	
Cluster 2	[GET/css/estilo.css]	
Cluster 3	[GET/css/wwwforum.css]	
Cluster 4	[GET/graficos/2002/cabecera_r5_c4.gif]	
Cluster 5	[GET/wwwforum/index.php?task=view_forum&forum_id=3]	
Cluster 6	[GET/alumnos/diegorp/canal.css]	
Cluster 7	[GET/includes/img/space.gif]	
Cluster 8	[GET/alumnos/diegorp/canalplus.html]	
	Entropía	0.00
	Coeficiente de Partición	0.99

Tabla 5.4: *Resultados grupos de páginas visitadas: Conjunto 1*

De los conjuntos de páginas encontrados, sobresalen de las demás páginas aquellas páginas que están relacionados con alguna imagen. Estos resultados no representan totalmente la realidad, ya que existen elementos ruidosos, en este caso particular, diversos grupos de páginas relacionadas con imagenes.

Para solucionar este problemas, hemos realizado un preprocesamiento eliminando los elementos que sean ruidosos dentro del análisis. Estos elementos ruidosos incluyen imágenes, páginas relacionadas con javascript. Al igual que el anterior experimento, utilizamos el análisis realizado en él, el cual se encuentra en la tabla 5.3 y utilizaremos el valor óptimo de clusters encontrado, el cual era 9.

5.9.3. Resultados con el conjunto 2

- **Conjunto 2:** el conjunto de datos para el análisis consta de 12920 transacciones. Este conjunto lo podemos ver en la tabla 5.2.

- **Resultados Conjunto 2:** hemos obtenido los siguientes grupos o cluster de páginas visitadas por el usuario (Ver tabla 5.5).

Cluster	[Centroides]	
Cluster 0	[GET/wwwforum]	
Cluster 1	[GET/web/tablon_anuncios]	
Cluster 2	[GET/css/estilo.css]	
Cluster 3	[GET/css/wwwforum.css]	
Cluster 4	[GET/web/tablon_anuncios/css/estilo.css]	
Cluster 5	[GET/wwwforum/index.php?task=view_forum&forum_id=3]	
Cluster 6	[GET/web/tablon_anuncios/css/tablon.css]	
Cluster 7	[GET/acta/asignaturas.css]	
Cluster 8	[GET/alumnos/diegorp/canalplus.html]	
	Entropía	0.00
	Coeficiente de Partición	0.99

Tabla 5.5: *Resultados grupos de páginas visitadas: Conjunto 2*

De esta manera los conjuntos encontrados son mucho más representativos de la navegación real de los usuarios en la página Web de la escuela. Los grupos encontrados refleja a los centros de cada grupo y estos reflejan las páginas que más ha utilizado el usuario dentro del sitio.

5.9.4. Resultados con el conjunto 3

Para el siguiente análisis utilizaremos la última información obtenida del nuevo sitio de la escuela, como ya lo hemos comentado anteriormente en el análisis de las reglas de asociación (Ver sección 4.1.1). Este conjunto lo podemos ver en la tabla 5.2. Lo único que cambia con los experimentos anteriores es la obtención del número de particiones iniciales de los datos, ya que es un nuevo conjunto. Para esto hemos utilizado la misma técnica que en el anterior experimentos, siendo este el Clustering Jerárquico. Para esto hemos utilizado los criterios analizados en la sección 5.3.4.

- **Conjunto 3:** el conjunto de datos iniciales antes del proceso de limpieza era de 98202 transacciones, luego de ese proceso hemos obtenido un conjunto objetivo para el análisis de 15676 transacciones (Ver tabla 5.2).
- **Obtención del número de particiones iniciales:** para este análisis lo único que ha cambiado con los dos anteriores es el número de centroides o clusters. Para este

caso hemos utilizado la misma metodología que en el análisis del conjunto 1. En la tabla 5.6 podemos ver el análisis realizado.

Para este caso la mejor opción sería el α-corte **0.76**, ya que nos entrega el valor más alto en H que es 0.28 que nos indica que es la partición más "estable" y porque si hacemos referencia a otro criterio de partición llamado el *más proxima a la relación*, que indica que será la mejor partición aquella que contenga el intervalo 0.5, que es el caso del α-corte **0.76**. Por lo tanto, según lo analizado anteriormente, el número óptimo para realizar el análisis de clustering es 12.

- **Resultados conjunto 3:** luego de saber el número de los clusters, hemos utilizado este valor para obtener los siguientes grupos de las páginas que el usuario ha utilizado dentro del sitio, estos grupos los podemos ver en la tabla 5.7.

  Como hemos comentado anteriormente, estos grupos reflejan las páginas más utilizada por el usuario dentro del sitio de la Escuela, en este caso podemos ver que los grupos de páginas que más sobresalen se relacionan con el Foro del sitio. En la siguiente sección analizaremos con más profundidad el Clustering del punto de vista difuso para la agrupación de las sesiones de usuarios.

5.9.5. Discusión de los resultados obtenidos en la agrupación de páginas similares

En los dos primeros experimentos relacionados con el conjunto 1 y 2 (Ver tabla 5.2), hemos analizado primero un conjunto de datos que no había sido preprocesado completamente, ya que sólo habíamos eliminados las transacciones idénticas y a partir de eso, obtuvimos grupos relacionados con páginas que contenía principalmente imágenes. Luego de realizar el preprocesamiento completo del conjunto de datos hemos obtenidos conjuntos más cercanos a la realidad de la navegación del usuarios, principalmente grupos de páginas relacionadas principalmente con la sección del Foro del sitio Web.

Luego hemos analizado el conjunto 3, que correspondía al conjunto de datos del nuevo sitio de la Escuela. En ella hemos encontrados resultados similares pero más representativos en el sentido que la gran mayoría de los grupos de páginas también estaban relacionados principalmente con la secciones del foro del sitio de la Escuela.

Podemos decir que según estos resultados que la etapa de preprocesamiento de datos es fundamental a la hora de realizar el clustering de páginas, ya que podemos eliminar los elementos ruidosos y así tener una visión más cercana a lo que realmente sucede en el sitio.

Nivel (α)	**Particiones**	**d2** $[1-\alpha_i]$	**D2** $[D2-d2]$	**H**
0.04	{98,99};{50,51};{48,49};{74,75};{52,53}; {24,25};{86,87};{13,12};{62,63};{2,3}; {36,37};{92,93};{6,7};{56,57};{96,97}; {42,43};{80,81};{18,19};{68,69};{30,31};; {46,47};{76,77};{22,23};{72,73};{26,27}; {88,89};{10,11};{38,39};{84,85};{14,15}; {64,65};{34,35};{94,95};{4,5};{54,55}; {44,45};{78,79};{20,21};{70,71};{28,29}; {90,91};{8,9};{58,59};{40,41};{82,83}; {16,17};{66,67};{32,33};{60,61}	0.96	1.0	0.04
0.28	{99,98,1}	0.72	0.96	0.24
0.48	{50,51,48,49};{74,75,24,25};{86,87,13,12}; {62,63,36,37};{92,93,6,7};{56,57,42,43}; {80,81,18,19};{68,69,30,31};{96,97,2,3}; {52,53,46,47};{76,77,22,23};{72,73,26,27}; {88,89,10,11};{60,61,38,39};{84,85,14,15}; {64,65,34,35};{94,95,4,5};{54,55,44,45}; {78,79,20,21};{70,71,28,29};{90,91,8,9}; {58,59,40,41};{82,83,16,17};{66,67,32,33}	0.52	0.72	0.20
0.76	{99,98,1,50,51,48,49};{86,87,13,12,62,63,36,37}; {92,93,6,7,56,57,42,43};{80,81,18,19,68,69,30,31}; {96,97,2,3,52,53,46,47};{76,77,22,23,72,73,26,27}; {88,89,10,11,60,61,38,39};{84,85,14,15,64,65,34,35}; {94,95,4,5,54,55,44,45};{78,79,20,21,70,71,28,29}; {90,91,8,9,58,59,40,41};{82,83,16,17,66,67,32,33}	0.24	0.52	**0.28**
0.84	{99,98,1,50,51,48,49,74,75,24,25}	0.16	0.24	0.08
0.96	{99,98,1,50,51,48,49,86,87,13,12,62,63,36,37}; {92,93,6,7,56,57,42,43,80,81,18,19,68,69,30,31}; {96,97,2,3,52,53,46,47,76,77,22,23,72,73,26,27}; {88,89,10,11,60,61,38,39,84,85,14,15,64,65,34,35}; {94,95,4,5,54,55,44,45,78,79,20,21,70,71,28,29}; {90,91,8,9,58,59,40,41,82,83,16,17,66,67,32,33}	0.04	0.16	0.12

Tabla 5.6: *Cálculo para la obtención de una partición óptima: conjunto 3*

Cluster	[Centroide]	
Cluster 0	[GET/apps/foro/index.php]	
Cluster 1	[GET/apps/tablon]	
Cluster 2	[GET/usuarios/jmlvega/idragon/formate.css]	
Cluster 3	[GET/apps/foro/index.php?action=foro&idforo=general]	
Cluster 4	[GET/alumnos/diegorp/canalplus.html]	
Cluster 5	[GET/apps/foro/index.php?action=foro&idforo=asignaturas]	
Cluster 6	[GET/js/protWindows/themes/default.css]	
Cluster 7	[GET/apps/foro/index.php?action=foro&idforo=escuela]	
Cluster 8	[GET/alumnos/mlii]	
Cluster 9	[GET/HTTP/1.1]	
Cluster 10	[GET/apps/foro/index.php?action=hebra&idhebra=1939]	
Cluster 11	[GET/apps/foro/index.php?action=foro&idforo=compra]	
	Entropía	0.00
	Coeficiente de Partición	1.00

Tabla 5.7: *Resultados grupos de páginas visitadas: Conjunto 3*

Por consiguiente, las grupos páginas más representativas en la navegación del usuario, como lo hemos comentado anteriormente estan relacionados con los foros del sitios Web de la Escuela.

A continuación veremos un enfoque para una posible agrupación de sesiones de usuarios para una futura creación de perfiles de los usuarios. En la próxima sección nos basaremos en el modelo de datos que hemos planteado en la sección 3.4 y utilizaremos el método timeout para la identificación de las sesiones de usuarios (ver sección 3.4.2.1).

5.10. Clustering difuso de sesiones de usuarios: Caso real

Este enfoque esta relacionado con agrupar las sesiones de usuarios que posean las mismas páginas Web o semejantes. La idea principal es agrupar las sesiones de usuarios que tengan cierto grado de pertenencias unas con otras, dependiendo de las páginas a las cuales se han conectado. De esta manera poder determinar el centroide de las páginas y así identificar diferentes perfiles (Ver Fig. 5.4). Para nuestro caso en especial, como los datos que tenemos corresponden a transacciones realizadas en la Web de la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicaciones (ETSIIT) de la Universidad de Granada, nuestra intención es identificar profesores de alumnos, esto lo analizaremos más adelante en algunos experimentos realizados para este enfoque.

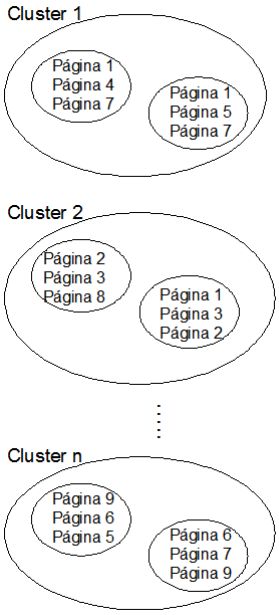

Figura 5.4: *Agrupación de sesiones que posean las mismas páginas Web o similares*

Como ya hemos mencionado anteriormente, la técnica de minería que utilizaremos para agrupar las sesiones de usuarios será el clustering difuso, cuyo modelo de datos recordamos a continuación.

5.10.1. Modelo de datos

Sea $S = \{s_1, s_2, \ldots, s_m\}$ las sesiones de usuarios donde cada sesión está representada por un conjunto de páginas $P = \{p_1, p_2, \ldots, p_n\}$ que representa un espacio d-dimensional (Ver sección 3.4). Sea $UU' = [u_{ij}]$ una matriz *cxm* que representa una partición de *S* en *c* cluster, siendo $V = [v_1, v_2, \ldots, v_c]$ el conjunto de centroides *c* (Ver sección 5.1.1).

Para ello los datos de partida serán el conjunto de sesiones que hemos creado a partir de las transacciones de usuarios que se encuentran en los ficheros log (Ver seccion 3.4). Hemos establecido como medida de similitud entre las sesiones la medida del coseno.

5.10.2. Medida del coseno

La medida del coseno, que es una medida de similaridad de patrones de datos, que permite comparar usuarios o documentos, ya que el coseno mide el ángulo entre dos vectores en un espacio N-dimensional (Ver figura 5.5) [Sal91], [KJNY01].

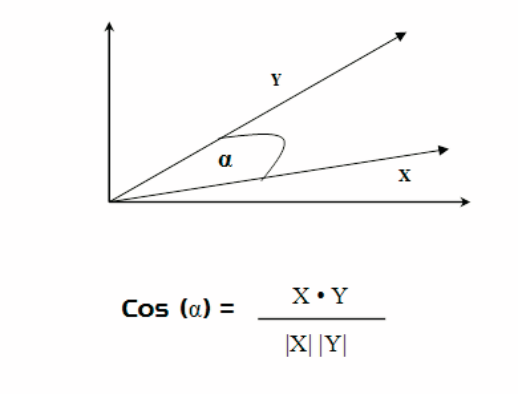

Figura 5.5: *Medida del Coseno*

Cuando los vectores están en la misma dirección el ángulo entre ellos será $\alpha = 0$. Si son perpendiculares el ángulo $\alpha = 90$. Tomando el coseno del ángulo el rango entre los extremos variará entre [0,1].

La medida del coseno se basa en las propiedades de vectores en un espacio euclídeo. La ecuación del coseno es la siguiente, siendo *N* el número de propiedades:

$$d\,(i,k) = \frac{\sum_{j=1}^{N} x_{ij} \cdot x_{kj}}{\sqrt{\sum_{j=1}^{N} x_{ij}^2}\sqrt{\sum_{j=1}^{N} x_{kj}^2}} \tag{5.19}$$

Cuando todas las propiedades son binarias (toman el valor 0 o el 1) se le puede dar una interpretación no geométrica, tomando la sumatoria del denominador como el número de atributos comunes entre las dos instancias y el denominador como la media geométrica del número de atributos que posee x_i y x_k, la medida se podría interpretarse como la relación de atributos que poseen ambas instancias.

Nosotros hemos dado una definición de la sesión de usuario en secciones anteriores (ver sección 5.6) y también hemos definido el umbral de tiempo para considerar una entrada perteneciente a una sesión de usuario, de esta manera hemos considerado como umbral de tiempo 30 minutos.

Nos centraremos en la medida del coseno para determinar la medida de similitud entre dos sesiones de usuarios para agruparlas según las páginas más similares entre ellas. Por lo tanto, definiremos la medida del coseno en función de dos sesiones de usuarios S_i y S_k, siendo *N* el número válido de *URL* como sigue [KJNY01]:

$$S_{1,kl} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} S_i^k \cdot S_j^l}{\sqrt{\sum_{i=1}^{N} S_i^k} \sqrt{\sum_{j=1}^{N} S_j^l}} \tag{5.20}$$

Siendo S_i^k y S_j^l vectores de similaridad entre las sesiones de usuario, y con estos vectores de similaridad podremos calcular el valor del coseno entre ambas. La ecuación del coseno no hace una diferencia en la representación sintáctica de las páginas Web o URL. Por ejemplo, si en una sesión puede tener la página $\{$*curso/alumno231*$\}$ y en otra sesión la página $\{$*curso/alumno234*$\}$, o tal vez las páginas $\{$*proyectos/webmining*$\}$ y $\{$*actas/electromagnetismo*$\}$, en ambas situaciones recibirá como valor de similitud cero, de acuerdo a la ecuación que hemos planteado anteriormente del coseno.

5.10.3. Medida del coseno extendido

Para solucionar este problema de la representación sintáctica de dos URL, se ha delimitado definiendo una medida alternativa que toma en cuenta la sintaxis de dos *URL* [KJNY01]. Esta medida alternativa la podemos ver en la siguiente ecuación:

$$S_n(i,j) = min\left(1, \frac{|p_i \cap p_j|}{max\left(1, max\left(|p_i|, |p_j|\right) - 1\right)}\right) \tag{5.21}$$

Donde p_i indica el camino desde la raíz al nodo correspondiente al *i*-ésimo URL y $|p_i|$ indica el largo de este camino o el número de aristas incluidas en el camino. Ahora la similaridad entre dos sesiones, incorporando la similitud de la sintaxis de dos URL, es definida por la asociación de todos los atributos de las URL y su similitud entre dos sesiones como sigue:

$$S_{2,kl} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} s_i^k s_j^l s_n(i,j)}{\sqrt{\sum_{i=1}^{N} s_k^i} \sqrt{\sum_{j=1}^{N} s_j^l}} \tag{5.22}$$

Esta medida la utilizaremos para obtener mejores resultados que utilizando la medida del coseno, ya que mejora la representación sintáctica que pueda haber entre las páginas

de cada sesión de usuario. A continuación veremos un conjunto de experimentos relacionados con los dos enfoques planteados anteriormente en esta sección.

5.10.4. Caso real: Estudio preliminar

El experimento que realizamos fue sobre los datos extraídos del servidor web de la E.T.S. de Ingenierías Informática y de Telecomunicación de la Universidad de Granada (http://etsiit.ugr.es).

El objetivo de este experimento fue obtener grupos de usuarios agrupados por la misma IP y así obtener información sobre cuales eran las preferencias de los usuarios que se han conectado más de alguna vez desde la misma IP en diferentes sesiones.

El tipo de archivo que se procesó, fue el archivo Web log del tipo ECLFF (Extended Common Log File Format). Este archivo consta de 6953 transacciones, al cuál denominaremos *conjunto preliminar*. Lo primero que se debe hacer en cualquier proceso de minería, si los datos no están preparados para el análisis, es realizar un preprocesamiento de los datos.

Este preprocesamiento consistió en obtener un fichero separado por transacciones de usuarios y también se han eliminado todas las transacciones repetidas. Una vez realizada el preprocesamiento, se identificaron las sesiones de los usuarios a través del método timeout, el cual explicamos en sesiones anteriores, siendo el umbral de corte máximo para el inicio de una nueva sesión de 30 minutos.

Una vez terminada esta etapa, se identificaron las sesiones de los usuarios, la cuales fueron 73. En estas sesiones el tiempo promedio de navegación de los usuarios fue de 16.61 minutos aproximadamente. La sesión más larga tuvo una duración aproximada de 6,5 horas y la más corta fue de cero. A continuación mostraremos un gráfico relacionado con el tiempo de cada sesión de usuario que fue identificada (Ver Figura 5.6).

Cabe destacar que las entradas de los usuarios al sitio web fueron realizadas en diferentes días y a diferentes horas. También dentro de cada sesión se presentaron múltiples páginas que son de bajo grado de importancia al momento de identificar las preferencias de los usuarios. Por ejemplo existían páginas relacionadas con los frame de la Web dinámicas, otras relacionadas con imágenes, y otras que sólo se diferenciaban de la no existencia de un agente de navegación. Muchas de estas entradas se realizaban en una misma hora en el mismo instante, en la figura 5.7 podemos ver un pequeño ejemplo de lo comentado.

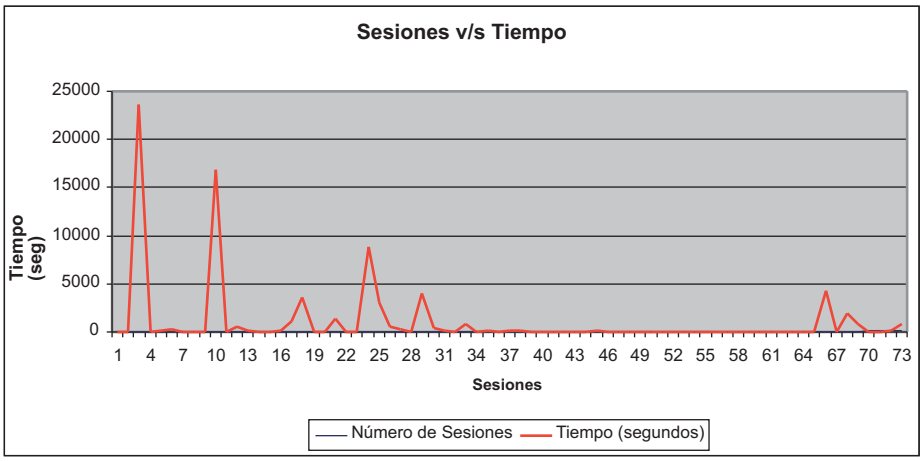

Figura 5.6: *Relación sesiones versus tiempo por sesión de usuario: caso 1*

```
33.red-83-33-8.dynamicip.rima- tde.net  [18/Jun/2006:07:41:14+0200]  GET/ graficos/2002/ pie _r6_c1. gif HTTP/1.1
33.red-83-33-8.dynamicip.rima- tde.net  [18/Jun/2006:07:41:14+0200]  GET/graficos/2002/pie _r8_c2. gif HTTP/1.1
33.red-83-32-6.dynamicip.rima- tde.net  [18/Jun/2006:07:41:14+0200]  GET/graficos/2002/pie_r6_c1. gif HTTP/1.1
33.red-83-32-6.dynamicip.rima- tde.net  [18/Jun/2006:07:41:14+0200]  GET/graficos/2002/pie _r3_c7. gif HTTP/1.1
```

Figura 5.7: *Ejemplo de entradas duplicadas*

Para los siguientes análisis hemos utilizado la información obtenida del nuevo sitio de la Escuela. Esta información ya ha sido preprocesada (ver secciones 4.4.1), así que utilizaremos el conjunto objetivo encontrado que consta de 15676 transacciones, al cual es *conjunto 3* (ver tabla 5.2). De este conjunto hemos obtenido 2780 sesiones de usuarios, con un promedio de tiempo por sesión de 360.72 segundos aproximadamente 6.012 minutos por sesión. El valor máximo de tiempo de navegación es una sesión de 6.27 horas aproximadamente y el tiempo mínimo fue de cero segundos. En la figura 5.8 podemos ver un gráfico que representa la relación de sesiones versus el tiempo de sesión de usuario.

Luego de haber realizado un pequeño análisis relacionado con la identificación de las sesiones de usuario y su tiempo de navegación, ahora veremos el enfoque relacionado con la agrupación de las sesiones por páginas similares.

5.10.5. Caso real: Experimentación del clustering para sesiones de usuarios por páginas similares

Ahora nos centraremos en el análisis del enfoque propuesto relacionado con las agrupaciones de sesiones de usuarios con páginas similares. Enfocaremos este punto utilizan-

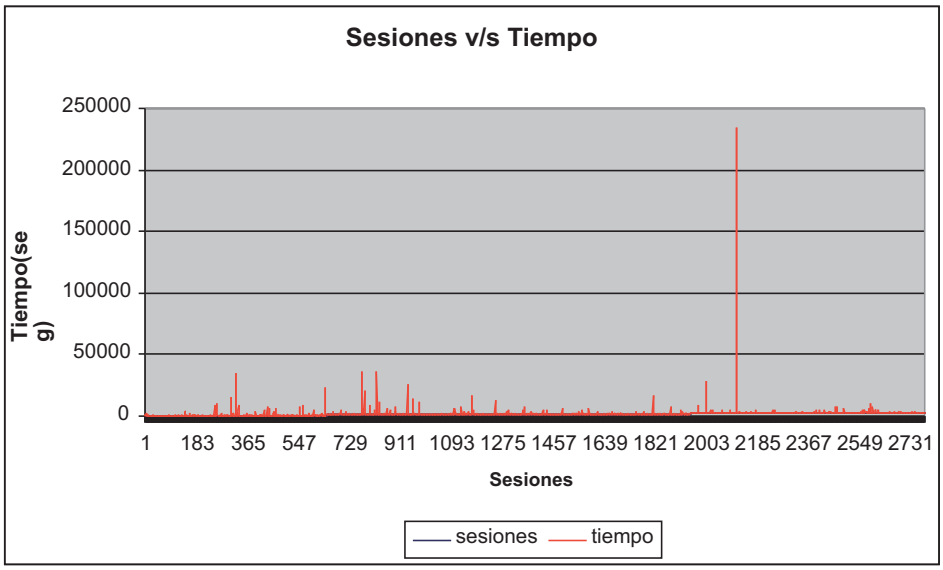

Figura 5.8: *Relación sesiones versus tiempo por sesión de usuario: Conjunto 3*

do el algoritmo del clustering difuso c-medias que hemos planteados anteriormente (Ver sección 5.11).

A continuación enumeraremos las páginas más frecuentes que hemos encontrado durante nuestro análisis con una frecuencia superior a 50, siendo 12920 la cantidad total de páginas que contiene el archivo (ver tabla 5.8).

A partir de esta información, hemos realizado el clustering de las sesiones que contengan las páginas más similares. Con esta técnica del clustering difuso c-medias obtendremos diferentes clusters y dentro de ellos estarán las diferentes sesiones de los usuarios más semejantes entre ellas. Algunas sesiones podrán ser parte en más de un cluster pero con diferente grado de pertenencia.

5.10.5.1. Características generales del experimento

- **Objetivo general:** el objetivo principal de estos análisis que realizaremos están enfocados en la obtencion de grupos de sesiones de usuarios por páginas similares.
- **Conjuntos de datos:** Para realizar los diferentes experimentos analizaremos diversos conjuntos de datos. En la tabla 5.9 podemos ver una descripción de los conjuntos de datos que utilizaremos para el análisis.

Página	**Frecuencia**
GET/wwwforum	2622
GET/web/tablon_anuncios	1503
GET/css/estilo.cssHTTP	540
GET/wwwforum/index.php?task=view_forum&forum_id=1	326
GET/css/wwwforum.css	292
GET/wwwforum/index.php?task=view_forum&forum_id=3HTTP/1.1	280
GET/alumnos/diegorp/canalplus.html	151
GET/web/tablon_anuncios/css/tablon.css	149
GET/HTTP/1.1	123
GET/alumnos/diegorp/canal.css	120
GET/wwwforum/index.php?task=view_forum&forum_id=2	114
GET/web/tablon_anuncios/css/estilo.css	92
GET/usuarios/kiakio/retoques/retoques.htm	81
GET/HEAD/HTTP/1.1	81
GET/acta/asignaturas.css	79
GET/alumnos/shin/sakura.htm	77
GET/usuarios/jmlvega/idragon/formate.css	72
GET/web/tablon_anuncios/?seccionSeleccionada=4	72
GET/acta/principal.css	64
GET/proyectos/silviaacid/basedatos/estilos.css	62
GET/web/tablon_anuncios?seccionSeleccionada=4	57
GET/alumnos/shin/marmboy.htm	56
GET/usuarios/other/esc//themes/default/style.css	55
GET/alumnos/mlii/Algoritmo.htm	54
GET/alumnos/mlii/Steve%20Jobs.htm	53
GET/estad-alum/usage_200512.html	52
Otras páginas	< 50

Tabla 5.8: *Frecuencias de páginas*

Conjuntos De Datos	Entrada de datos originales	Entradas de datos preprocesadas	N° Sesiones
Conjunto 1	100900	12910	2024
Conjunto 2	98202	15676	2780

Tabla 5.9: *Resumen de los conjuntos de datos para el análisis del Clustering de sesiones de usuarios*

- **Modelo de datos:** sea un conjunto de páginas Web $P = \{p_1, p_2, ..., p_n\}$, donde p_n corresponde a una página Web. Sea $S = \{s_1, s_2, ..., s_m\}$ las sesiones de usuarios donde cada sesión está representada por un conjunto de página *P* (Ver sección 5.5.1).

- **Medida:** En los diferentes experimentos utilizaremos tanto la medida del coseno como la medida del coseno extendido, principalmente para determinar cual de las dos medidas es la más óptima al momento de obtener los diferentes grupos o clusters.

- **Técnica utilizada:** utilizaremos la técnica de clustering difuso y el algoritmo a utilizar es el *c-medias difuso*. Con esto queremos obtener grupos más cercanos a la realidad de la navegación que realizan los usuarios por la Web.

5.10.5.2. Resultados con el conjunto 1

Este conjunto lo hemos separado en dos casos, el **caso a** esta relacionado con la medida del coseno y el **caso b** con la medida del coseno extendido.

- **Conjunto 1a:** el conjunto que analizaremos corresponde al conjunto 1 (Ver tabla 5.9). Este archivo ha pasado por una etapa de preprocesamiento, donde se eliminaron los elementos conflictivos del archivo como son las imágenes, javascripts, entradas repetidas, entre otros elementos. Para este caso hemos analizado el archivo anteriormente descrito, en el cual tenemos 2024 sesiones de usuarios.

- **Medida:** la medida a utilizar es la medida del *coseno* (Ver sección 5.7.2).

- **Resultados conjunto 1a:** en la tabla 5.10 podemos ver algunos resultados obtenidos después del análisis, veremos las sesiones más representativas de algunos de los clusters y las páginas que se relacionan con las sesiones. El valor que podemos

ver entre ”( )” corresponde al grado de pertenencia de la sesión a la sesión centroide. Por ejemplo, las sesión 39 tiene de un grado de pertenencia 1.0 a la sesión centroide.

N° Cluster	**Sesiones (Grado de Pertenencia)**	**Sesión Centroide**
Cluster 0	39 (1.0) 37(1.0) 77(1.0) 166 (1.0)	GET/depar/ccia/bd1/bd17.ppt GET/depar/ccia/bd1/bd14.ppt GET/depar/ccia/robotica/director.htmlHTTP GET/depar/ccia/neurocomputacion/Material GET/depar/ccia/neurocomputacion GET/depar/ccia/bd1/sqlconsulta.ppt
Cluster 3	150 (1.0) 511 (0.954) 28 (0.954) 31 (0.954)	GET/acta/info_asignatura.php?id=197 GET/acta/asignaturas.css GET/acta/info_asignatura.php? GET/acta/principal.css GET/css/estilo.cssHTTP/1.1
Cluster 6	103 (0.99) 109 (0.99) 147 (0.99) 118 (0.99) 152 (0.99) 556 (0.99) 805 (0.99) 156 (0.99)	 GET/estad-alum/usage_200512.html GET/alumnos/shin/marmboy.htm GET/alumnos/diegorp/canalplus.html GET/alumnos/mu01/guerraSoftware.html GET/alumnos/diegorp/canalplus.html
Cluster 8	71 (0.997) 775 (0.997) 949 (0.997) 630 (0. 994) 642 (0. 994) 749 (0. 994)	 GET/css/wwwforum.cssHTTP GET/wwwforum GET/wwwforum/index.php?task=buscar _mensajes

Tabla 5.10: *Clusters de Sesiones por páginas con la medida del coseno: Conjunto 1a*

Hemos seleccionado estos resultados para mostrar claramente que estas sesiones corresponden a alguno de los dos perfiles que queríamos identificar, al perfil de los alumnos o al perfil del profesor. Todos estos resultados han sido validados con las medidas de entropía y coeficiente de partición, siendo estos valores de 0.068 y 0,988 respectivamente. Estos valores nos indican que la agrupación se realizo correctamente.

Como se muestra en la tabla 5.10, podemos ver y suponer que los dos primeros clus-

ter corresponden al perfil de los profesores, ya que las páginas están relacionadas a consultas de actas y a un departamento específico de la universidad. Y los otros dos clusters o grupos corresponde claramente al perfil de alumnos. Veremos a continuación una extensión de este análisis, basado principalmente en una extensión de medida de similaridad del coseno.

Anteriormente hemos analizado con la ecuación del coseno sin la extensión sintáctica de las páginas Web. Así que haremos una comparación entre ambas para ver las diferencias evidentes al momento de agrupar las sesiones.

- **Conjunto 1b:** en este análisis hemos utilizado el mismo conjunto de datos que el anterior experimento, que corresponde al conjunto 1.
- **Técnica utilizada:** la medida de coseno extendido (Ver sección 5.7.3).
- **Resultados conjunto 1b:** en la tabla 5.11 podemos ver los resultados relacionados con este análisis, veremos las sesiones más representativas y a continuación haremos la comparación con los resultados anteriores. Al igual que el anterior experimento, las sesiones están acompañadas del grado de pertenencia a la sesión centroide.

  Los resultados son muy similares en relación a las sesiones más representativas de cada cluster, sólo han subido unas cuantas décimas, pero podemos decir que sesiones que forman parte de algunos clusters el grado de semejanza han aumentado y así hemos obtenido mejores resultados con respecto al anterior experimento. Los valores de entropía para este análisis fue de 0.037 y el coeficiente de partición fue de 0.998. Estos valores nos indica que esta agrupación es levemente mejor que la anterior realizada.

A continuación realizaremos otros experimentos relacionados con la información obtenida del nuevo sitio de la Escuela de Informática y Telecomunicaciones de la Universidad de Granada, analizaremos el conjunto de datos 2 (Ver tabla 5.9)

5.10.5.3. Resultados con el conjunto 2

Al igual que en los anteriores experimentos, hemos separado este análisis en dos casos. El **caso a** utilizaremos la medida del coseno y en el **caso b** la medida del coseno extendido.

N° Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 0	39 (1.0) 37(1.0) 77(1.0) 166 (1.0)	GET/depar/ccia/bd1/bd17.ppt GET/depar/ccia/bd1/bd14.ppt GET/depar/ccia/robotica/director.htmlHTTP GET/depar/ccia/neurocomputacion/Material GET/depar/ccia/neurocomputacion GET/depar/ccia/bd1/sqlconsulta.ppt
Cluster 3	150 (1.0) 511 (0.984) 28 (0.984) 31 (0.984)	GET/acta/info_asignatura.php?id=197 GET/acta/asignaturas.css GET/acta/info_asignatura.php? GET/acta/principal.css GET/css/estilo.cssHTTP/1.1
Cluster 6	103 (0.989) 109 (0.989) 147 (0.989) 118 (0.989) 152 (0.989) 556 (0.989) 805 (0.989) 156 (0.989)	GET/estad-alum/usage_200512.html GET/alumnos/shin/marmboy.htm GET/alumnos/diegorp/canalplus.html GET/alumnos/mu01/guerraSoftware.html GET/alumnos/diegorp/canalplus.html
Cluster 8	71 (0.997) 775 (0.997) 949 (0.997) 630 (0. 997) 642 (0. 99) 749 (0. 997)	GET/css/wwwforum.cssHTTP GET/wwwforum GET/wwwforum/index.php?task=buscar_ mensajes

Tabla 5.11: *Clusters de Sesiones por páginas por coseno extendido: Conjunto 1b*

- **Conjunto 2a:** el conjunto a analizar corresponde en este caso al conjunto 2 (Ver tabla 5.9).

- **Medida utilizada:** la medida a utilizar es la medida del *coseno* (Ver sección 5.7.2).

- **Resultados conjunto 2a:** en las tablas 5.12 y 5.13 podemos ver los diferentes cluster, relacionando las sesiones de usuarios con la sesión centroide. Al igual que los experimentos anteriores, cada sesión tiene un cierto grado de pertenencia a la sesión centroide, este valor lo podemos encontrar dentro de la tabla entre paréntesis.

N° Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 0	2 (0.75) 437 (0.90) 508 (0.90) 512 (0.85)	GET/apps/tablon GET/apps/foro/index.php GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=foro&idforo=general GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=hebra&idhebra=1937 GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=hebra&idhebra=1916H
Cluster 3	21 (0.97) 65 (0.97) 6 (0.97) 51(0.97) 136 (0.97) 13 (0.97) 68 (0.85) 569 (0.85)	 GET/js/protWindows/themes/default.css GET/apps/foro/index.php GET/apps/tablon GET/page.php?pageid=departamentos GET/apps/foro/index.php?action=hebra&idhebra=1583 GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php?action=foro&idforo=general

Tabla 5.12: Clusters 0 y 3 de Sesiones por páginas utilizando la medida del coseno: Conjunto 2a

Hemos tomado estas sesiones de cada cluster principalmente para notar la diferencia en los grados de pertenencia de estos con respecto al grupo que pertenece, ya que en las tablas 5.9 y 5.10 podemos ver una real mejoría en los resultados obtenidos en relación a la pertenecía de cada sesión al clusters que corresponde.

Ahora utilizaremos el mismo conjunto de datos para realizar el análisis con coseno extendido, al igual que lo hemos hecho con el conjunto de información anterior.

- **Conjunto 2b:** Utilizaremos el mismo conjunto que el anterior análisis, el que corresponde al conjunto 2.
- **Medida utilizada:** la medida a utilizar es la medida del *coseno extendido* (Ver sección 5.7.3).
- **Resultados conjunto 2b:** podemos ver los resultados obtenidos en las tablas 5.14 y 5.15. También podemos decir que existen mejoras en los resultados, del punto

N° Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 8	11 (0.80) 204 (0.86) 254 (0.66) 273 (0.87)	GET/apps/foro/index.php?action=hebra&idhebra=1916 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=hebra&idhebra=1892 GET/apps/foro/index.php?action=hebra&idhebra=1939 GET/apps/foro/index.php?action=foro&idforo=general GET/apps/tablon GET/apps/foro/index.php?action=hebra&idhebra=1922 GET/apps/foro/index.php?action=foro&idforo=deportes GET/js/protWindows/themes/alphacube.css GET/apps/foro/index.php?action=hebra&idhebra=1752
Cluster 11	147 (0.80) 79 (0.82) 409 (0.89) 447 (0.92) 700 (0.82)	GET/page.php?pageid=rss_base GET/apps/tablon GET/apps/foro/index.php GET/js/protWindows/themes/default.css GET/guias/actual/Guia.pdf GET/guias/actual/Indice.pdf GET/guias/actual/Presentacion.pdf GET/page.php?pageid=infocentro GET/page.php?pageid=descargas GET/apps/descargas GET/apps/descargas/styles/descargas.css GET/apps/descargas/index.php?id=guias GET/apps/descargas/index.php?id=secretaria GET/page.php?pageid=webinfo GET/page.php?pageid=wemap GET/apps/foro/index.php?action=hebra&idhebra=1935 GET/page.php?pageid=rriiDestFran GET/page.php?pageid=rriiExtranjero

Tabla 5.13: *Clusters 8 y 11 de Sesiones por páginas utilizando la medida del coseno: Conjunto 2a*

de vista del grado de pertenencia de las sesiones a los diferentes clusters encontrados en relación a los resultados obtenidos en el anterior análisis (ver tabla 5.12 y 5.13). Esto quiere decir que la medida del coseno extendido nos entrega una mejor representación de los datos en los clusters.

N° Cluster	**Sesiones (Grado de Pertenencia)**	**Sesión Centroide**
Cluster 0	2 (0.95) 437 (0.98) 508 (0.98) 512 (0.96)	GET/apps/tablon GET/apps/foro/index.php GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=foro&idforo=general GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=hebra&idhebra=1937 GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=hebra&idhebra=1916H
Cluster 3	21 (1.00) 65 (1.00) 6 (1.00) 51(1.00) 136 (1.00) 13 (1.00) 68 (0.939) 569 (0.939)	 GET/js/protWindows/themes/default.css GET/apps/foro/index.php GET/apps/tablon GET/page.php?pageid=departamentos GET/apps/foro/index.php?action=hebra&idhebra=1583 GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php?action=foro&idforo=general

Tabla 5.14: Clusters 0 y 3 de Sesiones por páginas utilizando la medida del coseno extendido: Conjunto 2b

Uno de los objetivos principales para este análisis era realizar una agrupación de sesiones de usuario. Hemos logrado obtener buenos resultados en la agrupación de las sesiones de usuario, obteniendo 12 grupos o cluster que serán la base principal para la creación de los perfiles de los usuarios. De esta manera, en el siguiente capítulo podremos relacionar e identificar los diferentes clusters o grupos de sesiones de usuarios a los diferentes perfiles de usuario.

N° Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 8	11 (0.95) 204 (0.98) 254 (0.98) 273 (0.99)	GET/apps/foro/index.php?action=hebra&idhebra=1916 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=hebra&idhebra=1892 GET/apps/foro/index.php?action=hebra&idhebra=1939 GET/apps/foro/index.php?action=foro&idforo=general GET/apps/tablon GET/apps/foro/index.php?action=hebra&idhebra=1922 GET/apps/foro/index.php?action=foro&idforo=deportes GET/js/protWindows/themes/alphacube.css GET/apps/foro/index.php?action=hebra&idhebra=1752
Cluster 11	147 (0.957) 79 (0.957) 409 (0.957) 447 (0.927) 700 (0.927)	GET/page.php?pageid=rss_base GET/apps/tablon GET/apps/foro/index.php GET/js/protWindows/themes/default.css GET/guias/actual/Guia.pdf GET/guias/actual/Indice.pdf GET/guias/actual/Presentacion.pdf GET/page.php?pageid=infocentro GET/page.php?pageid=descargas GET/apps/descargas GET/apps/descargas/styles/descargas.css GET/apps/descargas/index.php?id=guias GET/apps/descargas/index.php?id=secretaria GET/page.php?pageid=webinfo GET/page.php?pageid=wemap GET/apps/foro/index.php?action=hebra&idhebra=1935 GET/page.php?pageid=rriiDestFran GET/page.php?pageid=rriiExtranjero

Tabla 5.15: *Clusters 8 y 11 de Sesiones por páginas utilizando la medida del coseno extendido: Conjunto 2b*

5.10.6. Discusión de los resultados obtenidos en la agrupación de sesiones de usuarios

El objetivo de la experimentación era obtener diferentes grupos de usuarios con ciertas características o intereses comunes, para eso aplicamos el algoritmo c-medias difuso utilizando dos distintas medias, una la medida del coseno y la otra la del coseno extendido. La utilización de estas medidas era ver cual de las dos nos entregaba mejores agrupamientos más cercanas a la navegación que realiza el usuario por la Web.

Al revisar los resultados obtenidos en los diferentes experimentos, nos podemos dar cuenta que en cada uno de ellos la medida del coseno extendido fue la que siempre nos entregó mejores resultados con relación a los obtenidos con la medida del coseno.

También nos hemos dado cuenta que los grupos o clusters con la medida del coseno extendido, los elementos de los grupos eran mucho más similares entre ellos mismos, en el sentido que cada elemento poseía un grado de pertenencia alto al grupo y eso se refleja principalmente en que los grupos pueden representan diferentes perfiles de los usuarios que navegan por el sitio de la escuela.

Este último punto lo retomaremos en el siguiente capítulo donde haremos una relación entre los resultados obtenidos en este capítulo sobre la agrupación de las sesiones de usuarios con la obtención de los diversos perfiles de usuarios.

5.11. Conclusiones

En la Minería Web de Uso, una de las técnicas más aplicadas es el clustering. Hay varios algoritmos de clustering, pero uno de los más destacados es el algoritmo c-medias, que nosotros hemos aplicado además en su versión difusa: el algoritmo c-medias difuso, que nos permite manejar clusters de forma flexible.

Dentro de la metodología desarrollada, hemos realizado una preparación previa de los datos para su análisis. De esta manera, tras realizar una tarea de preprocesamiento para limpiar las entradas de los datos irrelevantes para nuestro análisis, hemos identificado las sesiones de usuarios e identificado las visitas de las sesiones de usuario a través de los clisckstreams. Además, para optimizar los grupos a obtener, previo al clustering de páginas y sesiones hemos obtenido una partición inicial de datos mediante un clustering jerárquico.

Los agrupamientos realizados han sido, por un lado sobre páginas similares para de-

terminar cuales eran las páginas más representativas en la navegación del usuario por el sitio de la escuela. Y por otro lado, hemos agrupado las sesiones de usuarios y hemos aplicado el algoritmo de c-medias difuso.

Además, hemos completado esta metodología con una validación de las agrupaciones obtenidas mediante medidas tales como el coeficiente de partición y la entropía.

En cuanto a la experimentación en nuestro caso real del sitio web de la E.T.S.I.I.T. de la Universidad de Granada, hemos conseguido determinar el comportamiento de los usuarios en la Web, pudiendo distinguir profesores de alumnos. Para ello, se han aplicado diferentes medidas de similitud como son el coseno y el coseno extendido, siendo ésta última la que aporta mejores resultados.

En el siguiente capítulo nos centraremos en el estudio de los perfiles de usuarios, identificando dichos perfiles con los grupos obtenidos en este capítulo.

Capítulo 6

Perfiles de usuario y lógica difusa: Modelo de representación en XML. Modelo de obtención de perfiles de usuario

En este capítulo estudiamos la construcción de perfiles de usuario donde se recogen el comportamiento o preferencias del usuario durante su navegación y así poder identificar diferentes grupos sociales y/o demográficos.

Para ello analizaremos de forma general el proceso de personalización, veremos definiciones relacionadas con el perfil de usuario que existen en la literatura y plantearemos una nueva representación de perfiles de usuarios en XML. También damos un nuevo modelo de obtención de perfiles basado en los procesos de minería vistos en los capítulos anteriores, derivando los perfiles a partir de los grupos demográficos obtenidos en los procesos de clustering.

6.1. El proceso de personalización

La personalización se refiere generalmente a ”la capacidad de proporcionar información diferente en función de los diferentes estereotipos definidos para clasificar a los usuarios sobre la base del conocimiento de sus preferencias y comportamientos a la hora

de interactuar" [Hag99].

En el caso de la Web, el proceso de personalización obtiene el conocimiento a partir de un conjunto de acciones realizadas por el usuario en su navegación, es decir que dicho proceso puede basarse en conocimiento adquirido a partir de un proceso previo de Minería Web de Uso.

La meta de la personalización basada en la Web es recomendar un conjunto de objetos para los usuarios actuales (activos); estos conjuntos pueden estar constituidos por enlaces, anuncios, textos, productos o servicios hechos a la medida de los usuarios según sus gustos o preferencias obtenidos a partir de sus patrones de navegación.

En general, las técnicas de Minería Web de Uso tales como las reglas de asociación ó clustering, pueden ser usadas dentro del proceso de personalización para obtener tendencias en el comportamiento de navegación o para identificar grupos de usuarios con ciertas características similares, y así generar las recomendaciones.

Todos los procesos de personalización de la Web basados en la Minería Web de Uso consisten en 3 fases: preparación y transformación de los datos, descubrimiento de patrones y recomendaciones (Ver figura 6.1). La **fase de preparación de datos** transforma los archivos del servidor Web en la forma intermedia [Jus04] adecuada para poder ser procesados luego por alguna técnica de minería. Esta fase también incluye la integración de datos de múltiples recursos o fuentes, como aplicaciones del servidor, bases de datos de "backend" y el contenido del sitio. Luego en la siguiente fase, diversas técnicas de minería pueden aplicarse a las transacciones de datos en la búsqueda de patrones. Los resultados de la **fase de minería** pueden ser transformados en perfiles de usuarios, estos son adecuados para ser utilizados en la fase de recomendación. El motor de **recomendación** considera la sesión activa de usuarios en conjunción con el descubrimiento de patrones y de los perfiles para proveer una personalización de contenido del sitio [Mob05].

Una posible forma de clasificar los tipos de personalización es dividirla en las siguientes categorías: *reconocimiento por nombre, la customización* y *adaptativa* [DGR03]:

- **El reconocimiento por nombre:** es la forma más básica, en la que el usuario se reconoce al llegar al sitio Web, normalmente a través de un formulario de identificación de usuario y contraseña.

- **La customización:** es normalmente llamada "personalización de caja de opción", porque el usuario define los parámetros de funcionamiento del servicio, seleccionando sus preferencias de una lista de cajas de opciones. Este tipo de personaliza-

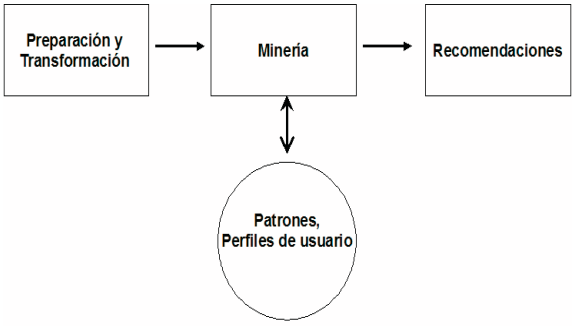

Figura 6.1: *Etapas del Proceso de Personalización*

ción es frecuentemente usada por los sistemas de hipermedia adaptables, en los que las opciones seleccionadas se almacenan en el perfil de usuario [MPR00].

- **La personalización adaptativa:** se realiza una personalización avanzada, seleccionando el contenido de las páginas que se visitan, de acuerdo con las acciones del usuario anteriormente realizadas en ese servicio. Esta información es procesada y guardada en su perfil de usuario durante la interacción del usuario con el sistema, y posteriormente analizada para adaptar la visita [DGR03].

6.1.1. Trabajos previos

En la literatura podemos encontrar algunos trabajos orientados a dar una visión general de la personalización en la web a través de herramientas y aplicaciones disponibles para llevar a cabo procesos de este tipo [EV03], [PPPS03]. En la mayoría de ellos se revela que las técnicas de minería más utilizadas dentro del proceso de personalización son el clustering y las reglas de asociación [Mob05], [EV03], [MCS00].

Por ejemplo en [GG02] se presenta un sistema de recomendación mediante la integración de técnicas de personalización y de minería a través de clustering, basadas en el comportamiento de usuarios junto a estrategias de marketing aplicado a las tiendas virtuales.

Quizás uno de los enfoques más representativos para la personalización de la web desde la minería está propuesto por Mobasher [MCS00], [Mob05]. El objetivo es capturar y modelar los patrones del comportamiento y perfiles de usuarios que interactúan con un sitio Web. Estos trabajos utilizan las técnicas de clustering y reglas de asociación tradicionales para obtener los patrones del comportamiento y los perfiles de usuarios.

El enfoque que nosotros adoptamos en esta tesis sigue las líneas generales de los trabajos de Mobasher, pero con la utilización de las técnicas con lógica difusa que nos permitirán una mayor flexibilidad tanto en el manejo de información como en la interpretación de los resultados. Hay otras diferencias entre el enfoque de Mobasher y el nuestro, ya que la agrupación que realiza el primero para determinar los perfiles de usuarios es a través de *páginas vistas*, mientras que nosotros lo hacemos a través de sesiones de usuario.

Además, nosotros planteamos un modelo para la obtención de los perfiles de usuario, utilizando la definición dada en [MBKV+02], que nos servirá para realizar la representación de los perfiles. Esta representación las almacenaremos utilizando el *lenguaje XML* que nos facilitará el manejo en un futuro de ellos.

6.2. Perfiles de usuario

Como ya hemos visto en la sección anterior, los perfiles de usuario se pueden enmarcar dentro del proceso de personalización como una estructura de almacenamiento de información sobre las preferencias del usuario obtenidas, en nuestro caso, a partir de un proceso de explotación de los ficheros log.

En general podemos definir un perfil como *una colección de datos acerca de un usuario*. Por ejemplo, éste puede ser rellenado con información relacionada con el método de conexión utilizado, el terminal utilizado, patrones de comportamiento e intereses del usuario. Esta información se puede completar con datos obtenidos directamente de los usuarios a través por ejemplo de formularios que soliciten información tal como la edad, la residencia habitual, e-mail, teléfono fijo o móvil, etc. Parte de la información del perfil de usuario puede ser estática, como la fecha de nacimiento, el nombre, etc.; y normalmente es introducida manualmente, de una sola vez por el usuario. Otra sin embargo es dinámica, como por ejemplo los intereses del usuario, que cambian y por consiguiente es aconsejable que sean determinados automáticamente. Esto significa que para obtener 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 [DGR03].

Hay perfiles de usuario que pueden almacenar por ejemplo, todo lo relacionado a la información relativa al ambiente de escritorio incluyendo el contenido del menú inicio, los iconos que aparecen en el escritorio y otras características acerca del ambiente GUI que a

los usuarios les está permitido personalizar. Estos perfiles son almacenados en servidores centrales y al identificarse en el sitio Web se descargan de forma local y así está disponible la información de configuración del usuario. A este tipo de perfiles se los denomina *roaming profile*. Las ventajas que puede presentar estos tipos de perfiles son que se necesita una mínima configuración, gran accesibilidad, los ajustes que realiza el usuario se mantienen y el almacenamiento centralizado para facilitar el backup, la recuperación y la administración. Las desventajas un pequeño incremento en el tráfico de la red, susceptible a la corrupción, y puede ralentizar el tiempo de conexión [SSTK07].

6.2.1. Trabajos previos

En la literatura podemos encontrar diferentes definiciones sobre los perfiles de usuarios. Por ejemplo en [MCS00] define un perfil de usuario como un grupo de páginas visitadas y donde *"cada uno de los grupos de URL puede ser mirado como un perfil de usuario virtual indicando qué tan diversos pueden ser los grupos que acceden a un conjunto de link en el sitio dentro de sus transacciones respectivas"*. Nasraoui también define el perfil de usuario como *"la información acerca de los atributos demográficos de los usuarios y preferencias que son obtenidas explicitamente o implicitamente"* [NFJK99], [NKJF00], [NK00], [NK02].

También en [VP07] se plantea un perfil de usuario a partir de un modelo de comportamiento de navegación. En dicho modelo, se utiliza tres variables para modelar el comportamiento: secuencias de páginas visitadas, el contenido de las secuencias de páginas y el tiempo que permanece en la página. Este modelo lo ha definido dentro de un vector de comportamiento definido: sea un vector $u = [(p_1, t_1), ..., (p_n, t_n)]$, donde el par (p_i, t_i) representa la $i^{th}4$ página visitada (p_i) en un porcentaje de tiempo de espera en una sesión (t_i).

Nosotros nos basaremos en el perfil descrito en [MBKV$^+$02], donde se distinguen dos tipos de perfiles: los *perfiles simples*, que son representados por un conjunto de términos extraídos de documentos estimados interesantes para ese usuario, y los *perfiles extendidos* que contienen conocimiento adicional acerca del usuario, tales como el nivel educativo, de grupo de edades, de idioma, el país, entre otras. La justificación para la elección de esta definición de perfil es porque recoge tanto las definiciones como la información que otros autores sugieren que debe tener el perfil, pero además dan una definición formal para un mejor manejo de la estructura desde un punto de vista computacional. Dicha definición formal se incluye a continuación.

6.2.2. Definición formal del perfil de usuario

Denotaremos como *E* el conjunto de los perfiles extendidos. Un perfil extendido e_1 pertenece a *E*, $1 \leq i \leq s$, siendo *s* el número de perfiles obtenidos, que puede ser representado en la tupla: [MBKV$^+$02]

$$e_i = \left(L_i, K_i, z_i^{'}, V_i\right) \tag{6.1}$$

donde:

- **Variables de identificación:** donde $L_i = (l_{i1}, l_{i2}, l_{i3}, \ldots, l_{ic})$ es el conjunto de identificación de variables de los archivos Web log, *c* es el número de variables, acerca de la identificación del usuario como por ejemplo el host (dominio o dirección IP), el agente de navegación (nombre y versión), entre otros que son almacenados en dichos ficheros.

- **Variables de clickstream:** donde $K_i = (K_{i1}, K_{i2}, K_{i3}, ..., K_{ir})$ es el conjunto de las variables de clickstream representado por el peso asociado a cada página *j* disponible, $1 \leq j \leq r$, donde *r* es el número de variables de cliskstream considerados, expresado en base a lapsos de tiempo en la página; si la página *j* no es visitada el valor de k_{ij} es 0.

- **Perfil simple:** donde un perfil simple $z_i \in Z$, siendo *Z* el conjunto de perfiles de usuario $Z = \{z_1, z_2, ..., z_d\}$, siendo *d* el número de perfiles. *T* es el conjunto de términos $T = \{t_1, t_2, .., t_n\}$ siendo *n* el número de términos, y $z^{'}$ la relación del conjunto de perfiles y los términos $z_i^{'} = (t_{i1}^{'}, t_{i2}^{'}, t_{i3}^{'}, \ldots, t_{ia}^{'})$ siendo $t_{ij}^{'} \in T$ donde *a* es el número de términos en el perfil definido por la función:

$$G : ZxT \longrightarrow [0,1] \, \forall z^{'} \in Z, t^{'} \in T, G\left(z^{'}, t^{'}\right) = \mu_{z^{'}}\left(t^{'}\right) \tag{6.2}$$

- **Variables demográficas:** donde $V_i = (v_{i1}, v_{i3}, v_{i3}, \ldots, v_{ib})$ representa el conjunto de variables demográficas, siendo *b* el número de variables a considerar. Las variables demográficas están relacionadas con aspectos demográficos y/o sociales del usuario, incluyendo el rango de edad del usuario, su nivel educativo, su idioma, entre otras. Como estas variables pueden ser imprecisas, las técnicas difusas pueden manejar estos datos con diferentes tipos de granularidades, dependiendo de la variable a modelizar.

En la figura 6.2 se puede ver un ejemplo relacionado con la definición que acabamos de dar, donde podemos ver las diferentes variables del perfil extendido con algunos valores. En el caso de las *variables de identificación*, los valores a considerar podrían ser, por ejemplo las IP, Host o los agentes, entre otras variables. Podemos ver también un ejemplo de las *variables de clickstream* donde encontramos diferentes páginas que el usuario ha visitado. En cuanto a las *variables demográficas*, encontramos campos que pueden ser tratados de manera difusa, tales como son la paciencia, el rango de edad, el nivel educativo, entre otros. Por último, podemos ver el *perfil simple*, que representa conceptos o temas relacionados con la navegación del usuario. En la siguiente sección veremos esta representación del perfil extendido del usuario en el lenguaje XML.

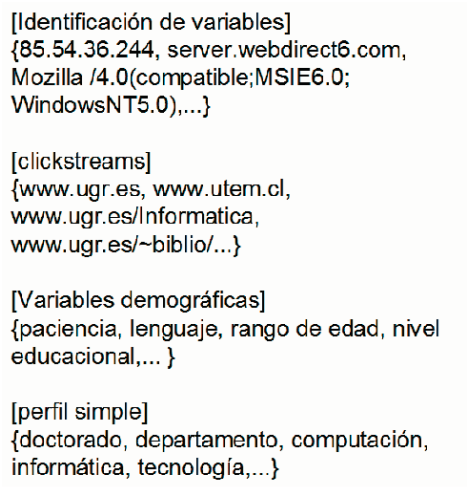

Figura 6.2: *Ejemplo práctico de representación del perfil del usuario*

6.3. Modelo de representación de perfiles de usuario: Representación en XML

Es bien conocido que el lenguaje XML es hoy en día uno de los lenguajes de representación e intercambio de información más relevante en la Web. Es por esta razón que presentamos un modelo de representación basado en este lenguaje que nos permita manejar con facilidad y automáticamente la información que ya tenemos almacenada en el perfil de usuario y que ya definimos en la sección anterior. De este modo, podemos determinar la representación general del perfil del usuario a través del lenguaje XML mediante

un esquema que se recoge en la figura 6.3.

```
<?xml version="1.0" encoding = "UTF-8" ?>
 <Perfil de Usuario>
      <Identificacion_Usuario>
        <Tipo Tipo={id_usuario} />
      </Identificacion_Usuario>
      <Paginas>
        <Pagina Pagina_Visitada="0"> Nombre página visitada  </Pagina>
        <Pagina Pagina_Visitada="1"> Nombre página visitada  </Pagina>
                          .
                          .
                          .
        <Pagina Pagina_Visitada="N">Nombre página visitada</Pagina>
      </Paginas>
      <Var_Demograficas>
        <Edad> {Rango de edad del usuario} </ Edad>
        <Género> {Género del usuario} <Género/>
        <Idioma> {Idioma de la página} </Idioma>
        <Paciencia>{Tiempo de navegación del usuario} </Paciencia>
        <Nivel_Educativo> {Dificultad de la página} <Nivel_Educativo />
      </Var_Demograficas>
      <Perfil_Simple>
        <Terminos Termino="0">Término </Terminos>
        <Terminos Termino="1"> Término </Terminos>
                          .
                          .
                          .
        <Terminos Termino="N"> Término </Terminos>
      </Perfil_Simple>
 </ Perfil de Usuario >
```

Figura 6.3: *Representación general del perfil del usuario en XML*

A continuación explicaremos los diferentes campos que hemos representados con el lenguaje XML y se ven en la figura 6.3.

- $<$ **Identificacion_Usuario** $>$**:** Este campo esta relacionada con la definición del conjunto de identificación de variables que hemos hecho en la sección 6.2.1 la cual es $L_i = (l_{i1}, l_{i2}, l_{i3}, \ \ldots, l_{ic})$.

  En esta parte del perfil se realiza la identificación del usuario, cuando decimos identificación hablamos de aquel usuario que se haya conectado aunque sea una vez al sitio y que se haya registrado o no en él. En este caso, surge un problema, el cual veremos en la sección 6.4.2, que trata de los usuarios registrados o los no registrado.

  Cuando el usuario está registrado, al usuario lógicamente lo podemos identificar, y de esta manera podremos realizar una personalización más adecuada a las pre-

ferencias del usuario. Y en al otro caso, cuando el usuario no esta registrado, el usuario es asignado a un perfil general relacionado con sus preferencias registradas al momento de navegar por la Web. No está demás señalar que las preferencias de los usuarios que navegan por la Web son almacenadas en los ficheros Log.

- <**Páginas**>**:** dentro de la definición del perfil también tenemos el campo llamado páginas, este campo se refiere principalmente a las páginas que el usuario ha visitado durante su navegación por el sitio Web y se relaciona con la definición del conjunto de *variables de clickstream* la cual es $K_i = (K_{i1}, K_{i2}, K_{i3}, ..., K_{ir})$ (Ver sección 6.4).

- <**Var_Demográficas**>**:** en el campo de las variables demográficas que hemos definido como $V_i = (v_{i1}, v_{i3}, v_{i3}, \ldots, v_{ib})$ en la sección 6.4 y aquí podemos ver algunas variables que representan ciertas características dentro de los perfiles del usuario como por ejemplo la edad, el género, el idioma, el nivel educativo ó la paciencia del usuario en su navegación. Dependiendo de la naturaleza de cada variable y de la fuente de origen, podemos determinar los valores que pueden tomar en las etiquetas correspondientes en XML.

  Por ejemplo, para la *edad*, su valor se podría tomar directamente de formularios que haya en el servidor de aplicaciones ó, a falta de un valor concreto especificado por el usuario, se podría estimar en función de las páginas que visita. En el caso de la web de la Escuela, por ejemplo, para accesos identificados en los que se conoce a priori si el usuario es alumno o no, se podrían establecer etiquetas lingüísticas relacionando los alumnos con la etiqueta *joven*; sin embargo, para accesos no identificador, se podría establecer la misma etiqueta si se identifica el perfil de usuario como perteneciente al grupo de los alumnos.

  Para la variable del *género*, sin embargo, es mucho más difícil establecer si el usuario es hombre ó mujer simplemente por su navegación, aunque en estudios muy orientados a dicha distinción se podría también estimar si el usuario es de un género ú otro dependiendo de las páginas que visita. Indudablemente, la obtención de dicha información a partir de formularios no dejaría duda alguna sobre el valor de la variable.

  En el caso de la variable del *idioma*, se puede suponer que si el usuario permanece un tiempo razonable en una página escrita en un determinado idioma es porque entiende ese idioma. No obstante, lo contrario no se puede afirmar, ya que si un usuario no permanece en la página no podemos saber si es porque la página no le interesa ó porque no entiende el idioma. Elementos adicionales pueden ayudarnos a

determinar el valor de este campo, como por ejemplo, que el usuario vea la misma página (respecto al contenido, no con el mismo nombre, obviamente) en una versión idiomática diferente.

Otra de las variables es la *paciencia*, la cuál está relacionada con el tiempo de navegación del usuario por el sitio Web. Obviamente, esta variable es altamente subjetiva, y puede ir determinada en función de la página en cuestión, ya que las páginas que contengan menos gráficos y que sean densas en texto llevarán asociado un mayor tiempo de lectura, por lo que el usuario necesitará más paciencia para leer la página completa. A estos valores de tiempo les hemos asociado etiquetas lingüísticas para entender de mejor manera el tiempo que utiliza en su navegación. En la figura 6.4 podemos ver la definición de las etiquetas para la variable de Paciencia.

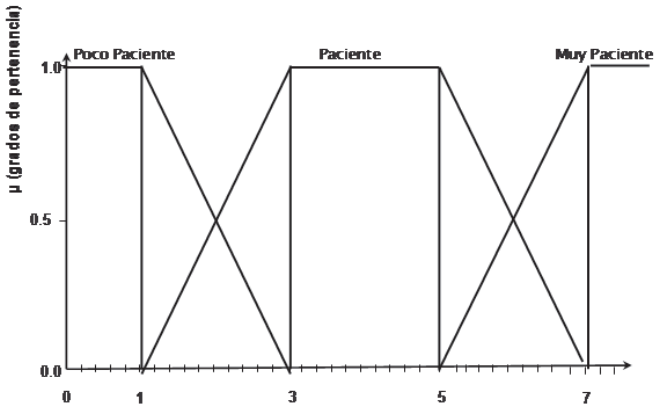

Figura 6.4: *Etiquetas lingüísticas para la variable paciencia.*

Como muestra la figura 6.5, hemos definido tres etiquetas: poco paciente, paciente y muy paciente en función del tiempo en minutos que el usuario permanece en la página. De esta manera podremos entender más fácilmente lo paciente que pueden llegar a ser los usuarios dentro de algún sitio Web.

Por último, la variable *nivel educativo* está relacionada con la dificultad que posee la página para su entendimiento o su lectura. Los valores de esta variable nos pueden ayudar también a hacer alguna estimación sobre la edad de los usuarios. Por ejemplo, si la navegación de un usuario se muestra impaciente a lo largo del sitio web, y éste está dedicado a la literatura, se puede suponer que el usuario no está interesado en los contenidos ó que no tiene una edad adecuada para los mismos porque es muy joven ó que su nivel educativo es bajo.

- <**Perfil_Simple**>**:** por último, está el campo del *perfil simple del usuario,* el cual hemos definido en la sección 6.2.1 como $z_i^{'} = (t_{i1}^{'}, t_{i2}^{'}, t_{i3}^{'}, \ldots, t_{ia}^{'})$ siendo $t_{ij}^{'} \in T$, donde *a* es el número de términos en el perfil.

  Este campo está relacionado principalmente con las páginas visitadas por el usuario, que se relacionan con la recuperación de información que el usuario haya realizado en el sitio Web o dicho de otra forma, el contenido de las páginas que haya visitado ó haya estado buscando.

  Para conocer el contenido de las páginas web de forma operativa, sin tener que leer la página completa, podemos ver las palabras clave ó más representivas de la página en la meta etiqueta *Keyword.*, que se encuentra en la sección *head* de las páginas web. Esta marca nos indica las palabras claves que se relacionan con el tema o términos importantes dentro del sitio o de la página web. Podemos ver un ejemplo de las palabras clave en la etiqueta Keyword del sitio Web de la Universidad de Granada (*http://www.ugr.es*). en la figura 6.5.

```
<meta name="keywords" content="titulaciones, centros, institutos,
departamentos, biblioteca universitaria, conozca granada, información
general, servicios, estudiantes, profesorado y PAs, relaciones
internacionales, investigación, extensión cultura cooperración,
evaluación y calidad, normativa, postgrado, acceso identificado,
webmail, directorio UGR, CSIRC, actualidad, actividades, agenda,
tablón, dossier de prensa. degrees, faculties, institutes,
departments, university library, Get to know Granada, general
information, services, students, teaching staff and administration and
services staff (PAS), international relations, research, extramural
studies: culture and cooperation, quality and evaluation, regulations,
postgraduate courses, authorized access, webmail, UGR directory,
Centre for Computer Services and Communication Networks (CSIRC), news,
activities, agenda, notice board, press releases"/>
```

Figura 6.5: *Ejemplo de palabras claves.*

Cabe resaltar que en muchas páginas no se encuentran este tipo de marcas, por lo que es necesario ver otras marcas que nos acerquen a determinar la información relacionada con el contenido de las páginas tales como las marcas $< H1 >$ ó $< TITLE >$. Al analizar este tipo de marcas es necesario quitar las *Stopwords o palabras prohibidas* consistentes en artículos, preposiciones, conjunciones y otras palabras del lenguaje que no aportan nada sobre el contenido semántico de la página. En la tabla 6.1 podemos ver un ejemplo de diferentes páginas Web con sus respectivas ”palabras claves”.

Páginas	Palabras claves
http://www.dcc.uchile.cl/ ljaramil/investigacion/	búsqueda, web, multimedia, dinámica semántica, mining, graph, universidad de chile
http://www.sadio.org.ar/	Sociedad Argentina de Informática
http://www.bits20.com/	blog, bits 20, 2.0, dos punto cero, dos cero, blogs, redes, redes sociales, adsense, adwords, diseño, buscadores, novedades, noticias, inversores, start-ups
http://www.acm.org/	scientific computing society, educational computing society, computing professionals, information technology, IT professionals, IT students, association for computing machinery, programming, computer programmers, algorithms and computational theory, Ada, APL, applied computing, computer architecture, artificial intelligence, biomedical computing, computer science

Tabla 6.1: *Ejemplo de páginas con sus respectivas palabras claves*

6.3.1. Caso real para la representación del perfil de usuario en XML

Para realizar la representación de perfiles de usuarios, hemos obtenido la información de los servidores Web, en este caso particular, de la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicaciones (ETSIIT) de la Universidad de Granada (*http://etsiit.ugr.es*).

La información que se extrae de los servidores Web es procesada principalmente para eliminar los elementos ruidosos para el análisis y de esta manera obtener un conjunto de datos más limpio para el proceso. Estos elementos ruidosos pueden ser las imágenes, javascripts entre otros elementos que pueden ser causante de un análisis poco eficiente y resultados poco representativos.

El tamaño del archivo analizado fue de 98202 entradas y tras el procesamiento del archivo se ha obtenido un archivo de 15676 entradas, el cuál pasa a ser el archivo objetivo del análisis.

Una vez que se haya identificado las entradas depuradas de los archivos log, identificaremos las sesiones de usuarios. Para ello, utilizaremos el *método Timeout*, que hemos explicado en la sección 3.4.3. Tras dicho proceso, aplicaremos sobre estas sesiones la técnica de Clustering Difuso, utilizando el algoritmo c-medias difuso para identificar los

grupos más similares entre las sesiones (Ver secciones 5.2.1 y sección 6.4). Una vez realizada esta etapa del proceso, lo que falta es representar los diferentes perfiles obtenidos en el proceso de minería.

A continuación veremos la representación para este caso particular del perfil del usuario a través del lenguaje XML, para luego explicar cada parte de esta representación y mostrar un ejemplo (Ver figura 6.6).

```
<?xml version="1.0" encoding = "UTF-8" ?>
 <Perfil de Usuario>
      <Identificacion_Usuario>
        <Tipo Tipo = {Alumno, Profesor} />
      </Identificacion_Usuario>
      <Paginas>
        <Pagina Pagina_Visitada="0"> Página visitada en la ETSIIT  </Pagina>
        <Pagina Pagina_Visitada="1"> Página visitada en la ETSIIT  </Pagina>
                        .
                        .
                        .
        <Pagina Pagina_Visitada="N"> Página visitada en la ETSIIT </Pagina>
      </Paginas>
      <Var_Demograficas>
        <Edad> {Rango de edad del usuario} < /Edad>
        <Idioma> {Idioma de la página Español} </Idioma>
        <Paciencia>{Tiempo de navegación del usuario} </Paciencia>
      </Var_Demograficas>
      <Perfil_Simple>
        <Terminos Termino="0">Término </Terminos>
        <Terminos Termino="1"> Término </Terminos>
                        .
                        .
                        .
        <Terminos Termino="N"> Término </Terminos>
      </Perfil_Simple>
 </ Perfil de Usuario >
```

Figura 6.6: *Representación de un caso particular de perfil de usuario*

Comenzaremos explicando la parte del perfil llamada *Identificación de usuario.* Para el caso de la Escuela Técnica Superior de Ingenierías Informática y de Telecomunicaciones dentro de sus ficheros log encontramos que los usuarios no están registrados y por lo tanto, no tenemos una identificación clara de cada uno. De esta manera lo que hemos hecho es asignar a cada usuario a un grupo con preferencias similares y así según estas preferencias poder identificar al usuario si es *alumno* o *profesor*.

Dentro de la definición del perfil también tenemos el campo de páginas, este campo se refiere principalmente a las páginas que caracteriza al perfil, o sea a las páginas que el usuario a visitado durante su visita al sitio de la Escuela.

En el campo de las variables demográficas hemos definido las etiquetas correspondientes a las variables definidas en la sección 6.3.1. Concretamente, la edad, el género, el idioma, la paciencia y el nivel educativo. De todas estas variables, podemos determinar a priori el valor de la variable idioma, ya que el sitio web que estamos analizando está en

español.

Por último, respecto a la definición del perfil simple del usuario formado por las palabras claves de las páginas, en nuestro caso de estudio, no ha sido posible extraerlas a partir de la marca *keywords* de las páginas, por no estar definida dentro del sitio. Debido a esto, hemos obtenido las palabras claves de las marcas $< H1 >$, y en ausencia de ellas, de las marcas $< TITLE >$. En la tabla 6.2 podemos ver diferentes páginas con sus palabras claves correspondientes.

Páginas	**Palabras claves**
GET/apps/foro/index.php	ETSIIT; Foros; Dudas; Redes
GET/alumnos/juliolo/genetica	ETSIIT; Foros; Economía; Empresa
GET/alumnos/juliolo/hormcrec.htmlHTTP/1.1	Ingeniería; Informática; Telecomunicación ;Planes ;estudios
GET/alumnos/mlii/eniac.htm	ETSIIT; Foros; Practica; BIO
GET/alumnos/mlii/Harvard%20Mark%20I.htmHTTP	ETSIIT; Foros; practica; periféricos

Tabla 6.2: *Ejemplo de páginas del sitio http://etsiit.ugr.es con sus palabras claves*

Una vez que ya hemos definido cada campo del perfil de usuario, veremos un ejemplo de los resultados obtenidos al analizar 15676 entradas de ficheros de log obtenidos del nuevo sitio de la Escuela, el cual hemos utilizado anteriormente en nuestros análisis. Con los resultados obtenidos aplicando el clustering difuso hemos conseguido 12 diferentes perfiles de usuarios. A continuación veremos un ejemplo de uno de estos perfiles. (Ver figura 6.7).

De los perfiles obtenidos, podemos decir que la gran mayoría de ellos son de alumnos, por lo que su edad es joven. Sólo uno de los perfiles corresponde a un profesor. Este perfil de profesor se relaciona principalmente con temas como horarios, planes de estudios, convocatorias, entre otros intereses (Ver Apéndice C). Además, los usuarios se muestran pacientes o muy pacientes al momento de realizar su navegación por la páginas del sitio web.

Si hacemos referencia a los perfiles de los alumnos podemos decir que la gran mayoría de los alumnos visitan las páginas de los foros. Dentro de estos foros podemos encontrar temas relacionados con la programación, la docencia, información general, asignaturas, prácticas, entre otros temas, los cuales se reflejan claramente dentro de los perfiles obtenidos (Ver Apéndice C).

De forma general podemos decir que los términos relacionados con las páginas que visita los usuarios de la Escuela son principalmente términos relacionados con la Escue-

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil4>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=escuelaHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1583HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=generalHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1874HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1709HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/page.php?pageid=departamentosHTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">tablón</Terminos>
    <Terminos Termino="5">"Index</Terminos>
    <Terminos Termino="6">js/protWindows/themes</Terminos>
    <Terminos Termino="7">"Escuela</Terminos>
    <Terminos Termino="8">comprar</Terminos>
    <Terminos Termino="9">coche</Terminos>
    <Terminos Termino="10">General</Terminos>
  </Perfil_Simple>
</Perfil4>
```

Figura 6.7: *Perfil de usuario obtenido en el caso real*

la, la Docencia, Temas Generales del Foro, y lo más lógico es que aparezcan términos relacionados con la Ingeniería, la Informática y las Telecomunicaciones.

6.4. Modelo para la obtención de perfiles de usuario

Los perfiles de usuario se pueden obtener de diferentes fuentes y a través de diferentes procesos. Los métodos principales para la creación de los perfiles son: *el método explícito o manual*, que responde principalmente a la introducción de los datos a través de formularios, el *método colaborativo* o de composición a partir de otros perfiles, en donde los perfiles se pueden crear o actualizar a través de la interacción colaborativa con otros perfiles, con los que se relacionan, recurriendo al conocimiento específico del dominio y heurísticas inteligentes; y por último, el *método implícito*, que utiliza técnicas específicas para extraer las características para crear ó modificar automáticamente los perfiles, recurriendo normalmente a técnicas de inteligencia artificial para realizar estas tareas [DGR03].

Del punto de vista del *perfil explícito*, un perfil se construye guardando la actividad directa del usuario, típicamente a través del llenado de formularios y cuestionarios. Cada perfil puede contener información genérica como la fecha de nacimiento y el código de área, también como alguna información dinámica, cuál deba probablemente cambiarse el tiempo como programas de televisión favoritos o las selecciones de fútbol. El perfil explícito requiere que los usuarios se involucren directamente y pongan la mayor parte del esfuerzo, y por lo tanto depende de la motivación del usuario.

Un ejemplo del método explícito es el sistema Doppelganger [Orw95], este sistema construye perfiles de usuarios explícitos utilizando métodos estadísticos y de aprendizaje automático. El sistema Doppelganger aplica un algoritmo de agrupamiento o clustering a los perfiles para descubrir usuarios semejantes, formando perfiles de grupos de usuarios.

En el *método colaborativo* se realizan predicciones automáticamente sobre los intereses de un usuario recogiendo la información de los intereses de muchos otros usuarios. Por ejemplo un sistema de colaboración para la recomendación sobre la preferencia de la música podría hacer predicciones sobre que música debe tener preferencia un usuario dado una lista parcial de las preferencias de ese usuario. Observar que estas preferencias son específicas al usuario, la información de uso recopilada proviene de muchos otros usuarios. El sistema Firefly [SM95] se basa en métodos colaborativos para recomendar música a los usuarios. Un ejemplo mucho más actual lo podemos ver claramente en la tienda de online llamada Amazon, que utiliza este tipo de método para adaptar sus páginas según las preferencias de sus usuarios o clientes. En la figura 6.8 podemos ver la interacción de

diferentes perfiles y sus fuentes de información, para su actualización y cooperación entre ellos.

En el *método implícito* los usuarios no están involucrados y se necesita de herramientas para poder obtener información útil de los usuarios y que se pueden obtener utilizando técnicas de minería como por ejemplo, el clustering.

En nuestro modelo, el método básico es el método implícito, aunque parte de la información que completa el perfil de usuario se obtiene de formularios y aplicaciones al igual que en el método explícito. En cuanto al modelo colaborativo, es aplicable una vez tenemos identificados ciertos perfiles de usuarios con ciertas clases demográficas para predecir a qué clases pertenecen otros usuarios sin identificar, pero con un perfil parecido a los ya determinados.

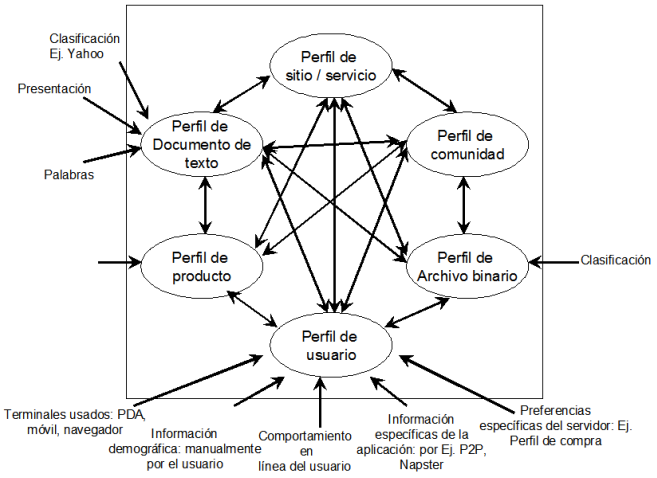

Figura 6.8: *Interacción entre diferentes perfiles y sus fuentes de información.*

6.4.1. Obtención de datos

El conocimiento para la creación de los perfiles lo podemos obtener de la navegación generada por el usuario. Los clicks que el usuario hace, nos indican el tiempo que una página Web es visible en el navegador de un usuario.

Estos datos (Web log) son conjuntamente llamados *entradas de la navegación.* Cada vez que un usuario se conecta a un sitio Web, una sesión nueva comienza. Cada click en un URL, en una imagen o un enlace general, en una sesión de usuario que representa

una entrada en el archivo Web log. Una sesión está cerrada cuando el tiempo transcurrido entre dos click es más alto que un umbral colocado como prefijo [NFJK99].

Mientras navega por la Web, el usuario va dejando registrado todas las acciones que realizó. El proceso de recoger los datos es efectuado por uso de elementos explícitos o implícitos relacionados con el usuario. Los elementos explícitos son básicamente formularios, encuestas de opiniones y registros de los usuarios cuando navegan por un sitio Web. Y los elementos implícitos incluyen las cookies y los archivos log.

En la figura 6.9 se muestra el proceso de actividad de un usuario mientras navega por un portal o por algún sitio Web. El servidor Web es la interfaz del sitio y maneja las peticiones del usuario, estas peticiones son registradas en los ficheros log. El servidor de aplicaciones permite la administración del sitio, la personalización y un motor de búsqueda de contenido y el servidor de contenido es el administrador del contenido y documentos del sitio. Cuando el usuario interactúa con el sitio, realizando algún click sobre una imagen, un botón, en un contenido, etc. es registrado en la tabla de actividad de la página. Con el cliscksteam y con los ficheros log, que se obtiene directamente del servidor Web, se puede realizar una identificación de sesión, para posteriormente identificar y clasificar a los usuarios (perfiles extendidos) a través de alguna técnica apropiada de minería [MBKV+02].

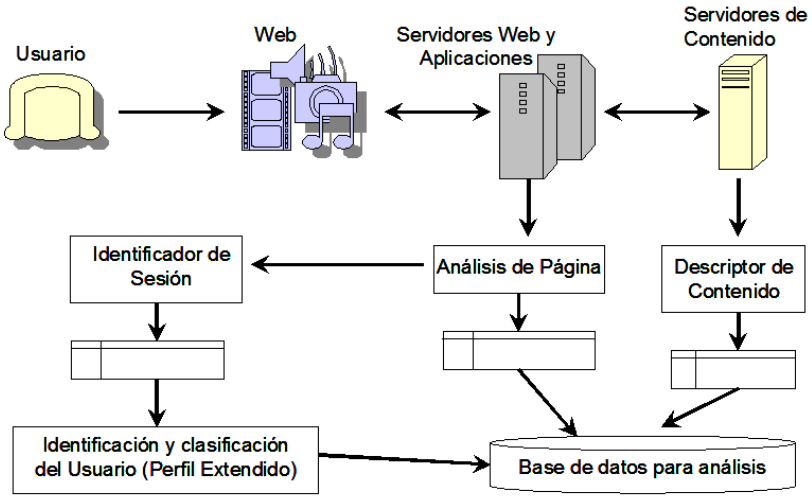

Figura 6.9: *Proceso de obtención de datos desde la actividad del usuario*

6.4.2. Usuarios registrados y no registrados

El inconveniente principal del manejo de perfiles del usuario en la Web es la falta de conocimiento acerca de la identidad del usuario, de esta idea surgen dos situaciones diferentes. La primera situación son *los usuarios sin registrar,* donde el perfil de usuario puede identificarse o personalizarse con un grupo social, asignando un perfil general relacionado con unas preferencias mostradas por el usuario mientras navega a través del sitio Web.

La segunda situación se refiere a los *usuarios registrados,* si un usuario es identificado de algún modo, entonces el sitio Web puede estar hecho a la medida según las preferencias del usuario; del punto de vista de los negocios si un usuario ha visitado el sitio Web antes y se ha registrado. El sistema sigue la pista al usuario de visitas previas junto con el perfil del usuario, así que puede usar esta información para realizar y personalizar el sitio Web. Donde la personalización Web no precisa interacción y retroalimentación explícita con usuarios, sin un conocimiento a priori, el método clustering es una herramienta válida para crear estos grupos de interés.

La personalización es una herramienta para atraer a los clientes probables, así para identificar a los usuarios con buenas experiencias en el sitio Web, podrían registrarse la próxima vez que se conecten al sitio.

El usuario puede ser caracterizado por un conjunto de perfiles de otros usuarios, relacionado a través de sus preferencias o intereses comunes con estos usuarios y así es identificado por algún grupo social. Cada grupo representa un grupo del usuario de interés con patrones similares de navegación.

Esto se realiza a través de alguna técnica de minería, como el Clustering, para descubrir usuarios semejantes y con ellos formar perfiles de grupos de usuarios. Y de esta manera, podemos inferir sobre aquellos usuarios que no se registren o no pertenezcan a un perfil determinado y asociarlos a alguna clase o grupo semejante a su área de interés.

Luego de haber realizado un análisis general relacionado con el proceso de personalización, revisado las diferentes definiciones de los perfiles de usuarios en la literatura y los principales inconvenientes que se presentan en el manejo de los perfiles, ahora veremos un modelo para la obtención de perfiles de usuario.

6.4.3. Modelo para la obtención de perfiles de usuario

Proponemos un modelo que nos permitirá la representación automática de perfiles de usuarios de algún sitio Web, este modelo lo podemos ver en la figura 6.10.

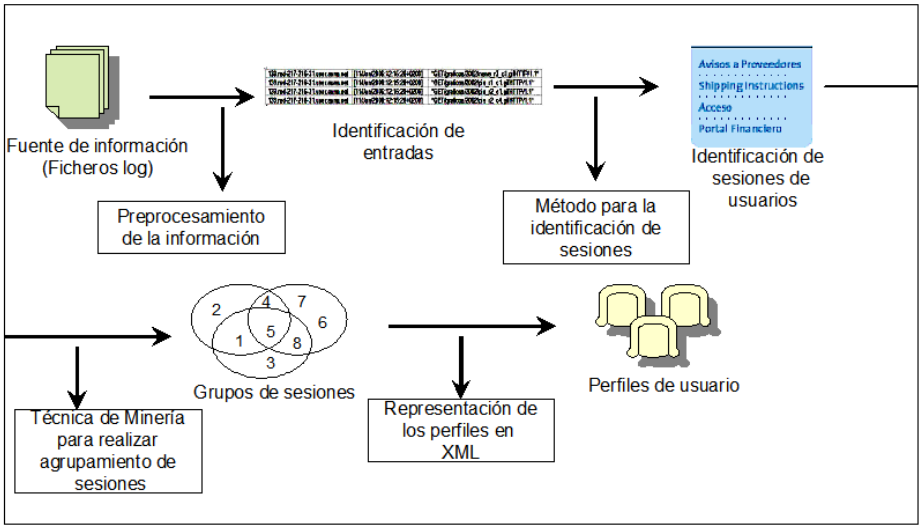

Figura 6.10: *Modelo para la obtención de perfiles.*

- **Fuente de información:** Este modelo nos permite analizar la información de lo servidores Web, esta información es principalmente obtenida de los ficheros Log que se almacenan en los servidores a medida que los usuarios vayan navegando por algún sitio Web.

  Existen diferentes tipos o formatos de ficheros log (ver sección 4.3.1), esto depende principalmente del tipo de servidor que se utilice. El principal problema que se presenta para analizar estos tipos de archivos es que carecen de una estructura determinada, ya que sólo se sabe los campos con los cuales están constituidos y es necesario realizar un proceso de identificación de transacciones para poder hacer un mejor análisis.

- **Preprocesamiento e identificación de transacciones:** Por consiguiente para realizar un buen proceso de Minería, en este caso Minería Web de Uso para la representación de perfiles de usuarios es necesario realizar un buen preprocesamiento de la información, ya que de esta manera nos garantizamos resultados más cercanos a

la realidad. Por esta razón, dentro del modelo que presentamos se realiza un proceso de limpieza de los datos principalmente para eliminar transacciones que nos impidan acercarnos más a la realidad de los clientes o usuarios.

La limpieza de los datos consiste principalmente en eliminar transacciones que contengan imágenes *(*.jpg, *.bmp, *.png, *.gif, etc)*, javascript *(*.js)*, transacciones que por alguna razón se repiten o sea son replicas y también hojas de estilos *(*.css)* que no representen ningún significado importante para el análisis.

Al mismo tiempo que se realiza la limpieza de los datos, se identifican las entradas o transacciones que va dejando registradas el usuario durante su navegación. De esta manera se irá formando un conjunto de datos que posea una cierta estructura para realizar el análisis, ya que la información obtenida de los servidores Web es en bruto, o sea que carece de alguna estructura.

Continuando con el método que planteamos en la figura 6.10, la siguiente fase del proceso luego de la identificación de las transacciones es la identificación de sesiones de usuarios.

- **Identificación de sesiones de usuarios:** Para realizar esta identificación de usuario existe en la literatura diferentes métodos, los cuales hemos comentados en la sección 6.1.1. Dentro de estos métodos podemos mencionar al *reference length*, el método *maximal forward references*, el *timeout*, entre otros. Basándose en alguno de estos métodos, se identificará las sesiones de usuarios, y también se determinará el umbral óptimo de tiempo que permitirá definir cuando comienza y cuando termina una sesión. Según estudios realizados (Ver sección 6.1.1) se estima que el tiempo óptimo es de 30 minutos.

  Cuando ya tenemos el conjunto de datos con una cierta estructura, o sea sin elementos ruidosos para el análisis, y también identificadas las transacciones y las sesiones de usuarios, es el momento de aplicar alguna técnica de Minería para procesar el conjunto de datos objetivo para el agrupamiento de sesiones de usuarios.

- **Minería:** Como ya hemos comentado en capítulos anteriores, las técnicas más utilizadas en la Minería Web de Uso según la literatura existente en esta área son las *reglas de asociación* (Ver Apéndice B) y el *clustering*.

  Las reglas de asociación por ejemplo nos permitirían determinar las tendencias de navegación de los usuarios y el clustering nos permitiría agrupar el conjunto de transacciones por sesiones que posean similares páginas de navegación o quizás

podríamos agrupar por el mismo número de IP (Ver sección 6.1.2), como ha sido comentado en otras secciones y capítulos utilizaremos estas técnicas del punto de vista difuso. Estos patrones encontrados pueden ser almacenados en los perfiles de usuario y el clustering a su vez, nos permitirá la creación y obtención de los perfiles de usuarios, a través de los clusters o grupos de usuarios que hemos encontrados en el análisis de agrupamiento de sesiones de usuarios por páginas Web similares.

6.4.4. Obtención de los perfiles de usuario a partir del clustering: Caso real

Para la representación de los perfiles, los clusters encontrados en el análisis anterior (Ver sección 5.6.1) están fuertemente relacionados, ya que a partir de ellos podemos saber cuales son las preferencias e intereses de los usuarios, y con la información entregada por los mismos podremos crear los diferentes perfiles.

Por consiguiente, estos cluster nos llevarán a la creación de los perfiles a partir de la información que el usuario ha dejado registrada durante su navegación, en este caso particular, por el sitio Web de la Escuela. Y a través de esta representación de los perfiles podremos inferir que perfil puede corresponder a los intereses o preferencias de los alumnos ó los profesores.

La figura 6.11 representa la relación entre los clusters encontrados y la creación de los perfiles, ya que a partir de los cluster, como lo hemos mencionado anteriormente, se realizará la creación y representación de los diferentes perfiles.

A través del análisis del clustering podremos obtener los diferentes conjuntos de las sesiones de usuarios. Cada grupo o clusters posee un centroide, el cual es el elemento que mejor representa al grupo encontrado. Cuando decimos un centroide no quiere decir que sea solo ”uno”, sino que pueden existir varias sesiones que poseen las mismas características y sean estas las que representen de mejor manera al grupo o clusters.

De esta manera al obtener los diferentes centroides de cada grupo encontrado, podremos realizar una representación general de los perfiles de los usuarios. Esta representación reflejará diferentes características importantes de los usuarios que navegan por el sitio Web y que esta información puede ser muy útil al momento de realizar algún tipo personalización dentro del sitio Web. Y de esta manera identificando los centroide de cada grupo podremos identificar si las preferencias corresponde a un profesor o a un alumno.

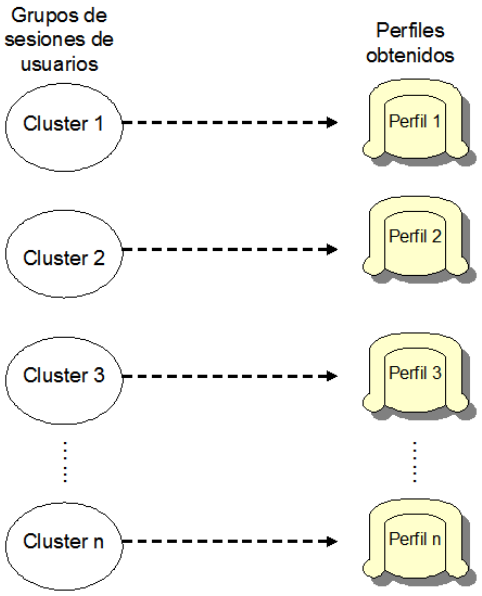

Figura 6.11: *Cluster vs Perfiles*

6.4.4.1. Caso real

Es importante señalar que el análisis que hemos realizado en el capítulo anterior es fundamental para la creación y luego representación de los diferentes perfiles de usuarios, para este caso particular de la Escuela de Informática y Telecomunicaciones de la Universidad de Granada.

Hemos planteado un enfoque para la agrupación de sesiones de usuarios, y a partir de este análisis poder lograr identificar los diferentes perfiles de usuarios a través de los grupos o clusters de sesiones que hemos logrado obtener. En la figura 6.12 podemos ver este enfoque relacionado con la agrupación de las sesiones de los usuarios.

Con la agrupación de las sesiones de usuario el objetivo era resolver la problemática que se presenta en la falta de conocimiento acerca de la identidad del usuario, en donde en el Web podemos encontrarnos con usuarios que se registran y con otros que no se registran (Ver sección 6.4.2). De esta manera se puede identificar o personalizar a un grupo a los usuarios no registrados, y así asignar un perfil general que relaciona todas sus preferencias o intereses.

En la tabla 6.3 podemos ver la identificación de los distintos grupos o clustering con

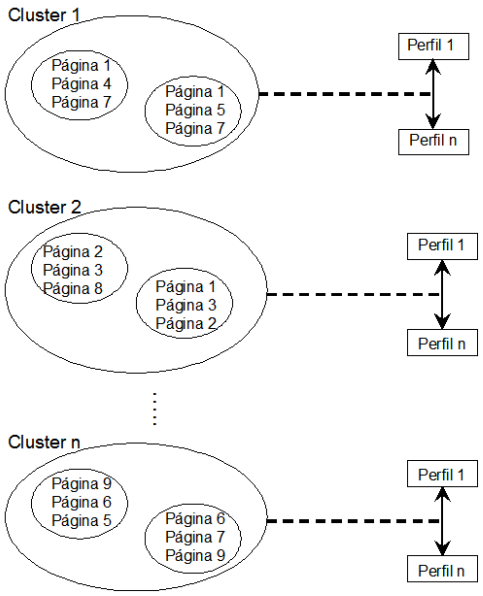

Figura 6.12: *Agrupación de sesiones para la creación de perfiles*

los diferentes perfiles de usuarios.

Otro de los objetivos planteados era identificar a los usuarios que tengan un perfil de alumno con un usuarios que pudiese ser un profesor. Con los diferentes grupos obtenidos del análisis de la agrupación de sesiones hemos podido inferir entre que grupos tenían más características de profesor y que grupos tenían más características de alumnos.

Podemos ver un ejemplo de lo planteado en las siguientes figuras (Ver figuras 13 y 14). En estos perfiles que identificamos como perteneciente a alumnos y profesores respectivamente, hemos hecho esta identificación principalmente basándonos en las páginas que visitaron durante su navegación por el Web. Por ejemplo, si las páginas que fueron visitadas tenían alguna relación con secciones del sitio Web de la escuela como el Foros, tablón de anuncios, entre otras podíamos asumir que el usuario que había hecho la navegación era un alumno. Y por el contrario, si las páginas visitadas estaban relacionados con planes, actas, entre otras, se podía asumir que el usuario que había hecho la navegación correspondía a un profesor.

Por ejemplo, si el usuario visita páginas como {*profesores/jmaroza/*} o {*depar/ccia/mp1/index.htm*}, asumiremos que el usuario conectado corresponde a un profesor, en cambio si el usuario visita páginas como { *alumnos/mlii_/eniac.htm* } o { *apps/foro/index.php?action=foro&idforo=asignatura*}

Número del Cluster	Perfil que identifica al Cluster
0	Perfil 1
1	Perfil 10
2	Perfil 6
3	Perfil 4
4	Perfil 2
5	Perfil 3
6	Perfil 11
7	Perfil 5
8	Perfil 9
9	Perfil 8
10	Perfil 7
11	Perfil 12

Tabla 6.3: *Identificación de los clusters con su respectivo perfil*

o la página está relacionado con algún foro del sitio, por lo tanto podemos asumir que el usuario conectado corresponde a un alumno.

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil1>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=foro&amp;idforo=generalHTPP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;idHebra=1937HTPP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.pho?action=hebra&amp;idhebra=1916HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=foro&amp;idforo=escuelaHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1709&amp;page=1HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/apps/foro/index.php?action=foro&amp;idforo=asignaturasHTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">General</Terminos>
    <Terminos Termino="5">"Index</Terminos>
    <Terminos Termino="6">js/protWindows/themes</Terminos>
    <Terminos Termino="7">"ETSIIT</Terminos>
    <Terminos Termino="8">Escuela</Terminos>
    <Terminos Termino="9">Asignaturas</Terminos>
    <Terminos Termino="10">Docencia</Terminos>
    <Terminos Termino="11">tablón</Terminos>
  </Perfil_Simple>
</Perfil1>
```

Figura 6.13: *Ejemplo de un Perfil de Usuario para un alumno*

6.4.5. Obtención de la clasificación de los perfiles de usuarios a través de páginas web: Caso real

Existe otra manera de mirar los perfiles de usuarios, es a través de la clasificación de páginas Web que clasifiquen en ciertos perfiles de usuarios.

Para este análisis de clasificación es necesario identificar las diferentes sesiones de usuarios con las respectivas páginas y a la vez asociarlas a los diferentes perfiles de usuario. De esta manera podremos clasificar los diferentes perfiles de usuarios con las páginas que más se le asocien a ellos.

Esta clasificación nos servirá para determinar en una futura navegación de algún usuario a qué perfil podría pertenecer, esto va depender claramente a la página que el usuario estaría visitando durante su navegación.

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil10>
  <Identificacion_Usuario>
    <Tipo Tipo="profesor" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/usuarios/jmlvega/idragon//formate.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/convocatorias/styles/convocatorias.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/profesores/jmaroza/anecdotario/chmanual.htmHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/profesores/jmaroza/anecdotario/anecdotario-z.htmHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/planes/index.php?id=3&amp;amp;id2=127HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/page.php?pageid=horarioHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1617HTTP/1.1*</Pagina>
    </Paginas>
  <Var_Demograficas>
    <Edad>Adulto</Edad>
    <Paciencia>Muy Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">"Index</Terminos>
    <Terminos Termino="4">skin/reloaded</Terminos>
    <Terminos Termino="5">convocatorias</Terminos>
    <Terminos Termino="6">ubuntu"</Terminos>
    <Terminos Termino="7">planes"</Terminos>
    <Terminos Termino="8">estudio"</Terminos>
    <Terminos Termino="9">Horario"</Terminos>
  </Perfil_Simple>
</Perfil10>
```

Figura 6.14: *Ejemplo de un Perfil de Usuario para un profesor*

Para este análisis se han considerado las páginas de las distintas sesiones de usuarios, para clasificar a los perfiles con ciertas páginas que se encuentran en dichas sesiones, y de esta manera tener una mejor idea de que páginas son las de mayor trascendencias dentro de los perfiles.

Luego de crear el conjunto de datos para el análisis, hemos realizado una clasificación de los distintos perfiles encontrados con las páginas de cada sesión de usuario. Esta clasificación la hemos hecho a través del algoritmo C4.5 (ver Apéndice D) que en el programa Weka existe una versión más avanzada denominada J4.8, que es un algoritmo de aprendizaje basado en un árbol de decisión. Para este análisis hemos evaluado la calidad del clasificador mediante validación cruzada y siendo esta opción con la que hemos obtenido los mejores resultados.

En el experimento realizado se han podido clasificar correctamente el 58.3 % de las instancias con un valor Kappa de 0.5299 (que mide lo que se ajusta la predicción a la clase real; 1.0 significa ajuste total) que se considera bueno para el caso de análisis.

De esta clasificación podemos decir que existen varios perfiles dentro del análisis que están determinados por algunas páginas específicas como:

- el **perfil 10** está clasificado por la página *GET/apps/convocatorias*
- el **perfil 12** está clasificado por la página *GET/apps/descargas*,
- el **perfil 3** está clasificado por la página *GET/alumnos/shin/shin.htm*
- el **perfil 7** está clasificado por la página *GET/alumnos/mlii*
- el **perfil 4** está clasificado por la página *GET/apps/foro/index.php*
- el **perfil 11** está clasificado por la página *GET/apps/foro/index.php?action=hebra&idhebra=1819*
- el **perfil 5** está clasificado por la página*GET/alumnos/shin/shin.htm*
- el **perfil 1** está clasificado por la página *GET/alumnos/mlii*

En cambio los demás perfiles están siendo clasificados no sólo por una página en especial sino por varias:

- el **perfil 6** está clasificado por la página
  - *GET/apps/foro/index.php*
  - *GET/apps/tablon*
  - *GET/alumnos/diegorp/canalplus.html*
  - *GET/alumnos/diegorp/canal.css*
- el **perfil 2** está clasificado por la página
  - *GET/apps/foro/index.php*
  - *GET/apps/tablon*

6.4.5.1. Discusión de los resultados de la clasificación de los perfiles de usuarios a través de páginas web

El análisis realizado de clasificación de los perfiles de usuarios a través de las páginas web, nos entregó valores buenos para el modelo del árbol resultante. Es importante señalar que los perfiles 6 y 2 tenían la mayor cantidad de elementos erróppertenncneos clasificados.

Un aspecto interesante es que este análisis, corrobora los resultados que hemos obtenido con relación a los perfiles, por ejemplo podemos decir que que el perfil 10, que representa al perfil del profesor, está clasificado en este análisis por la página *GET/apps/convocatorias*, y esta página dentro de la Universidad de Granada es muy utilizada por los profesores para realizar las convocatorias de los exámenes.

También podemos decir que los demás perfiles de usuario han sido clasificados por páginas relacionadas principalmente por la navegación que realiza un alumno; esto también justifica los resultados relacionados con los perfiles de usuario a través del clustering.

Con respecto a las páginas que clasifican a los perfiles, podemos señalar que las páginas del foro (GET/apps/foro/index.php)y tablón de anuncios (GET/apps/tablon) son páginas que han estado presentes durante todos los análisis que hemos realizado, desde la obtención de las reglas hasta la clasificación de los perfiles. Por lo tanto, estos representan un comportamiento habitual y particular en la navegación de los alumnos por el sitio web de la Escuela.

Otra página interesante es la de GET/alumnos/diegorp/canalplus.html, la cuál representa una clasificación diferente para los perfiles, ya que la gran mayoría están relacionadas con aspectos académicos dentro del sitio de la escuela por ejemplo GET/apps/descargas, GET/apps/foro/index.php?action
=hebra&idhebra=1819, GET/apps/convocatorias, entre otras; en cambio esta representa el lado menos académico del sitio, ya que se relaciona con aspectos extracurricular como son programas o aspectos relacionados con la señal de canal plus.

6.5. Conclusiones

El perfil de usuario puede almacenar información interesante acerca de los usuarios o clientes que navegan por un sitio web, y así podemos conocer sus hábitos y preferencias de navegación, y poder ofrecer o recomendar al usuario información personalizada. Mediante la lógica difusa, se puede manejar y representar de forma más flexible la información del

usuario que sea vaga e imprecisa tal como la edad que tiene o cuán paciente es en su navegación.

En este capítulo, hemos partido de una definición formal de perfil de usuario dada en [MBKV+02] para nosotros proponer un modelo de representación del perfil en lenguaje XML, que es uno de los lenguajes más éstandar hoy día en la Web.

Para obtener los perfiles, hemos detallado un modelo de obtención de los datos basándonos en las técnicas de minería mostradas en este trabajo, principalmente el clustering. Es más, desde un punto de vista experimental, hemos obtenido perfiles de usuario reales a partir de los grupos demográficos resultados de los experimentos del capítulo anterior. Además, hemos completado este modelo caracterizando los perfiles mediante las páginas web que mejor los representaban, lo cuál nos ha permitido revalidar los resultados obtenidos.

El siguiente capítulo se recogen las conclusiones sobre este trabajo y las líneas de investigación a seguir en el futuro.

Capítulo 7

Conclusiones y trabajos futuros

7.1. Conclusiones

A lo largo de este trabajo nuestros objetivos han sido mostrar el potencial del uso de la lógica difusa en el desarrollo de distintas herramientas en la Minería Web de Uso. Nos hemos centrado principalmente en la aplicación de técnicas tales como las reglas de asociación difusas y el clustering difuso. Ello nos ha permitido la obtención de patrones de navegación y análisis demográfico de los usuarios que navegan por un determinado sitio web.

Además, la información obtenida con las técnicas anteriores nos ha permitido la obtención de perfiles de usuario representados en XML a través de un modelo de obtención desarrollado. A partir de estos perfiles de usuario se pueden llevar a cabo procesos de personalización para la mejora del sitio web.

Para lograr esto hemos tenido que profundizar en cada uno de los procesos implicados en la tarea de la Minería Web del Uso, estudiando, definiendo, modelando e implementando las técnicas más adecuadas para la consecución de nuestros objetivos. De forma más detallada, la labor realizada queda descrita a continuación:

- Obtención de patrones de navegación:
  - Hemos planteado un modelo de obtención de reglas de asociación difusas mediante un análisis de preprocesamiento a partir de los ficheros log con una representación transaccional de los datos, permitiéndole al usuario configurar el contenido de las reglas a extraer.

- A partir de las reglas obtenidas, hemos llevado a cabo un proceso de interpretación semántica aplicando tanto medidas de interés objetivas como medidas de interés subjetivas basadas en creencias obtenidas, entre otras cosas, a partir de una encuesta real.
- Hemos experimentado sobre un sitio web real, que nos permitió validar los modelos anteriores y obtener patrones de navegación que nos describen el comportamiento de los usuarios que navegaban por el sitio analizado.

- Análisis demográfico:

  - Utilizando diversas técnicas del clustering hemos podido establecer una metodología para realizar diferentes agrupaciones de los elementos que participan en un sitio web. Dicha metodología implica no sólo la aplicación de algoritmos de clustering al uso sino también la obtención de una partición inicial de datos y la aplicación de medidas de validación de los procesos realizados.
  - Concretamente hemos agrupado tanto páginas como sesiones de usuario para llevar a cabo nuestro análisis:
    - Mediante un algoritmo de clustering de c-medias, hemos agrupado páginas similares para la obtención de los contenidos más representativos en la navegación de los usuarios.
    - Así mismo, hemos utilizado un algoritmo de clustering difuso de c-medias para agrupar sesiones de usuarios y obtener diferentes grupos de usuarios con características similares y así identificar a los usuarios que se conectan a la web. Con este fin, una vez definido el modelo de datos hemos tomado las medidas del coseno y del coseno extendido como medidas de similaridad de patrones, siendo ésta última la más adecuada para tratar el carácter sintáctico de las direcciones de las páginas.
  - Como una etapa previa a los dos procesos anteriores, hemos utilizado el clustering jerárquico para la obtención de la partición inicial de los datos y hemos utilizado el coeficiente de partición y la entropía como medidas de validación para las técnicas anteriores.
  - Por último, hemos experimentado sobre un caso real que nos permitió identificar diferentes grupos demográficos de usuarios. Concretamente, para el sitio analizado de la web de la E.T.S de Ingenierías Informática y de Telecomunicación de la Universidad de Granada (http://etsiit.ugr.es), se han identificado grupos con características asociadas al alumnado y otros con características más propias del profesorado.

- Construcción de perfiles de usuario:
  - Hemos planteado una nueva representación de los perfiles de usuarios en XML, basándonos en la definición formal propuesta por [MBKV+02].
  - Hemos definido un modelo de obtención de los perfiles de usuarios, basándonos en el análisis demográfico realizado anteriormente en los procesos de minería. Concretamente, hemos establecido una identificación de los diferentes perfiles a partir de los grupos de usuarios recogidos en el anterior análisis de clustering de sesiones de usuario.
  - Hemos realizado una clasificación de los perfiles de usuarios a través de las páginas web más representativas, la cual confirmó la caracterización de los grupos de usuario obtenidos en el análisis del caso real.

7.2. Trabajos futuros

Finalizaremos este trabajo señalando algunas líneas de investigación que han surgido del estudio realizado y que no han sido abordadas, debido a que sobrepasan los objetivos de este proyecto de investigación, o que han aparecido como consecuencia de sus resultados. Estas líneas de trabajo se recogen a continuación:

- Extender los resultados obtenidos a otros sitios web hasta desarrollar una herramienta integrada que incluya tanto los procesos de análisis descritos como la actualización dinámica y on-line de los perfiles de usuario.
- Ampliar el estudio de otras agrupaciones, asociaciones y relaciones entre los elementos que participan en el sitio web, como por ejemplo las IP's y los perfiles de usuarios, para así completar el conocimiento que se puede extraer a partir de la información inicial.
- Hay que hacer notar que este trabajo ha tenido como uno de sus objetivos principales la utilización y estudio de diferentes técnicas de minería para la creación y representación de perfiles de usuario. En un futuro, extenderemos el uso de los perfiles de usuario para desarrollar un sistema de recomendación que complete el proceso de personalización aquí iniciado.

Apéndice A

Lógica Difusa

La lógica difusa o borrosa parte del principio de que las cosas no son blancas o negras, tal como establece la lógica clásica, sino con tonalidades y con múltiples valores, lo cual se adapta mejor al comportamiento humano.

La información con la cual trabajamos diariamente no siempre presenta el grado de perfeccion que caracteriza a los modelos matematicos que utilizamos para su tratamiento automatico. En muchas ocaciones la información puede ser *incompleta* (sólo describe parcialmente la realidad), *imprecisa* (el valor de una variable se encuentra en un conjunto de valores, pero no podemos precisar cual es) e *incierta* (no tenemos total certerza de que la información sea verdadera).

Esto ha hehco que a lo largo de los años se hayan involucrados modelos matemáticos que permiten representar información imperfecta tales como la Teroría de la probabilidad, la Teoria de la Evidencia de Dempster/Shafer [Sha76] y la Teoría de Factores de Certeza [SB75]. Uno de dichos modelos es la Teoria de los Conjuntos Difusos [Zad75], propuesta por L.A. Zadeh en 1965 y que desde entonces ha experimentado un fuerte auge debido a las aportaciones de muchos otros investigadores.

La logica difusa se ha convertido en una potente herramienta a la hora de modelar sentencias de lenguaje natural, y razonar con las mismas tal y como lo hacen el ser humano, tipo de razonamiento que se ha dado en llamar razonamiento aproximado. La teoría de subconjuntos difusos y la logica difusa, ambos propuestos por L. A. Zadeh [Zad75], constituye los cimientos de la formalizacion de este tipo razonamiento.

A.1. Conjuntos Difusos

La Lógica Difusa actualmente está relacionada y fundamentada en la teoría de los Conjuntos Difusos. Según esta teoría, el grado de pertenencia de un elemento a un conjunto viene determinado por una función de pertenencia, que puede tomar todos los valores reales comprendidos en el intervalo [0,1]. La representación de la función de pertenencia de un elemento a un Conjunto Difuso se representa según la figura A.1.

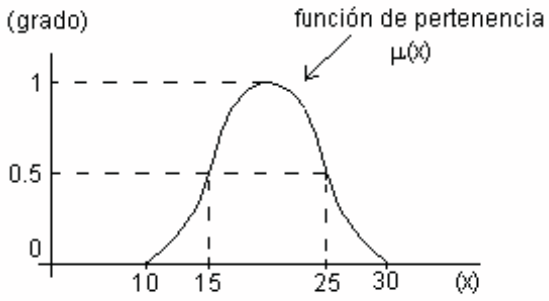

Figura A.1: *Formulación matricial de la Regresión Probabilística Lineal*

De una forma mas formal podemos decir que el concepto de un conjunto difuso sobre un universo X (tambien llamado preferecial) es una generalización del concepto gráfico de conjunto en el que la funcion indicadora (el que indica si un elemento pertenece o no a un elemento) tiene el intervalo real [0,1] en lugar del conjunto 0,1. Así el conjunto difuso A viene descrito por una funcion de pertenencia μ_A

$$\mu_A : X \longrightarrow [0,1] \tag{A.1}$$

Por lo general se suele identificar al conjunto con su función indicadora, que en el contexto de la teoria de conjuntos difusos se denomina funcion de pertenencia. Por lo tanto, seguiremos la siguiente notación.

$$A : X \longrightarrow [0,1] \tag{A.2}$$

de esta forma $A\,(x)$, donde $x \in X$, representa el grado de pertenencia del elemento *x* al conjunto *A* . Cuando el referencial es finito, $X = \{x_1, x_2, ..., x_n\}$, usaremos la

siguiente notación para representar a un conjunto difuso *A* sobre *X*:

$$A = \{A(x_1)/x_1 + A(x_2)/x_2 + ... + A(x_n)/x_n\} \tag{A.3}$$

Ahora comentaremos brevemente algunos conceptos basicos sobre los conjuntos difusos, donde:

- Se dice que un conjunto difuso *A* es normal si existe al menos un $x \in X$ tal que $A(x) = 1$.
- Se llama soporte del conjunto difuso *A* al conjunto $Sop(A) = \{x \in X | A(x) > 0\}$.
- Se llama núcleo del conjuntodifuso *A* al conjunto $Ker(A) = \{x \in X | A(x) = 1\}$.
- El conjunto de subconjuntos difusos sobre un referencial *X* se nota por $\ddot{P}(X)$.Es obvio que $P(X) \subset \ddot{P}(X)$.
- Se dice que un conjunto difuso *A* es convexo si verifica $x \leq y \leq z \Rightarrow A(y)min(A($ $x), A(z))$

A.1.1. Operaciones básicas con conjuntos difusos

La extensión del concepto de conjunto no tendría sentido sin extender simultáneamente las operaciones que podemos realizar con ellos. Las principales operaciones sobre conjuntos son la unión, intersección y el complemento. Estas operaciones pueden generalizarse al caso de los conjuntos difusos de diversas formas. La condición indispensable que deben verificar las extensiones es que, cuando los conjuntos inplicados son ordinarios, éstas deben comportarse como los operadores ordinarios. Las familias de operadores difusos mas importantes son los llamados *t-norma, t-conormas y negaciones*, que extienden las operaciones de *intersección, unión y complemento respectivamente.*

A.1.2. Operadores de interseccion: t-normas

Una t-norma es una función:

$$i : [0, 1] \times [0, 1] \longrightarrow [0, 1] \tag{A.4}$$

que verifica las siguientes propiedades para cualesquiera $a, b, c \in [0, 1]$:

- Frontera:$i(a,1) = a$
- Monotonía:$b \leq c \rightarrow i(a,b) \leq i(a,c)$
- Conmutatividad:$i(a,b) = i(b,a)$
- Asociatividad:$i(a,i(b,c)) = i(i(a,b),c)$

La interseccion de dos conjuntos difusos A y B mediante una t-norma i se define como:

$$(A \cap B)(x) - i(A(x), B(x)) \tag{A.5}$$

Algunas de las t-normas más utilizadas son las siguientes:

- Intersección Estandar: $i(a,b) = min(a,b)$
- Producto Algebráico: $i(a,b) = ab$
- Resta Acotada:$i(a,b) = max(0, a+b-1)$
- Intersección drástica:

$$i(a,b) = \begin{cases} a & \text{si} \quad b = 1 \\ b & \text{si} \quad a = 1 \\ 0 & \text{si} \quad otro\ caso \end{cases}$$

A.1.3. Operadores de unión: t-conorma

Una t-conorma es una función

$$u : [0,1] \times [0,1] \longrightarrow [0,1] \tag{A.6}$$

que verifica las siguientes propiedades para cualesquiera $a, b, c \in [0,1]$:

- Frontera:$u(a,0) = a$
- Monotonía:$b \leq c \rightarrow u(a,b) \leq u(a,c)$
- Conmutatividad:$u(a,b) = u(b,a)$

- Asociatividad:$u(a, i(b, c)) = u(i(a, b), c)$

La union de dos conjuntos difusos A y B mediante una t-norma *u* se define como:

$$(A \cup B)(x) - i(A(x), B(x)) \tag{A.7}$$

Algunas de las t-conormas más utilizadas son las siguientes:

- Intersección Estandar: $u(a, b) = min(a, b)$
- Suma Algebráica: $u(a, b) = a + b - ab$
- Resta Acotada:$u(a, b) = max(1, a + b)$
- Intersección drástica:

$$u(a,b) = \begin{cases} a & \text{si} \quad b = 1 \\ b & \text{si} \quad a = 1 \\ 1 & \text{si} \quad otrocaso \end{cases}$$

A.1.3.1. Operadores de complemento: negaciones

Una negación es una funcion

$$c : [0, 1] \longrightarrow [1, 0] \tag{A.8}$$

que verifica las siguientes propiedades para cualesquiera $a, b, c \in [0, 1]$:

- Frontera:$c(0) = 1$ y $c(1) = 0$
- Monotonía: si $a \leq b \rightarrow c(a) \geq c(b))$

Estas propiedades son mínimas que deben ser verificadas por una negación, pero se suelen exigir algunas otras propiedades para definir mejor negaciones desde un punto de vista práctico:

- Continuidad:c debe ser una función continua
- Propiedad involutiva: $c(c(a)) = a \forall a \in [0, 1]$

A.2. Funciones de implicación

La lógica difusa generaliza los operadores de conjunción, disyunción y negación de la lógica difusa clásica mediante el uso de t-normas, t-conormas y negaciones. El operador de implicancia de la lógica clasica se extiende mediante una familia de operadores difusos llamados implicaciones. Los operadores de implicación difusa verifican que en el caso de conjuntos ordinarios se reducen a la implicación clasica. Una implicación difusa es una función:

$$I : [0,1] \times [0,1] \longrightarrow [0,1] \tag{A.9}$$

la cual verifica las siguientes propiedades para todo $a, b \in [0,1]$ [TV84]:

- Si $a \leq b$ entonces $I(a,x) \geq I(b,x)$
- $I(0,x) = 1$
- $I(1,x) = x$
- $I(a, I(b,x)) = I(b, I(a,x))$

Entre los operadores de implicación destacan dos importantes subfamilias de operadores llamadas S-implicaciones y R-implicaciones.

Las S-implicaciones se definen mediante el uso de una t-conorma u y una negación c como:

$$I(a,b) = u(c(a), b) \tag{A.10}$$

Algunas S-implicaciones son las siguientes:

- Kleenes-Dianes:$I_b(a,b) = max(1-a, b)$
- Reichenbach:$I_r = 1 - a + ab$
- Lukasiewiez:$I_a = min(1, 1 - a + b)$

- Intersección drástica:

$$I_{ls} = \begin{cases} b & \text{si} \quad a = 1 \\ 1 - a & \text{si} \quad b = 1 \\ 1 & \text{si} \quad otrocaso \end{cases}$$

La implicación drástica es la mayor de las S-implicaciones. Las R-implicaciones se definen mediante una t-norma continua *i* según la expresión.

$$I(a, b) = sup\{x \in [0, 1] i(a, x) \leq b\} \tag{A.11}$$

Algunas R-implicación son las siguientes:

- Godel:

$$I_g(a, b) = \begin{cases} 1 & a \leq b \\ b & a > b \end{cases}$$

- Goguen:

$$I_\Delta(a, b) = \begin{cases} 1 & a \leq b \\ b/a & a > b \end{cases}$$

- Lukasiewiez:$I_a(a, b) = min(1, 1 - a + b)$
- Drástica:

$$I_{LR}(a, b) = \begin{cases} b & a = 1 \\ 1 & otro \ \ caso \end{cases}$$

Como puede apreciase la implicación de Lukasiewiez pertenece a ambas familias. La implicación drástica es el menor de las cotas superiores de R-implicaciones.

A.3. Variables Lingüísticas

La Teoría de Conjuntos Difusos puede utilizarse para representar expresiones lingüísticas que se utilizan para describir conjuntos o algoritmos. Los Conjuntos Difusos son capaces de captar por sí mismos la vaguedad lingüística de palabras y frases comúnmente aceptadas, como *gato pardo o ligero cambio*. La habilidad humana de comunicarse mediante definiciones vagas o inciertas es un atributo importante de la inteligencia

En la lógica difusa, los conceptos imprecisos se representan mediante una *etiqueta lingüística* [Zad75]. Las etiquetas lingüísticas son en definitivas conjuntos difusos mediante lo que se da semántica a un identificador que representa un concepto o un calificativo. Un ejemplo de etiqueta es alto, esta etiqueta se representa mediante un conjunto difuso que tiene como referencial el conjunto de las alturas posibles. Una representación subjetiva de esta etiqueta la podemos ver en la Figura A.2.

De manera informal, una variable lingüística es una variable que toma valores dentro de un dominio formado por etiquetas lingüísticas, definidas sobre un referencial adecuado. Por ejemplo, la variable lisngüística altura de una persona podría tomar valores dentro del dominio $\{bajo, medio, alto, muy\ alto\}$. Estas etiquetas son conjuntos difusos cuyo referencial es el conjunto (numerico) de las alturas posibles de una persona. Una representación subjetiva de las etiquetas mensionadas se muestran en la Figura A.2.

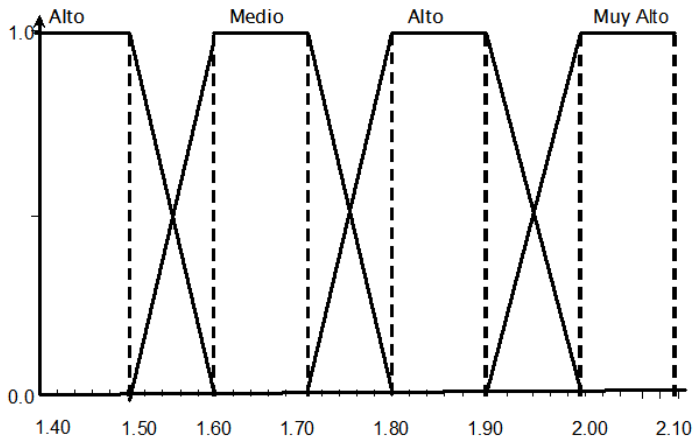

Figura A.2: *Etiquetas lingüística para la variable altura.*

El uso de estas variables permite un razonamiento más cercano al nuestro, ya que disminuye la granularidad del dominio de los valores de una variable. De esta forma es más sencillo definir reglas y conceptos, semánticamente intuitivos, que involucren variables con un referencial numérico o con un alto número de valores, como es el caso de la variable .ªltura".

Para cuantificar el grado de relacion entre los valores (etiquetas) de distintas variables lingüísticas en base a los datos almacenados en la base de datos, basta con medir los cardinales y cardinales relativos de los conjuntos difusos inducidos por dichas etiquetas, y utilizarlos para la evaluacion de sentencias cuantificadas.

Apéndice B

Reglas de Asociación

Dado un conjunto de ítems, las reglas de asociación describen como varias combinaciones de ítems están apareciendo juntas en los mismos conjuntos. Una típica aplicación de regla de asociación está dentro del análisis llamado ”datos del carro del supermercado”. El objetivo es encontrar regularidades en los comportamientos de los clientes dentro de términos de combinaciones de productos que son comprados muchas veces en un conjunto, o sea reglas que reflejen relaciones entre los atributos presentes en los datos.

El descubrimiento de reglas de asociación busca relaciones o afinidades entre los conjuntos de items (ítem sets). Un conjunto de artículos se define como cualquier combinación formada por dos o más artículos diferentes de todos los artículos disponibles.

Una regla de asociación se forma por dos conjuntos: *el antecedente y el consecuente*. Las reglas generalmente se escriben con una flecha apuntando hacia el consecuente desde el antecedente, por ejemplo $0123 \longrightarrow 4567$. Una regla de asociación indica una afinidad entre el antecedente y el consecuente, y generalmente está acompañada por estadísticos basados en frecuencias que describen esta relación.

Las medidas más utilizadas para describir las relaciones entre antecedente y consecuente son el *Soporte (Supp)*, y *la Confianza (Conf)*, los cuales son valores numéricos. Para describirlos necesitamos de algunas definiciones previamente.

Podemos definir a*I* como un conjunto de ítems y a *T* como un conjunto de transacciones con ítems en *I* , ambos conjuntos finitos. Una regla de Asociación que es de la forma $A \longrightarrow C$, donde $A, C \subseteq I$, $A, C \neq 0$ y $A \cap C = 0$. La regla $A \longrightarrow C$ significa que ”toda transacción *T* que *A* contiene *C* contiene”.

El soporte de un ítemset $I_0 \subseteq I$ es:

$$supp(I_0, T) = \frac{|\{t \in T | I_0 \subseteq t\}|}{|T|} \tag{B.1}$$

es la probabilidad de que una transacción *T* contenga I_0. Y el Soporte de una regla de asociación $A \longrightarrow C$ en *T* es:

$$Supp(A \longrightarrow C, T) = supp(A \cup C) \tag{B.2}$$

y la confianza es:

$$Conf(A \longrightarrow C, T) = \frac{supp(A \cup C)}{supp(A)} = \frac{Supp(A \longrightarrow C)}{supp(A)} \tag{B.3}$$

El soporte indica que porcentaje de los atributos de una regla aparecen con valor positivo dentro de transacciones de un conjunto de datos. Y la confianza es la razón probabilística de *C* con respecto a *A*, en otras palabras es la cardinalidad relativa de *C* con respecto a *A*.

Las técnicas utilizadas para poder encontrar las reglas de asociación, intentan descubrir reglas cuyo soporte y confianza es mayor o igual a dos umbrales, los cuales son determinados por el usuario, llamados *Minsupp* y *Minconf*, respectivamente. Tales reglas son llamadas reglas fuertes.

A continuación se ilustra el cálculo del soporte y la confianza con un pequeño grupo de transacciones:

{ciruela, lechuga, tomates}
{apio, dulcería}
{dulcería}
{manzanas, zanahorias, tomates, papas, dulcería}
{manzanas, naranjas, lechuga, tomates, dulcería}
{duraznos, naranjas, apio, papas}
{frijoles, lechuga, tomates}
{naranjas, lechuga, zanahorias, tomates, dulcería}
{manzana, plátanos, ciruelas, zanahorias, tomates, cebolla, dulcería}
{manzana, papas}

Se puede ver, si se quiere obtener el soporte de manzana, de 10 transacciones disponibles 4 contienen a manzana, por lo que supp(manzana)=$4/10 = 0{,}4$, igualmente para el soporte de la zanahoria, que hay 3 transacciones que la contienen, así el:

- supp(zanahoria)=$3/10 = 0{,}3$
- supp(dulcería)=$0{,}6$
- Supp(manzana $\rightarrow$ dulcería)=$0{,}3$
- Supp(manzana $\rightarrow$ tomate)=$0{,}3$

Si el soporte es suficientemente alto y el conjunto de transacciones es grande, entonces la confianza es un estimado de la probabilidad que cualquiera transacción futura que contenga el antecedente, contendrá el consecuente. Del ejemplo se ve que:

- Conf(manzana $\rightarrow$ dulcería)=Supp(manzana $\rightarrow$ dulcería)$\div$ supp(manzana)=$0{,}3/0{,}4 = 0{,}75$
- Conf(manzana $\rightarrow$ tomates)=$0{,}75$
- Conf(zanahoria $\rightarrow$ tomates)=$1{,}0$

El algoritmo de asociación tratará de descubrir todas las reglas que excedan las cotas mínimas especificadas para el soporte y confianza. La búsqueda exhaustiva de reglas de asociación consideraría simplemente todas las combinaciones posibles de los elementos, poniéndolos como antecedentes y consecuentes, entonces se evaluaría el soporte y la confianza de cada regla, y se descartarían todas las reglas que no satisfacen las restricciones.

B.1. Algoritmo APrioriTID

El objetivo en todo algoritmo de búsqueda de reglas de asociación es encontrar las reglas que satisfacen con la condición de confianza y soporte mínimo, y en este caso también el factor de certeza mínimo. Esto es necesario por que sino la búsqueda se haría exhaustiva, encontrándose al final una cantidad demasiado grandes de reglas generadas, y no sabiendo que reglas son las que mejor representan el conjunto de datos.

Cuando se desea realizar una búsqueda en grandes conjuntos de datos, se debe tratar de minimizar la cantidad de tiempo que se emplea en acceder las mismas, por cuanto estas operaciones de acceso a disco son por lo general las más lentas del proceso.

El problema de la búsqueda y extracción de reglas de asociación en general suele descomponerse en dos pasos:

- *Items relevantes (Large itemsets).* Encontrar todos los conjuntos de ítems con relevancia por encima del mínimo soporte. Los conjuntos de ítems cuya relevancia quede por encima del umbral serán los conjuntos de ítems relevantes (large itemsets) y mientras los que queden por debajo no nos interesa (serán los "small itemsets").
- *Reglas de asociación.* Genera las reglas de asociación utilizando los conjuntos de ítems relevantes obtenido.

Por lo tanto, el algoritmo APrioriTID presentado por Agrawal en [AR94], se caracteriza porque no accede a la base de datos para obtener la relevancia de los candidatos. Para ello utiliza los conjuntos auxiliares *CT[k]*.

Cada miembro del conjunto auxiliar $CT[k]$ es de la forma $< TID, X >$, donde cada *X* es un k-itemset potencialmente relevante (un candidato) presente en la transacción identificada por TID. Evidentemente, *CT[1]* se corresponde a la Base de Datos original en la cual cada ítem *i* es reemplazado por el ítemset$\{i\}$. El elemento de *CT[k]* correspondiente a la transacción *t* es el par $< TID_t, \{c \in C[k] c \subseteq t\} >$. Si una transacción no contiene ningún k-ítemset candidato no tendrá una entrada en *CT[k]*. Se muestra en la figura B.1 el algoritmo *APrioriTID*.

La característica principal del APrioriTID es que, en cada iteración, se recorre el conjunto $CT[k-1]$ en vez de la Base de Datos completa para obtener la relevancia de los itemsets de *C[k]*.

El tamaño de los conjuntos auxiliares CT[k] puede llegar a ser mucho menor que el de la Base de Datos original tras unas cuantas iteraciones del algoritmo (para valores grandes de k), lo que ahorra esfuerzo consumido en realizar operaciones de entrada y salida. Sin embargo, para valores pequeños de k (especialmente cuando k vale 2 o 3), las entradas correspondientes a cada transacción en CT[k] puede ocupar más espacio que las propias transacciones en la Base de Datos original: los conjuntos CT[k] puede llegar a ser mayores que la base de datos inicial para valores pequeños de k.

Luego de describir el algoritmo también es importante comentar que la búsqueda inicial de las reglas permite encontrar todas las asociaciones que satisfagan las restricciones

```
L [1] = { large 1-itemsets}
CT [1] = Base de Datos D
K = 2
Mientras L [k-1]≠Ø
        C [k] = candidatosAPRIORI ( L [k-1])
        CT [k] = Ø
        Para cada entrada t ∈ CT [k-1]
                C_t = conjunto de candidatos de C [k] contenidos en t (usando TID)
                Para cada candidato c ∈ Ct
                    c.contador++
                Si C_t≠Ø
                    CT [k] +=< t.TID, Ct >
        L [k] = { c ∈ C [k] | c.contador > MinSupport}
        k++
```

Figura B.1: *Algoritmo AprioriTID*

iniciales de soporte y confianza. Esto puede llevar a obtener una gran cantidad de reglas de asociaron a partir de los datos, los cuales no serían manejables. Luego, es importante reducir el número de reglas de manera que solo queden las más interesantes. Para eso utilizaremos otra medida de las reglas de asociación, ya antes definida, que es el Factor de Certeza.

La generación del conjunto candidato C[k] se realiza directamente a partir de los conjuntos de ítems relevantes $L[k-1]$. En primer lugar, se generan posibles candidatos a partir del producto cartesiano $L[k-1] \times L[k-1]$ imponiendo las restricciones de que los $k-2$ primeros ítems de los elementos de $L[k-1]$ han de coincidir. Acto seguido se eliminan del conjunto de candidatos aquellos itemsets que contienen algún $[k-1]-itemset$ que no se encuentre en $L[k-1]$.

Por ejemplo, supongamos que L[3]={$\{123\}$,$\{124\}$,$\{134\}$,$\{135\}$,$\{234\}$}.
Tras la reunión C[4] será {$\{1234\}$,$\{1345\}$}. La poda del conjunto de candidatos eliminará el ítemset $\{1345\}$ porque el ítemset $\{145\}$ no está en L[3]. Por lo tanto, C[4] sólo incluirá a $\{1234\}$.

Apéndice C

Resultados de los Perfiles encontrados en el sitio ETSIIT

En este apartado podemos ver los resultados obtenidos de los perfiles de usuario encontrados automáticamente del sitio de la Escuela Técnica Superior de Ingeniería Informática y de Telecomunicaciones (ETSIIT).

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil1>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=foro&amp;idforo=generalHTPP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;idHebra=1937HTPP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.pho?action=hebra&amp;idhebra=1916HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=foro&amp;idforo=escuelaHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1709&amp;page=1HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/apps/foro/index.php?action=foro&amp;idforo=asignaturasHTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">General</Terminos>
    <Terminos Termino="5">"Index</Terminos>
    <Terminos Termino="6">js/protWindows/themes</Terminos>
    <Terminos Termino="7">"ETSIIT</Terminos>
    <Terminos Termino="8">Escuela</Terminos>
    <Terminos Termino="9">Asignaturas</Terminos>
    <Terminos Termino="10">Docencia</Terminos>
    <Terminos Termino="11">tablón</Terminos>
  </Perfil_Simple>
</Perfil1>
```

Figura C.1: *Perfil 1*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil2>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1939HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Muy Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">tablón</Terminos>
  </Perfil_Simple>
</Perfil2>
```

Figura C.2: *Perfil 2*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil3>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
  <Pagina Pagina_Visitada="0">GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1317HTTP/1.1</Pagina>
  <Pagina Pagina_Visitada="1">*GET/alumnos/shin/shin.htmHTTP/1.1*</Pagina>
  <Pagina    Pagina_Visitada="2">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=compra&amp;amp;page=1HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=gnulinuxHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="4">*GET/alumnos/mu01/guerraSoftware.htmlHTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Poco Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
  <Terminos Termino="0">"Ingeniería</Terminos>
  <Terminos Termino="1">Informática</Terminos>
  <Terminos Termino="2">Telecomunicación</Terminos>
  <Terminos Termino="3">Foros</Terminos>
  <Terminos Termino="4">busco</Terminos>
  <Terminos Termino="5">chips</Terminos>
  <Terminos Termino="6">wii</Terminos>
  <Terminos Termino="7">compra</Terminos>
  <Terminos Termino="8">venta</Terminos>
  <Terminos Termino="9">proximo</Terminos>
  <Terminos Termino="10">robo</Terminos>
  <Terminos Termino="11">portatil</Terminos>
  </Perfil_Simple>
</Perfil3>
```

Figura C.3: *Perfil 3*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil4>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=foro&amp;idforo=escuelaHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1583HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.php?action=foro&amp;idforo=generalHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1874HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1709HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/page.php?pageid=departamentosHTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">tablón</Terminos>
    <Terminos Termino="5">"Index</Terminos>
    <Terminos Termino="6">js/protWindows/themes</Terminos>
    <Terminos Termino="7">"Escuela</Terminos>
    <Terminos Termino="8">comprar</Terminos>
    <Terminos Termino="9">coche</Terminos>
    <Terminos Termino="10">General</Terminos>
  </Perfil_Simple>
</Perfil4>
```

Figura C.4: *Perfil 4*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil5>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/js/protWindows/themes/default.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1939HTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Poco Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Index</Terminos>
    <Terminos Termino="1">js/protWindows/themes</Terminos>
    <Terminos Termino="2">tablón</Terminos>
    <Terminos Termino="3">"Ingeniería</Terminos>
    <Terminos Termino="4">Informática</Terminos>
    <Terminos Termino="5">Telecomunicación</Terminos>
    <Terminos Termino="6">Foros</Terminos>
  </Perfil_Simple>
</Perfil5>
```

Figura C.5: *Perfil 5*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil6>
 <Identificacion_Usuario>
  <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
 <Paginas>
  <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1939HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="1">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1874&amp;amp;page=2HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="2">*GET/usuarios/jmlvega/idragon//formate.cssHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1617&amp;amp;page=1HTTP/1.*</Pagina>
  <Pagina Pagina_Visitada="4">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1709&amp;amp;page=1HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="5">*GET/alumnos/diegorp/canalplus.htmlHTTP/1.1*</Pagina>
  </Paginas>
 <Var_Demograficas>
  <Edad>Joven</Edad>
  <Paciencia>Paciente</Paciencia>
  <Idioma>Español</Idioma>
  </Var_Demograficas>
 <Perfil_Simple>
  <Terminos Termino="0">"Ingeniería</Terminos>
  <Terminos Termino="1">Informática</Terminos>
  <Terminos Termino="2">Telecomunicación</Terminos>
  <Terminos Termino="3">Foros</Terminos>
  <Terminos Termino="4">sitio</Terminos>
  <Terminos Termino="5">Diego</Terminos>
  <Terminos Termino="6">Rodero</Terminos>
  <Terminos Termino="7">Internet</Terminos>
  </Perfil_Simple>
</Perfil6>
```

Figura C.6: *Perfil 6*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil7>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">* *GET/apps/foro/index.php?action=foro&amp;idforo=asignaturasHTTP/1.1**</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/alumnos/juliolo/principal.htmlHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4"> *GET/apps/foro/index.php?action=hebra&amp;idhebra=1752HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1885HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6"> *GET/page.php?pageid=planesHTTP/1.1</Pagina>
    </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Muy Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">tablón</Terminos>
    <Terminos Termino="5">ETSIIT</Terminos>
    <Terminos Termino="6">Página</Terminos>
    <Terminos Termino="7">Culturista</Terminos>
    <Terminos Termino="8">Planes</Terminos>
    <Terminos Termino="9">Estudio</Terminos>
  </Perfil_Simple>
</Perfil7>
```

Figura C.7: *Perfil 7*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil8>
 <Identificacion_Usuario>
  <Tipo Tipo="alumno" />
 </Identificacion_Usuario>
 <Paginas>
  <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=foro&amp;idforo=escuela&amp;page=3HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="1">*GET/apps/foro/index.php?action=foro&amp;idforo=programacion&amp;page=1HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="2">*GET/alumnos/mlii/Algoritmo.htmHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1628HTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="4">*GET/page.php?pageid=grupos_investigacionHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="5">*GET/alumnos/mlii/ArqVonNeumann.htmHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="6">*GET/alumnos/mlii/Lenguajes.htmHTTP/1.1*</Pagina>
  <Pagina Pagina_Visitada="7">*GET/alumnos/mlii/index.htmlHTTP/1.1*</Pagina>
  </Paginas>
 <Var_Demograficas>
  <Edad>Joven</Edad>
  <Paciencia>Muy Paciente</Paciencia>
  <Idioma>Español</Idioma>
 </Var_Demograficas>
 <Perfil_Simple>
  <Terminos Termino="0">"Ingeniería</Terminos>
  <Terminos Termino="1">Informática</Terminos>
  <Terminos Termino="2">Telecomunicación</Terminos>
  <Terminos Termino="3">Escuela</Terminos>
  <Terminos Termino="4">Foro</Terminos>
  <Terminos Termino="5">Programación</Terminos>
  <Terminos Termino="6">Asignatura</Terminos>
  <Terminos Termino="7">Docencia</Terminos>
  <Terminos Termino="8">Grupos</Terminos>
  <Terminos Termino="9">Investigacion</Terminos>
  <Terminos Termino="10">Lenguajes</Terminos>
  <Terminos Termino="11">Computacion</Terminos>
 </Perfil_Simple>
</Perfil8>
```

Figura C.8: *Perfil 8*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil9>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1916HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.php?action=foro&amp;idforo=asignaturasHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/foro/index.php?action=foro&amp;idforo=escuelaHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1709&amp;vpage=1HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1874&amp;page=2HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1892HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1939HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/apps/foro/index.php?action=foro&amp;idforo=generalHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="9">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="10">*GET/apps/foro/index.php?action=foro&amp;idforo=deportesHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="11">*GET/apps/foro/index.php?action=hebra&amp;idhebra=1752HTTP/1.1*</Pagina>
  </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Muy Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">Foros</Terminos>
    <Terminos Termino="4">Prácticas</Terminos>
    <Terminos Termino="5">Asignatura</Terminos>
    <Terminos Termino="6">Docencia</Terminos>
    <Terminos Termino="7">Escuela</Terminos>
    <Terminos Termino="8">Plataforma</Terminos>
    <Terminos Termino="9">Examenes</Terminos>
    <Terminos Termino="10">actualizada</Terminos>
    <Terminos Termino="11">tablón</Terminos>
    <Terminos Termino="12">Deportes</Terminos>
  </Perfil_Simple>
</Perfil9>
```

Figura C.9: *Perfil 9*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil10>
  <Identificacion_Usuario>
    <Tipo Tipo="profesor" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/usuarios/jmlvega/idragon//formate.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/convocatorias/styles/convocatorias.cssHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/profesores/jmaroza/anecdotario/chmanual.htmHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/profesores/jmaroza/anecdotario/anecdotario-z.htmHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/planes/index.php?id=3&amp;amp;id2=127HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/page.php?pageid=horarioHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1617HTTP/1.1*</Pagina>
    </Paginas>
  <Var_Demograficas>
    <Edad>Adulto</Edad>
    <Paciencia>Muy Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
    <Terminos Termino="0">"Ingeniería</Terminos>
    <Terminos Termino="1">Informática</Terminos>
    <Terminos Termino="2">Telecomunicación</Terminos>
    <Terminos Termino="3">"Index</Terminos>
    <Terminos Termino="4">skin/reloaded</Terminos>
    <Terminos Termino="5">convocatorias</Terminos>
    <Terminos Termino="6">ubuntu"</Terminos>
    <Terminos Termino="7">planes"</Terminos>
    <Terminos Termino="8">estudio"</Terminos>
    <Terminos Termino="9">Horario"</Terminos>
  </Perfil_Simple>
</Perfil10>
```

Figura C.10: *Perfil 10*

```
<?xml version="1.0" encoding="UTF-8"?>
<Perfil11>
  <Identificacion_Usuario>
    <Tipo Tipo="alumno" />
  </Identificacion_Usuario>
  <Paginas>
    <Pagina Pagina_Visitada="0">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="1">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=asignaturasHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="2">*GET/apps/foro/index.php?action=foro&amp;amp;idforo=generalHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="3">*GET/profesores/jmaroza/anecdotario/anecdotario-z.htmHTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="4">*GET/apps/tablon/HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="5">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1819&amp;amp;page=0HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="6">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=696HTTP/1.1*</Pagina>
    <Pagina Pagina_Visitada="7">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1349HTTP/1.0*</Pagina>
    <Pagina Pagina_Visitada="8">*GET/page.php?pageid=googlemapsHTTP/1.1*</Pagina>
    </Paginas>
  <Var_Demograficas>
    <Edad>Joven</Edad>
    <Paciencia>Paciente</Paciencia>
    <Idioma>Español</Idioma>
  </Var_Demograficas>
  <Perfil_Simple>
   <Terminos Termino="0">"Ingeniería</Terminos>
   <Terminos Termino="1">Informática</Terminos>
   <Terminos Termino="2">Telecomunicación</Terminos>
   <Terminos Termino="3">Foros</Terminos>
   <Terminos Termino="4">Asignatura</Terminos>
   <Terminos Termino="5">General</Terminos>
   <Terminos Termino="6">Anecdotario</Terminos>
   <Terminos Termino="7">Googlemaps</Terminos>
  </Perfil_Simple>
</Perfil11>
```

Figura C.11: *Perfil 11*

```
<?xml version="1.0" encoding="UTF-8" ?>
<Perfil12>
<Identificacion_Usuario>
<Tipo Tipo="alumno" />
</Identificacion_Usuario>
<Paginas>
<Pagina Pagina_Visitada="0">*GET/page.php?pageid=rss_baseHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="1">*GET/apps/tablon/HTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="2">*GET/apps/foro/index.phpHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="3">*GET/guias/actual/Guia.pdfHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="4">*GET/guias/actual/Indice.pdfHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="5">*GET/guias/actual/Presentacion.pdfHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="6">*GET/page.php?pageid=infocentroHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="7">*GET/page.php?pageid=descargasHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="8">*GET/apps/descargas/HTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="9">*GET/apps/descargas/styles/descargas.cssHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="10">*GET/apps/descargas/index.php?id=guiasHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="11">*GET/apps/descargas/index.php?id=secretariaHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="12">*GET/page.php?pageid=webinfoHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="13">*GET/page.php?pageid=wemapHTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="14">*GET/apps/foro/index.php?action=hebra&amp;amp;idhebra=1935HTTP/1.1*</Pagina>
<Pagina Pagina_Visitada="15">*GET/page.php?pageid=rriiExtranjeroHTTP/1.1*</Pagina>
</Paginas>
<Var_Demograficas>
<Edad>Joven</Edad>
<Paciencia>Paciente</Paciencia>
<Idioma>Español</Idioma>
</Var_Demograficas>
<Perfil_Simple>
<Terminos Termino="0">Noticias</Terminos>
<Terminos Termino="1">RSS</Terminos>
<Terminos Termino="2">Base</Terminos>
<Terminos Termino="3">Ingenierías</Terminos>
<Terminos Termino="4">Informática</Terminos>
<Terminos Termino="5">Telecomunicación</Terminos>
<Terminos Termino="6">tablón</Terminos>
<Terminos Termino="7">Foros</Terminos>
<Terminos Termino="8">Index</Terminos>
<Terminos Termino="9">/guias/actual</Terminos>
<Terminos Termino="10">Infirmación</Terminos>
<Terminos Termino="11">Centro</Terminos>
<Terminos Termino="12">Descargas</Terminos>
<Terminos Termino="13">Nueva</Terminos>
<Terminos Termino="14">Web</Terminos>
<Terminos Termino="15">Estudio</Terminos>
<Terminos Termino="16">extranjero</Terminos>
</Perfil_Simple>
</Perfil12>
```

Figura C.12: *Perfil 12*

Apéndice D

Weka

El sistema WEKA (Waikato Environment for Knowledge Analysis) fue desarrollado en la Universidad de Waikato en Nueva Zelanda. Está implementado en el lenguaje de programación Java y ha sido probado en los ambientes operativos Windows, Linux y Macintosh. Implementa algoritmos de minería de datos que pueden aplicarse a bases de datos desde su línea de comando o bien desde su interfaz gráfica.

Este sistema incluye una variedad de herramientas para transformar conjuntos de datos. Permite realizar preprocesamientos de datos para transformarlos en un esquema de aprendizaje, a fin de que sus resultados puedan ser analizados. Una manera de usar WEKA es aplicar un método de aprendizaje a conjuntos de datos y analizar los resultados para extraer información. Otra es aplicar varios métodos de aprendizaje y comparar sus resultados en orden de escoger una predicción. Estos métodos son llamados clasificadores.

La implementación de los esquemas de aprendizaje son los recursos más valiosos del sistema. Las herramientas para el Preprocesamiento de datos, llamados filtros, son los segundos más importantes. La atención de WEKA se centra en los algoritmos de clasificación y filtro, sin embargo, también incluye la implementación de algoritmos para el aprendizaje de reglas de asociación y el agrupamiento de datos (clustering).

Este software de distribución gratuita es posible obtenerlo del siguiente sitio de Internet:{ *http://w*
ww.cs.waikato.ac.nz/ ml}, donde se encuentra una versión con java incluido para que pueda correr en cualquier Sistema Operativo.

Una vez instalado el software weka, se deben preparar los datos para poder hacer minería sobre ellos. Para esto se crea un archivo, que es un formato de relaciones de

atributos (ARFF); este archivo es un archivo de texto de ASCII que describe una línea de casos que comparten un set de atributos. ARFF archivos fueron desarrollados por la Machines Learning del Proyecto del Departamento de Informática de la Univerisdad de Waikato para el desarrollo de esta herramienta de software llamada WEKA.

Se indicará una breve descripción de cómo es el formato de este archivo. ARFF archivo tienen dos secciones distintas; la primera es la información del encabezado, que es seguida por la información de los Datos. El encabezado del archivo ARFF contiene el nombre de la relación, una lista de los atributos (las columnas en los datos), y sus tipos. Un ejemplo de encabezado de un archivo ARFF se puede ver en la figura D1: (las lineas que empiezan con % son comentarios)

```
% 1.Título Base de Datos de Plantas de Lirio (Iris)
% 2.Fuentes: a) Creador: R. A. de Pescador; b) Donante: Michael Marshall
% fecha: junio 1998
@ relation Lirio (Iris)

@ ATTRIBUTE  sepallength
@ ATTRIBUTE NUMERICO sepalwidth
@ ATTRIBUTE NUMERICO petallength
@ ATTRIBUTE NUMERICO petallength
Clase numérica @ ATTRIBUTE (Lirio ( iris)-setosa, Lirio ( iris)- versicolor,Li)
```

Figura D.1: *Formato ARFF*

La declaración @relation es definido como la primera linea en el archivoARFF. La sintaxis es: @relation $< realtion - nombre >$, donde lo que va entre <> se escribe el nombre de la base de datos o archivo que se va a ocupar; la declaraciones de atributo toman la forma de una secuencia de orden de declaraciones @attribute que unicamente define el nombre de aquel atributo y el tipo de dato. El orden de los atributos declarados indican la posicion de las columnas en la seccion de datos del archivo. Por ejemplo, si un atributo es el tercero declarado entonces WEKA espera que todo los valores serán encontrados en la tercera columnas. La sintaxis para la declaración de @attribute es: @attribute $< attribute - nombre >< tipodedato >$, donde se debe comenzar con un carácter alfabético. Los atributos pueden ser numericos , nominales, fechas o string.

La seccion de los datos ARFF swl archivo contiene la linea de declaración de los datos y las líneas de los casos reales. La declaracion @data es uan línea sola que denota

el principio del segmento de datos en el archivo.

Los datos reales se representan sobre una sola línea, conretorno de carro que denotan el final del dato real; los valores de los atributos para cada caso o valor son delimitados por comas. Ellos deben aparecer en el orden que ellos fueron declarados en la seccion de encabezado del archivo. En el caso que algun dato sea desconocido debe ser representado con un signo de interrogacion (?)

WEKA se centra en los algoritmos de clasificación y filtro, como también incluye la implementación de algoritmos para el aprendizaje de reglas de asociación y el agrupamiento de datos (clustering). Cualquier algoritmo de estudio en WEKA es sacado de la base de Clasificador; sorprendentemente poco es necesario para un clasificador básico: una rutina que genera un modelo de clasificador de un entrenamiento de una dataset (buildClassifier) y otra rutina que evalúa el modelo generado sobre una prueba no vista en el dataset (classifyInstance). Un modelo clasificador es una traza de un mapa de complejo arbitrario ”de todos excepto un”dataset atributo al atributo de la clase. La forma especifica y la creación de esta traza, o el modelo, se diferencian del clasificador al clasificador. Por ejemplo, el modelode ZeroR solamente consiste en un solo valor: la clase más común, o la mediana de todos los valores numéricos en caso de predicción de un valor numérico. ZeroR es un clasificador trivial, pero esto da un mas bajo funcionamiento de un dato de dataset que considerablemente debería ser mejorado por clasificadores más complejos.

Dentro del software también nos da la facilidad de filtrar los datos, de esa forma estos filtros se preocupan de las clases que transforman los conjuntos de datos o dataset. WEKA ofrece el apoyo útil para el proceso previo de datos. Se debe tener en cuenta que la mayor parte de los clasificadores en WEKA utilizan la transformación de los datos con los filtros internamente.

Los clasificadores están en el corazón del WEKA. Hay muchas opciones comunes para clasificadores, la mayor parte del cual son relacionados con objetivos de evaluación. Ahora se muestra una lista de selección de los clasificadores en WEKA, los cuales son:

- *Weka.classifiers.trees.j48*. Para crear árboles de decisión podados a través del algoritmo C4.5.

- *Weka.classifiers.bayes. NaivesBayes-k*. Valores precisos numéricos son escogidos basados en el análisis basados en los datos que se entrenan, usando una presicion del 0.1 para atributos numéricos cuando llaman buildclassifier con el cero que entrena casos.

```
Función C4:5
(R: conjunto de atributos no clasificados,
C: atributo clasificador,
S: conjunto de entrenamiento ) devuelve  un árbol de decisión.
Comienzo
Si S está vacío,
   Devolver un único nodo con valor falla.
Si todos los registros de S tienen el mismo valor para el atributo clasificador.
    Devolver un único nodo con dicho valor.
Si R está vacío
    Devolver un único nodo con el valor más frecuente del atributo clasificador en los
    registros de S [Nota: habrá errores, es decir, registros  que no estarán bien]
    clasificados en este caso];
Si R no está vacío
    D ?  atributo con mayor Proporción  de Ganancia (D,S) entre los atributos de R:
    Sea {dj | j = 1,2,...,m} los valores del atributo D
    Sea {dj | j = 1,2,...,m} los subconjuntos de S correspondientes a lso valores de dj
    respectivamente.
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados como d1, d2,...,dj
que van respectivamente a los árboles
C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm),
```

Figura D.2: *Pseudo-código algoritmo C4.5*

- *Weka.classifiers.functions.* SMO (algoritmo optimización secuencial minima de Platt, para entrenar un vector clasificador que usa polinomios, en si el algoritmo es capaz de normalizar los datos perdidos y transformar datos nominales a binarios. Para obtener estimaciones de probabilidad apropiadas, usa la opción que encaja modelos de regresión logísticos a las salidas del support vector machine).

- *Weka.classifiers.lazy.kstar.* KStar (K* es un clasificador basado por caso, que es la clase de un caso de prueba lo es basado en la clase de aquellos casos de entrenamiento similares, como determinado por alguna función de semejanza. La asunción subyacente de clasificadores basados por caso como la K *, IB1, PEBLS, etc., es que casos similares tendrán clases similares).

- *Weka.classifiers.rules.* JRip (Esta clase pone en práctica a un principiante de regla lógico, la Poda Repetida Incremental para Producir la Reducción de Error (RIPPER)).

- *Weka.classifiers.fuctions.* LinearRegresion (multirespuesta regression lienal) (Clase para usar regresión linear para predicción. Usa el criterio Akaike para la selección modela, y es capaz de tratar con casos ponderados)

- *Weka.classifiers.meta.* ClassificationViaRegression-W (Clase para hacer clasificación que usa métodos de regresión)

Después de haber mencionado la parte de los clasificadores del WEKA comentaremos la parte de clustering que trae el sistema, aquí el software trabaja con 3 tipos de algoritmos o cluster los cuales son EM, K-means y el Cobweb. Se dará ciertas definiciones para entender cada uno de estos métodos de clustering.

EM, este método empieza obteniendo los parámetros de las distribuciones y los usa para calcular las probabilidades de que cada objeto pertenezca a un cluster y usa es probabilidad para re-estimar los parámetros de las probabilidades , hasta converger. Es recomendado hacer varias veces este proceso para garantizar la convergencia de los datos. K-means, este método selecciona elementos aleatorios, los cuales representan el centro o media de cada cluster. A cada objeto restante se le asigna el cluster con los cuales más se parece, basándose en una distancia entre el objeto y la media del cluster. Después calcula la nueva media del cluster e itera hasta no cambiar de medias. Cobweb, este método crea un cluster jerárquico con un árbol de clasificación. En el árbol de clasificación cada nodo es un concepto que tiene una descripción probabilística de ese concepto que resume los objetos clasificados bajo ese nodo. La descripción probabilística incluye la probabilidad del concepto $(P(C_i))$ y las probabilidades condicionales de pares atributos-valor dado por el concepto $(P(A_i = V_{ij}|C_k))$. Entre más grande es la proporción de elementos de las clase que tiene ese atributo-valor, ese atributo-valor es más predicativo sobre la clase. Cobweb también considera en cada iteración, unir los dos mejores nodos evaluados y dividir el mejor nodo evaluado. Cobweb depende del orden de los objetos. El método se puede extender a valores numéricos usando distribuciones gaussianas.

Al ocupar alguno de estos métodos, los resultado pueden ser guardados, para luego poder ver esos valores agrupados a través de la visualización, que es una herramienta grafica de representación de datos y de esa forma se puede determinar las tendencias de los datos de una forma visual.

Nos queda solamente la parte de reglas de asociación y que existe un solo método de asociación, y el esquema de asociación es el A Priori. El algoritmo APriori su objetivo es obtener itemsets (conjuntos de valores que se repiten) de un determinado tamaño, para combinarlos en reglas. Posee algunas ventajas como, que el algoritmo A Priori y sus variantes son los más usados dentro de este tipo de análisis. Su eficiencia para grandes volúmenes de datos es muy elevada y Ciertos SGBD son capaces de ejecutar este algoritmo dentro del núcleo del gestor. También este algoritmo presenta deficiencias como para ciertos datos de entrada, los resultados intermedios consumen gran cantidad de recursos

(memoria).

Apéndice E

Glosario

- **Apache:** es un software (libre) servidor HTTP de código abierto para plataformas Unix (BSD, GNU/Linux, etc.), Windows, Macintosh y otras, que implementa el protocolo HTTP/1.1 y la noción de sitio virtual.
- **ATM:** El Modo de Transferencia Asíncrona o Asynchronous Transfer Mode (ATM) es una tecnología de telecomunicación desarrollada para hacer frente a la gran demanda de capacidad de transmisión para servicios y aplicaciones.
- **Cookies:** Una cookie es un fragmento de información que se almacena en el disco duro del visitante de una página web a través de su navegador, a petición del servidor de la página. Esta información puede ser luego recuperada por el servidor en posteriores visitas. Las inventó Lou Montulli, un antiguo empleado de Netscape Communications. Al ser el protocolo HTTP incapaz de mantener información por sí mismo, para que se pueda conservar información entre una página vista y otra (como login de usuario, preferencias de colores, etc), ésta debe ser almacenada, ya sea en la URL de la página, en el propio servidor, o en una cookie en el ordenador del visitante.
- **CSV:** Los ficheros CSV (del inglés comma-separated values) son un tipo de documento sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal: España, Francia, Italia...) y las filas por saltos de línea. Los campos que contengan una coma, un salto de línea o una comilla doble deben ser encerrados entre comillas dobles. El formato CSV es muy sencillo y no indica un juego de caracteres concreto, ni cómo van situados los bytes, ni el formato para el salto de línea. Estos

puntos deben indicarse muchas veces al abrir el fichero, por ejemplo, con una hoja de cálculo.

- **Datawarehouse:** es una colección de datos orientadas a un dominio, integrado, no volátil y variable en el tiempo que ayuda a la toma de decisiones de la empresa u organización.

- **DNS:** El Domain Name System (DNS) es una base de datos distribuida y jerárquica que almacena información asociada a nombres de dominio en redes como Internet. Aunque como base de datos el DNS es capaz de asociar distintos tipos de información a cada nombre, los usos más comunes son la asignación de nombres de dominio a direcciones IP y la localización de los servidores de correo electrónico de cada dominio.

- **Gateway:** es un equipo que permite interconectar redes con protocolos y arquitecturas completamente diferentes a todos los niveles de comunicación. La traducción de las unidades de información reduce mucho la velocidad de transmisión a través de estos equipos.

- **GUI:** La interfaz gráfica de usuario (en inglés Graphical User Interface, GUI) es un tipo de interfaz de usuario que utiliza un conjunto de imágenes y objetos gráficos para representar la información y acciones disponibles en la interfaz. Habitualmente las acciones se realizan mediante manipulación directa para facilitar la interacción del usuario con la computadora. Surge como evolución de la línea de comandos de los primeros sistemas operativos y es pieza fundamental en un entorno gráfico. Como ejemplo de interfaz gráfica de usuario podemos citar el escritorio o desktop del sistema operativo Windows y el entorno X-Window de Linux.

- **Host:** El término host (equipo anfitrión) en informática o computación puede referirse a:

  - A una máquina conectada a una red de ordenadores y que tiene un nombre de equipo (en inglés, hostname). Es un nombre único que se le da a un dispositivo conectado a una red informática. Puede ser un ordenador, un servidor de archivos, un dispositivo de almacenamiento por red, una máquina de fax, impresora, etc. Este nombre ayuda al administrador de la red a identificar las máquinas sin tener que memorizar una dirección IP para cada una de ellas.
  - A veces también se llama así al dominio del equipo (Un dominio es la parte de una URL por la que se identifica al servidor en el que se aloja).

- También es el nombre de un fichero (fichero Hosts) que se encuentra en los ordenadores y resuelve algunos DNS.

- **Html:** es el acrónimo inglés de HyperText Markup Language, que se traduce al español como Lenguaje de Marcas Hipertextuales. Es un lenguaje de marcación diseñado para estructurar textos y presentarlos en forma de hipertexto, que es el formato estándar de las páginas web. Gracias a Internet y a los navegadores como Internet Explorer, Opera, Firefox, Netscape o Safari, el HTML se ha convertido en uno de los formatos más populares y fáciles de aprender que existen para la elaboración de documentos para web.

- **NSCA:** Existen cuatro formatos disponibles para los registros de transacciones. Los formatos de registro y los nombres predeterminados de archivos son los mismos que los utilizados por otros servicios de IIS. Piense sobre qué desea realizar el seguimiento en todos los servicios, cuántos archivos desea utilizar y sobre la manera de establecer el tamaño de los archivos antes de elegir un formato para el servicio SMTP.

- **Path Analysis:** análsis de camino de navegación.

- **Petabytes:** Un petabyte es una unidad de almacenamiento de información. Corresponde a 1200 terabytes, mil doscientos millones de megabytes, o mil doscientos billones de bytes. Se representa con el símbolo PB.

- **Proxy:** el término proxy hace referencia a un programa o dispositivo que realiza una acción en representación de otro. La finalidad más habitual es la del servidor proxy, que sirve para permitir el acceso a Internet a todos los equipos de una organización cuando sólo se puede disponer de un único equipo conectado, esto es, una única dirección IP.

- **Roaming Profile (Perfiles Móviles:** almacenarán los cambios en la configuración del perfil indicados por el usuario, tales como los archivos almacenados en ”Mis Documentos” o los iconos existentes en el ”Escritorio”, o sea los usuarios pueden acceder a sus ficheros y carpetas mientras trabajan, pudiendo conectarse desde distintos puestos.

- **Samba:** es una implementación libre del protocolo de archivos compartidos de Microsoft Windows (antiguamente llamado SMB, renombrado recientemente a CIFS) para sistemas de tipo UNIX. De esta forma, es posible que ordenadores con Linux o Mac OSX se vean como servidores o actúen como clientes en redes de Windows.

- **URL:** significa Uniform Resource Locator, es decir, localizador uniforme de recurso. Es una secuencia de caracteres, de acuerdo a un formato estándar, que se usa para nombrar recursos, como documentos e imágenes en Internet, por su localización.

- **XML:** sigla en inglés de eXtensible Markup Language («lenguaje de marcas extensible»), es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C). Es una simplificación y adaptación del SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML). Por lo tanto XML no es realmente un lenguaje en particular, sino una manera de definir lenguajes para diferentes necesidades. Algunos de estos lenguajes que usan XML para su definición son XHTML, SVG, MathML.

Bibliografía

[Abi97] Abiteboul S. (1997) Querying semi-structured data. *En Proceedings of the International Conference on Databases Theory (ICDT)* páginas 1–18.

[AFJM95] Armstrong R., Freitag D., Joachims T. y Mitchell T. (1995) Webwatcher: A learning apprentice for the world wide web. *AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Enviroments* IBM Almaden Research Center: 6–12.

[AGGR98] Agrawal R., Gehrke J., Gunopulos D. y Raghavan P. (1998) Automatic subspace clustering of high dimensional data for data mining application. *En Proceedings of the 1998 ACM SIGMOD International* 27 n°2: 94–105.

[AM04] Arotaritei D. y Mitra S. (2004) Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems* 148: 5–19.

[AMW04] Araya S., M.Silva y Weber R. (2004) A methodology for web usage mining and its application to target group identification. *Fuzzy Sets and Systems* 148: 139–152.

[And73] Anderberg M. (1973) Cluster analysis for applications. *Academic Press Inc New York* .

[AR94] Agrawal R. y R.Srikant (1994) Fast algorithms for mining association rules. *En Proceedings of the 20th International Conference on Very Large Data Base IBM Almaden Research Center* páginas 487–499.

[AR03] Abraham A. y Ramos V. (2003) Web usage mining using artificial ant colony clustering and genetic programming. *En: CEC03-Congress on evolutionary computation, IEEE Press* páginas 1384– 1391.

[Arl] Arlitt M.Characterizing web user sessions. *IEEE Network 14* 3: 30–37.

[Bac95] Backer E. (1995) Computer-assisted reasoning in cluster analysis. *Prentice Hall Internacional (UK) Ltd* Hertfordshire UK.

[Bal96] Baldwin F. (1996) Knowledge from data using fuzzy methods. *Pattern Recognition Lett* 17: 593–600.

[Bar02] Barbara D. (2002) Requirements for clustering data streams. *CM SIGKDD Explorations Newsletter* 3(2): 23–27.

[BBD+02] Babcock B., Babu S., Datar M., Motwani R. y Widom J. (June 2002) Models and issues in data stream systems, principles of database systems (pods'02). *En Proc. 2002 ACM Symp* 16: 1–16.

[BDH+94] Bowman M., Danzig B., Hardy D., Manher U. y Shwartz M. (1994) The harvest information discovery and access system. *En Proc.* 2^{nd} *Internacional World Wide Web Conference* páginas 763–771.

[Bez81] Bezdek J. C. (1981) Pattern recognition with fuzzy objective function algorithms. *Plenum, NY* .

[BFR+99] Burt M., Fowlkes C., Roden J., Stechert A. y Mukhtar S. (April 1999) Diamond eye: A distributed architecture for image data mining. *En SPIE DMKD, Orlando* páginas 71–80.

[BKM+00] Broder A., Kumar S., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. y Wiener J. (2000) Graph structure in the web. *Computer Networks* 33(1-6): 309–320.

[BLCGP] Berners-Lee T., Cailliau R., Groff J. y Pollermann B.World-wide web: The information universe. *Electronic Networking: Research, Applications and Policyr, volume = 1, number = 2, pages = 74–82, year = 1992,* .

[BR99] Baeza R. y Ribeiro B. (1999) Modern information retrieval. *ACM Press Addision-Wesley New York* .

[BW01] Babu S. y Widom J. (2001) Continuous queries over data streams. *En: SIGMOD Record'01l Conference on Very Large Data Base* página 109–120.

[CDF+98] Craven M., DiPasquo D., Fretag D., McCallum A., Mitchell T., Nigam K. y Slattery S. (1998) Learning to extract symbolic knowledge from the world wide web. *En Proceedings of the Fifteenth National Conference on Artificial Intelligende (AAAI98)* páginas 509–516.

[CDF+00] Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K. y Slattery S. (2000) Learning to construct knowledge bases from the world wide web. *Artificial Intelligence* 118(1-2): 69–114.

[CDH+02] Chen Y., Dong G., Han J., Wah B. y Wang J. (2002) Multidimensional regression analysis of time-series data streams. *En: 2002 International Conference on Very Large Data Bases* páginas 323–334.

[CDI98] Chakrabarti S., Dom B. y Indyk P. (1998) Enhanced hypertext categorization using hyperlinks. *En Proceedings of the ACM SIGMOD International Conference on Management on Data, ACM Press Seattle, WA* página 307–318.

[CH05] Cohen A. y Hersh W. (2005) A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 6(1): 55–57.

[CHMW01] Chang G., Healy M., McHugh J. y Wang J. (2001) Mining the world wide web: An information search approach. *Kluwer Academic Publishers* páginas 100–104.

[CM04] Cernuzzi L. y Molas M. (Septiembre 2004) Integrando diferentes tecnicas de data mining en procesos de web usage minig. *30th Conferencia Latinoamericana de Informatica (CLEI2004)* 3: 140–149.

[CMS97] Cooley R., Mobasher B. y Srivastava J. (1997) Grouping web page references into transactions for mining world wide web browsing patterns. *Knowledge and information Systems* páginas 2–11.

[CMS99] Cooley R., Mobasher B. y Srivastava J. (1999) Data preparation for mining world wide web browsing patterns. *Knowledge and information Systems* 1(1): 5–32.

[Coh99] Cohen W. (1999) What can we learn from the web? *En Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)* páginas 515–524.

[CP95] Catledge L. y Pitkow J. (1995) Characterizing browsing behaviors on the world wide web. *En: Computer networks and ISDN systems* 27(6): 1.065–1.073.

[CPY96] Chen M., Park J. y Yu P. (1996) Data mining for path traversal patterns in a web environment. *En: Proc. 16th International conference on distributed computing systems* páginas 385–392.

[Cro95] Croft W. (1995) What do people want from information retrieval? *D-Lib Magazine* páginas 22–44.

[CS96] CHeeseman P. y Stutz J. (1996) Bayesian classification (autoclass): Theory and results. *En Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy , R. (Eds.) Advances in Knowledge Discovery and Data Mining* AAAI Press/MIT Press: 5–32.

[CSN98] Craven M., Slattery S. y Nigam K. (1998) First-order learning for web mining. *C. Nedellec and C. Rouveirol editors Proceedings of the 10th European Conference on Machine Learning (ECML-98)* 1, Springer-Verlag,Chemnitz Germany(1): 250–255.

[CTS99] Cooley R., Tan P. y Srivastava J. (August 1999) Websift: The web site information filter system. *Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99).En Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag* páginas 385–392.

[CVS04] Cerda L., Vila M. y Sánchez D. (2004) Reglas de asociacion aplicadas a la deteccion de fraude con tarjetas de credito. *XXII Congreso Español sobre Tecnologias y Logica Difusa* páginas 1–6.

[Dan02] Dandretta G. (Junio, 2002) Web mining: Implementando técnicas de data mining en un servidor web. *Technical Report, Universidad de Belgrano, Buenos Aires* .

[DEW96] Doorenbos R., Etzioni O. y Wield D. (1996) A scalable comparison shopping agent for the world wide web. *IEEE Technique Report, University od Washintong , Dept. Of Computer Science and Engineering* páginas 39–48.

[DGR03] DaCruz R., García F. y Romero L. (Enero, 2003) Perfiles de usuario: en la senda de la personalización. *Informe Técnico, Technical Report DPTOIA-IT-2003-001* .

[DGSV96] Delgado M., Gómez-Skarmeta A. y Vila M. A. (1996) On the use of the hierarchical clustering in fuzzy modeling. *International Journal of Approximate Reasoning* 14: 237–257.

[DHS00] Duda R., Hart P. y Stork D. (2000) Pattern classification. *John Wiley & Sons, Nueva York, Estados Unidos* .

[DKS95] Dougherty J., Kohavi R. y Sahami M. (1995) Supervised and unsupervised discretization of continuos features. *Proc. of the 12th International Conference Machine Learning* páginas 194–202.

[DL01] Dzeroski S. y Lavrac N. (2001) Relational data mining: Inductive logic programming for knowledge discovery in databases. *Springer-Verlag* .

[DO74] Duran B. y Odell P. (1974) Cluster analysis: A survey. *Springer-Verlag, New York* páginas 194–202.

[DOF03] DeCaceres M., Oliva F. y Font X. (2003) Ginkgo, un programa de análisis multivariante orientado a la clasificación basada en distancias. *27 Congreso Nacional de Estadística e Investigación Operativa, Lleida* páginas 1–9.

[DSV00] Delgado M., Sanchez D. y Villa M. A. (2000) Fuzzy cardinality based evaluation of quantified sentences. *Int. J. Aproxx.Reaason* 3: 23–66.

[DSVM03] Delgado M., Sanchez D., Vila M. A. y Marín N. (2003) Fuzzy association rules: General model and application. *IEEE Transactions on Fuzzy Systems* 11(2): 214–225.

[EJMBSV06a] Escobar-Jeria V., Martin-Bautista M., Sanchez D. y Vila M. A. (2006) Minería web: Aplicaciones con lógica difusa. *3th Congreso Español sobre Tecnología y Lógica Fuzzy* páginas 235–240.

[EJMBSV06b] Escobar-Jeria V., Martin-Bautista M., Sanchez D. y Vila M. A. (2006) Web mining: Aplications with fuzzy logic. *Primera Conferencia Internacional sobre Ciencias Y Tecnologias Multidiciplinarias de la Informacion (InScit)* páginas 77–81.

[Etz96] Etzioni O. (1996.) The world wide web: Quagmire or gold mine. *Communications of the ACM* 39(11): 65–68.

[EV03] Eirinaki M. y Vazirgianis M. (2003) Web mining for web personalization. *ACM Transactions on Internet Tehnology (TOIT)* 3(1): 1–27.

[Fa00] Fukuda M. y Ñanri I. (2000) Mining from literary texts: Pattern discovery and similarity computation. *Progress in Discovery Science 2000* páginas 518–531.

[Fis87] Fisher D. (1987) Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2: 139–172.

[FLG00] Flake G., Lawrence S. y Giles C. L. (August 2000) Efficient of web communities. *En Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts* páginas 150–160.

[FLM98] Florescu D., Levy A. y Mendelzon A. (1998) Database techniques for the world-wide web: A survey. *SIGMOD Record* 27(3): 59–74.

[FPSM92] Frawley W., Piatetsky-Shapiro G. y Matheus C. (1992) Knowledge discovery in databases: An overview. *SAAAI/MIT Press* páginas 57–70.

[FPSSU96] Fayyad U., Piatetsky-Shapiro G., Smyth P. y Uthurusamy P. (1996) Advances in knowledge discovery and data mining. *AAAI/MIT Press* .

[FR99] Fraley C. y Raftery A. (1999) Mclust: Software for model-based cluster and discriminant analysis. *Tech Report 342 Dept. Statistics Univ. of Washington* 16: 297–306.

[GCS+05] Garre M., Cuadrado J., Sicilia M., Charro M. y Rodríguez D. (2005) Comparacion de diferentes algoritmos de clustering en la estimacion de coste en el desarrollo de software. *the ISBSG database, Information Technology Interfaces, Croacia* ISBN: 84-607-5827-3: 20–23.

[GCS07] Garre M., Cuadrado J. y Sicilia M. (2007) Comparacion de diferentes algoritmos de clustering en la estimacion de coste en el desarrollo de software. *Revista Española de Innovación, Calidad e Ingeniería del Software* 3: 20–27.

[GG02] García F. y Gil A. (2002) Personalización y recomendación en aplicaciones de comercio electrónico. *En Avances en Comercio Electrónico, F. García Peñalvo (Ed.)* ISBN: 84-607-5827-3: 137–148.

[GGR02] Garofalakis M., Gehrke J. y Rastogi R. (June 2002) Querying and mining data streams: You only get one look (a tutorial). *En Proc. 2002 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'02)* ISBN: 84-607-5827-3: 635–645.

[GK00] Gedeon T. y Koczy L. (2000) A model of intelligent information retrieval using fuzzy tolerance relations based on hierarchical co-occurrence of words. *Lecture Notes in Artificial Intelligence* 50: 48–74.

[GM98] Gibbons P. y Matias Y. (June 1998) New sampling-based summary statistics for improving approximate query answers. *En Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)* 2393: 331–342.

[GMLL02] Gelbukh A., Montes M. y López-López A. (2002) Text mining at dateil level using conceptual graphs. *En Soft Computing in Information Retrieval: Techniques and Applications, Germany: Physica-Verlag* 2393.

[GMMO00] Guha S., Mishra N., Motwani R. y O'Callaghan L. (2000) Clustering data streams. *IEEE Symposium on Foundations of Computer Science (FOCS'00)* página 359.

[GO03] Golab L. y Ozsu M. (2003) Issues in data stream management. *En SIGMOD Record* 32(2): 5–14.

[GRS97] Guha S., Rastogi R. y Shim K. (1997) Cure: A clustering algorithm for large databases. *Technical Report, Bell Laboratories, Murray Hill* páginas 73–84.

[Gye00] Gyenesei A. (2000) A fuzzy approach for mining quantitative association rules. *Univ. Turku, Dept. Comput. Sci., Lemminkisenkatu 14, Finland, TUCS Tech. Rep. 336* Bell Laboratories, Murray Hill.

[Hag99] Hagen P. (1999) Smart personalization. *Forrester Report* .

[Har75] Hartigan J. (1975) Clustering algorithms. *John Wiley & Sons* .

[HBML95] Hammond K., Burke R., Martin C. y Lytinen S. (1995) Faq-finder: a case based approach to knowledge navigation. *En Working notes of the AAAI Spring Symposium: Information Gathering from Heteregeneous Distributed Environmets AAAi Pres* páginas 69–73.

[HCC93] Han J., Cai Y. y Cercone N. (1993) Data-driven discovery of quantitative rules in relational databases. *En IEEE Transactions on Knowledge and Data Eng.* 5: 29–40.

[HG00] He D. y Goker A. (2000) Detecting session boundaries fromweb user logs. *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research, UK: British Computer Society Cambridge* página 57–66.

[HH03] Hilderman R. y Hamilton H. (2003) Measuring the interestingness of discovered knowledge: A principal approach. *Intelligent Data Analysis* 7(4): 347–382.

[HLT99] Hussain F., Liu H. y Tan C. (1999) Discretization : an enabling technique. *Technical Report TRC6/99, The National University of Singapore* páginas 1022–1027.

[HPS04] Huang X., Peng F. y Schuurmans D. (2004) Dynamic web log session identification with statistical language models. *En: Journal of American society for information science and technology* 55(13): 1.290–1.303.

[HSCL01] Haruechaiyasak C., Shyu M., Chen S. y Li X. (2001) Web document classification based on fuzzy association. *Proceedings of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment* páginas 487 – 492.

[JD98] Jain A. y Dubes R. (1998) Algorithms for clustering data. *Prentice-Hall advanced references The National University of Singapore series Inc., Upper Saddle River* .

[Jim07] Jimenez M. R. (2007) Algunos aspectos formales en representación y evaluación de reglas de asociación. *Reporte Interno del Depto. Ciencias de la Computacion e Inteligencia Artificial, Universidad de Granada.* .

[Jus04] Justicia C. (Granada, 2004) Formas intermedias de representacion en mineria de texto. *Memoria para el Diploma de Estudios Avanzados, Universidad de Granada* .

[Kan] Kandel A.Fuzzy techniques in pattern recognition. *New York: John Wiley & Sons* página 356.

[Kar01] Kargupta H. (2001) Career: Ubiquitous distributed knowledge discovery from heterogeneous data. *NSF Information and Data Management (IDM) Workshop* páginas 18–23.

[Kar03] Kargupta H. (2003) Vehicle data stream mining (vedas proyect). *NSF Information and Data Management (IDM) Workshop* páginas 37–46.

[KB00] Kosala R. y Blockeel H. (June 2000) Web mining research: A survey. *ACM SIG KDD Explorations Newsletter of the SCMA Special Interest Group on Knowledge Discovery and Data Mining* 2(1): 1–15.

[KC01] Kim K. y Cho S. (2001) Personalized mining of web documents using link. *IFSA World Congress and 20th NAFIPS International Conference* 1: 81–86.

[KC03] Kim K. y Cho S. (2003) Fuzzy integration of structure adapttive soms for web content mining. *Fuzzy Sets and System* 148: 43–60.

[KHK99] Karypis G., Han E. y Kumar V. (1999) Chameleon: A hierarchical clustering algorithm using dynamic modeling. *Journal Computer* 32(8): 68–75.

[Kim96] Kimball R. (1996) The data warehouse toolkit. *John Wiley & Sons, Inc. Publication:1996 New York* .

[KJNY01] Krishnapuram R., Joshi A., Nasraoui O. y Yi L. (August 2001) Low-complexity fuzzy relational clustering algorithms for web mining. *Journal IEEE-FS* 9: 595–607.

[KKR99] Kleinberg J., Kumar R. y Raghavan P. (1999) The web as a graph: measurements, models, and methods. *Lecture Notes in Computer Science* (1627): 1–17.

[Kle99] Kleinberg J. (1999) Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5): 604–632.

[KLR+98] Kennedy R., Lee Y., Roy B., Reed C. y Lippman R. (1998) Solving data mining problems through pattern recognition. *Prentice Hall, Upper Saddler River, New Jersey* .

[KLSS95] Kirk T., Levy A., Sayiv Y. y Srivastava D. (1995) The information manifold. *En working notes of the AAAI Sping Simposyum: Infomation Gathering from Heterogeneous, Distributed Enviroments.AAAI Press* páginas 85–91.

[KR87] Kaufman. L. y Rousseeuw P. (1987) Clustering by means of medoids. *En Dodge,Y. (ed.), Statistical Data Analysis Based on the L1-norm and Related Methods* North Holland, Amsterdam: 405–416.

[KR90] Kaufman L. y Rousseeuw P. (1990) Finding groups in data: An introduction to cluster analysis. *John Wiley & Sons, New York* .

[KRRT99] Kumar S., Raghavan P., Rajagopalan S. y Tomkins A. (1999) Trawling emerging cyber-communities automatically. *Proc 8th World Wide Web Conference, newsletter of the SCMA Special Interest Group on Knowledge Discovery and Data Mining* páginas 1481–1493.

[KS95] Konopnicki M. y Shmueli O. (1995) W3qs: A query systems for the world wide web. *En Proc. of the 21th VLDB Conference* páginas 54–65.

[KW96] Kwork C. y Weld D. (1996) Planning to gather information. *En Proc. 14th National Conference on AI* 9: 32–39.

[LHCM00] Liu B., Hsu W., Chen S. y Ma Y. (2000) Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems* 15(5): 47–55.

[LHML99] Liu B., Hsu W., Mun L. y Lee H. (1999) Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engeneering* 11(6): 817–832.

[LW67] Lance G. y Williams W. (1967) A general theory of classificatory sorting strategies clustering systems. *Computer Journal* 10: 271–277.

[Mae94] Maes P. (1994) Agents that reduce work and information overload. *Communications of the ACM* 37(7): 30–40.

[MAFM99] Menascé D., Almeida V., Fonseca R. y Mendes M. (November 1999) A methodology for workload characterization of e commerce sites. *Proceedings of ACM Conference on Electronic Commerce (EC-99)Denver, CO* páginas 119–128.

[MAM97] Merilado P., Atzeni P. y Mecca G. (1997) Semistructured and structured dara in the web: Going back and forth. *En Proccedings of the Workshop on the Management of Semistructured Data (in conjunction with ACM SIGMOD* .

[MBKV+02] Martín-Bautista M., Kraft D., Vila M. A., Chen J. y Cruz J. (August 2002) User profiles and fuzzy logic for web retrieval issues. *Springer Berlin-Heidelberg, ISSN: 1432-7643 (Paper) 1433-7479 (Online)* 6(5): 365 – 372.

[McG67] McGarry J. (1967) Some methods for classification and analysis of multivariate observations. *Fifth Berkeley Symposium on Mathematical Statistics and Probability 1* páginas 281–297.

[McG05] McGarry K. (2005) A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review* 20(1): 39–61.

[McQ67] McQueen J. (1967) Some methods for classification and analisys of multivariate observations. *En Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability* páginas 281– 287.

[MCS99] Mobasher B., Cooley R. y Srivastava J. (1999) Creating adaptive web sites through usage-based clustering of urls. *En Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)* páginas 19–25.

[MCS00] Mobasher B., Cooley R. y Srivastava J. (2000) Automatic personalization based on web usage mining. *Communcation of the ACM* 43(8): 142–151.

[MH02] Mack R. y Hehenberger M. (2002) Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discovery Today* 7(11 (Supl.)): S89–S98.

[Mit97] Mitchell T. (New York, 1997) Machine learning. *Book McGraw-Hill* .

[MJ96] Mao J. y Jain A. (1996) A self-organizing network for hyperellipsoidal clustering (hec). *EEE Trans. Neural Netw* 7: 16–29.

[MJF99] Murty M., Jain A. y Flynn P. (1999) Data clustering: A review. *ACM Computing Survey* 31: 264–323.

[MKH97] Mobasher B., Kumar V. y Han E. (1997) Clustering in a high dimensional space using hypergraph models. *Minneapolis: Univ. Minnesota, Tech. Rep. TR-97-063* .

[Mob05] Mobasher B. (2005) Web usage mining and personalization. *Capítulo 4 de Practical Handbook of Internet Computing M.P.Singh ed. CRC Press LLC* .

[Mol02] Molina L. (2002) Data mining: torturando a los datos hasta que confiesen. *http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html* .

[MPM02] Mitra S., Pal S. y Mitra P. (January 2002) Data mining in soft computing framework: a survey. *IEEE Transactions on Neural Networks* 13(1): 3–14.

[MPR00] Manber U., Patel A. y Robison J. (2000) Experience with personalization on yahoo!! *Communications of the ACM* 43: 35–39.

[MS96] Maarrek Y. y Shaul L. (1996) Automatically organizing bookmarks per contents. *En Proc. of 5th International World Wide Web Conference* páginas 1321–1333.

[MT96] Manila H. y Toivonen H. (1996) Discovering generalized episodes using minimal occurences. *En: Proc. Second international conference on knowledge discovery and data mining* páginas 146–151.

[Mth03] Mthukrishnan S. (Jan. 2003) Data streams: algorithms and applications. *En Proc. 2003 Annual ACM-SIAM Symp, Discrete Algorithms (SODA'03)* Baltimore, MD: 413–413.

[MTV94] Mannila H., Toivonen H. y Verkamo A. I. (July 1994) Efficient algorithms for discovering association rules. *En Proceedings of the AAAI Workshop on Knowledge in Databases* Seattle, Washington: 181–192.

[NCRG03] Nasraoui O., Cardona C., Rojas C. y González F. (August 2003) Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm. *En Proc of webKDD 2003-KDD workshop on web mining as a premise to effective and Intelligent web application, Washintong DC* páginas 40–47.

[NFJK99] Nasraoui O., Frigui H., Joshi A. y Krishnappuram R. (1999) Mining web accsess logs using relational competitive fuzzy clustering. *En Proceedings of the Eight International Fuzzy Systems Association World Congress* páginas 531–547.

[NK00] Nasraoui O. y Krishnapurum R. (may 2000) A novel approach to unsupervised robust clustering using genetic niching. *Proc. Of the 9th IEEE International Conf. on Fuzzy Systems, San Antonio, TX* páginas 170–175.

[NK02] Nasraoui O. y Krishnapurum R. (Sep. 2002) A new evolutionary approach to web usage and context sensitive associations mining. *International Journal on Computational Intelligence and Applications – Special Issue on Internet Intelligence Systems* 2(3): 339–348.

[NKJF00] Nasraoui O., Krishnapurum R., Joshi A. y Frigui H. (2000) Extraction web user profiles using relational competitive fuzzy clustering. *International Journal on Artificial Intelligence Tools* 9(4): 509–526.

[OML00] Oh H., Myaeng S. y Lee M. (Greece, 2000) A practical hypertext categorization method using links and incrementally available class information. *En Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval (SIGIR-00)* 2(3): 264–271.

[OPW94] Oostendorp K., Punch W. y Wiggins R. (1994) A tool for individualizing the web. *En Proc 2nd Internacional World Wide Web Conference* páginas 49–57.

[Orw95] Orwant J. (1995) Heterogeneous learning in the doppelganger user model system. *User Modeling and User Adapted Interaction* 4(2): 107–130.

[PB95] Pal N. y Bezdek J. (1995) On cluster validity for the fuzzy c-means model. *IEEE transactions on Fuzzy System* 3(3): 370–379.

[PB02] Paralic J. y Bednar P. (2002) Knowlelge discovery in texts supporting e-democracy. *En 6th IEEE International Conference o Intelligent Engineering Systems,INES 2002* páginas 327–332.

[PBMW99] Page L., Brin S., Motwani R. y Winograd T. (1999) The pagerank citation ranking: Bringing order to the web. *Stanford Digital Library working paper SIDL-WP-1999-0120* .

[PE95] Perkowitz M. y Etzioni O. (1995) Category traslation: learning to understand information on the internet. *En Proc. 15th International Joint Conference on AI, Montreal, Canada* páginas 930–936.

[PMB96] Pazzani M., Muramatsu J. y Billsus D. (1996) Syskill & webert: Identifying interesting websites. *En Proc AAAI Spring Simposyum on Machine Learning in Information Access, Portland* páginas 54–61.

[PPPS03] Pierrakos D., Paliouras G., Papalheodorou C. y Spyropoulos C. (2003) Web usage mining as a tool for personalization:a survey. *Kluwer Academic Publishers, Hingham, MA, USA*, 13(4): 311–372.

[PTM02] Pal S., Talwar V. y Mitra P. (September 2002) Web mining in soft computing framework: relevance, state of the art and future directions. *IEEE Transactions on Neural Networks* 13: 1163– 1177.

[Pyl99] Pyle D. (1999) Data preparation for data mining. *Book Morgan Kaufmann, San Francisco California* .

[RB03] Runklerm T. y Bezdek J. (2003) Web mining with relational clustering. *International Journal of Approximate Reasoning* 32: 217–236.

[RIS+94] Resnik P., Iacovou N., Sushak M., Bergestrom P. y Riedl J. (1994) Grouplens: an open architecture for collaborative filtering of netnews. *En Proc. of the 1994 Computer Supported Cooperative Work Conference* 40(3): 77–87.

[RJ02] Raymond T. y Jiawei H. (2002) Clarans: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering* 14(5): 1003–1016.

[RM92] Richards B. y Mooney R. (1992) Learning relations by path finding. *En proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92), AAAI Press* página 50–55.

[Sah99] Sahar S. (1999) Interestingness via what is not interesting. *En Proceedings of the 5th International Conference on Knowledge and Data Mining, San Diego, CA* páginas 332–336.

[Sah01] Sahar S. (2001) Interestingness preprocessing. *En Proceedings of the 2001, IEEE International Conference on Data Mining, San Jose, CA* páginas 489–496.

[Sah02] Sahar S. (Japan, 2002) On incorporating subjective interestingness into the mining process. *En Proceedings of the 2002, IEEE International Conference on Data Mining, Maebashi City* páginas 681–684.

[Sal91] Salton G. (1991) Developments in automatic text retrieval. *Science* 253: 974–980.

[San99] Sanchez D. (1999) Adquisición de relaciones entre atributos en base de datos relacionales.

[SB75] Shortlife E. y Buchaman B. (1975) A model of inexact reasoning in medicine. *Mathematical Biosciences* 23: 351–379.

[SCDT00] Srivastava J., Cooley R., Deshpande M. y Tan P. (2000) Web usage mining: Discovery and applications of usage pattern from web data. *SIGKDD Explotarions* 1(2): 1–12.

[Ser03] Serrano J. (2003) Knowledge fusion in relational databases: Aggregation and summarization measures. *Proceedings of the 1st ICEIS Doctoral Consortium (DCEIS-2003) in conjunction with ICEIS 2003, Angers, France* páginas 1–4.

[SH01] Subasic P. y Huettner A. (Aug 2001) Affect analysis of text using fuzzy semantic typing fuzzy systems. *IEEE Transactions* 9: 483–496.

[Sha76] Shafer G. (1976) A mathematical theory of evidence. *Princeton University Press* .

[SM95] Shardannad U. y Maes P. (1995) Social information filtering: Algorithms for automating word of mouth. *En Proc. of 1995 Confrence on Human Factors in Computing Systems* páginas 210–217.

[SMBN03] Spiliopoulou M., Mobasher B., Berendt B. y Nakagawa M. (2003) A framework for the evaluation of session reconstruction heuristics in web usage analysis. *En: Informs journal on computing* 15(2): 171–190.

[Spa80] Spath H. (1980) Cluster analysis algorithms for data reduction and classification. *Ellis Horwood, Upper Saddle River, NJ.* .

[Spe97] Spertus E. (1997) Parasite: mining structure information on the web. *En Proc. of 6th international World Wide Web Conference* páginas 1205–1215.

[SRK98] Sudipto G., Rajeev R. y Kyuseok S. (1998) Cure: an efficient clustering algorithm for large databases. *En SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International conference on Management of data* páginas 73–84.

[SRK00] Sudipto G., Rajeev R. y Kyuseok S. (2000) Rock: A robust clustering algorithm for categorical attributes. *Information Systems* 25(5): 345–366.

[SS03] Srivastava A. y Stroeve J. (2003) Onboard detection of snow, ice, clouds and other geophysical processes using kernel methods. *Proceedings of the ICML'03 workshop on Machine Learning Technologies for Autonomous Space Applications* .

[SSTK07] Shacham R., Schulzrinne H., Thakolsri S. y Kellerer W. (2007) Ubiquitous device personalization and use: The next generation of ip multimedia communications. *ACM Trans. Multimedia Comput. Commun* 3(12): 1–20.

[Swa94] Swanson D. (1994) Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *En : Neuroscinse Research Communications* (15): 1–9.

[Tan99] Tan A. (1999) Text mining: Promises and challenges. *En Pacic Asia Conference on Knowledge Discovery and Data Mining PAKDD'99 workshop on knowledge Discovery from Advanced Databases* páginas 63–70.

[TCZ+03] Tatbul N., Cetintemel U., Zdonik S., Cherniack M. y Stonebraker M. (2003) Load shedding in a data stream manager. *En Proceedings of VLDB, Berlin, Germany* .

[TK00] Tan P. y Kumar V. (2000) Interestingness measures for association patterns: A perspective. *En Technical Report TR00-036 (KDD 2000 Workshop on Postprocesing in Machine Learnig And Data Mining)* páginas 63–70.

[TV84] Trillas P. y Valverde L. (1984) On implication and indistinguishability in the setting of fuzzy logic. *En R.R. Yager and J. Kacprzyk, editors, Managament Decisión Support Systems Using Fuzzy Set and Possibility Theory, Verlag TUV Rheinland* páginas 198–212.

[TZL96] T. Zhang R. R. y Livny M. (1996) Birch: an efficient data clustering method for very large databases. *En SIGMOD '96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data* páginas 103–114.

[Val06] Valenzuela R. (2006) Aplicaciones del soft computing al análisis de ficheros log de sitios web. *Tesis Doctoral Ed. Universidad de Granada ISBN 84-338-4166-1* páginas 1–184.

[Van79] VanRjsbergen C. (1979) Information retrieval. *Butterworths, London Second Edition* 7(1): 2–10.

[VDS00] Vila M. A., Delgado M. y Sanchez D. (2000) Acquisition of fuzzy association rules from decimal data. *En: Barro S, Marín R, editors. Fuzzy logic in medicine, Physical-Verlag* .

[VN02] Viglas S. y Naughton J. (2002) Rate based query optimization for streaming information sources. *En Proc. of SIGMOD* páginas 37–48.

[VP07] Velásquez J. y Palade V. (2007) A knowledge base for the maintenance of knowledge extracted from web data. *Knowledge-Based Systems* 20(3): 238–248.

[WCF+00] Wong P., Cowley W., Foote H., Jurrus E. y Thomas J. (2000) Visualizing sequential patterns for text mining. *Proc. IEEE Information Visualization* páginas 43–50.

[WD94] Wallace C. y Dowe D. (1994) Intrinsic classification by mml. the snob program. *En the Proceedings of the 7th Australian Joint Conference on Artificial Intelligence, UNE, World Scientific Publishing Co., Armidale* páginas 37–44.

[WSP01] Wong C., Shiu S. y Pal S. (2001) Mining fuzzy association rules for web access case adaptation. *En Workshop Proceedings of Soft Computing in Case-Based Reasoning Workshop, in conjunction with the 4th International Conference in Case-Based Reasoning, Vancouver* páginas 213–220.

[WVS+96] Weiss R., Velez B., Sheldon M., Namprempre C., P.Szilagvi, Duda A. y Gifford D. (1996) Hypursuit: a hierarchical network search engine that exploits content-link hypertext clustering. *EN Hypertext'96: The 7th ACM Conference on hypertext* páginas 180–193.

[XKPS02] Xu F., Kurz D., Piskorski J. y Shmeier S. (2002) Term extraction and mining of term relations from unrestricted texts in the financial domain. *En Proceedings of BIS 2002* .

[XMZ05] Xin J., Mobasher B. y Zhou Y. (2005) A web recommendation system based on maximum entropy. *ITCC* 1: 213–218.

[Yag96] Yager R. (1996) Database discovery using fuzzy sets. *Int. J. Intell. Syst.* 11: 691–712.

[Yag00] Yager R. (2000) A framework for linguistic and hierarchical queries for document retrieval. *Soft Computing in Information Retrieval: Techniques and Applications.* páginas 172–180.

[YSG02] Yang Y., Slattery S. y Ghani R. (2002) A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems,Special Issue on Automatic Text Categorization* 18(2-3): 219–241.

[Zad75] Zadeh L. (1975) The concept of linguistic variable and its application to approximate reasoning. *Information Sciences* 8: 199–251.