Preprocesamiento de bases de datos masivas y multi-dimensionales en minería de uso web para modelar usuarios: comparación de herramientas y técnicas con un caso de estudio

Gustavo González Sánchez, Sonia Delfín Ávila, Josep Lluís de la Rosa

Institut d'Informàtica i Aplicacions

Agents Research Lab

Universitat de Girona

Campus Montilivi, Edifici P4

E-17071, Girona. España.

{gustavog, slizzeth, peplluis}@eia.udg.es

Resumen

El modelado de usuarios en Sistemas de Recomendación en Internet a través de sus acciones de uso del sitio web contribuye a aliviar el problema de la sobrecarga de información al usuario. Sin embargo, la extracción de los datos útiles para construir modelos de usuarios en la Minería de Uso Web (WUM) constituye sin lugar a dudas la fase más compleja y costosa en términos de tiempo y recursos computacionales. Conocida como fase de preprocesamiento de datos, esta etapa de selección, limpieza, mejoramiento, reducción y transformación de las bases de datos masivas y multi-dimensionales y de los logs requiere de una combinación sinérgica de la experiencia del analista en la aplicación de técnicas de aprendizaje automático, aplicación de algoritmos de minería de datos y el uso de herramientas específicas para obtener datos fiables. Analizamos con un caso real esta fase de preprocesamiento en la Minería de Uso Web y desarrollamos una comparación entre las herramientas más usadas para este propósito creando una taxonomía de caracteríticas y los algoritmos usados en cada una. El caso de estudio tiene 2.161.159 usuarios que pueden ser modelados a partir de un máximo de 984 atributos y los logs son del orden de 50 Gb.

1. Introducción

Como resultado del tremendo crecimiento de la World Wide Web (WWW), los datos brutos de la Web, se han convertido en una vasta fuente de información. Por consiguiente, el uso de las técnicas de data mining en estos datos se ha hecho necesario para descubrir información oculta. La Minería Web consiste en aplicar las técnicas de minería de datos a documentos y servicios de la Web. En particular, la creación, extraccción y mantenimiento de los modelos de usuario en Sistemas de Recomedación en Internet mejora la experiencia del usuario en relación con la información que es relevante para él o ella reduciendo el problema conocido como sobrecarga de información.

Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (Log). Las herramientas de Web mining analizan y procesan estos logs para producir información significativa. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, meta datos o hipervínculos, investigaciones recientes usan el término multimedia data mining (minería de datos multimedia) como una instancia de la Minería de uso Web, para tratar ese tipo de datos [13]. Los accesos totales por dominio, ho-

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.193-202
ISBN: 84-9732-449-8

rarios de accesos más frecuentes y visitas por día, entre otros datos, son registrados por herramientas estadísticas que complementan todo el proceso de análisis del Web mining.
La Minería Web puede ser clasificada, dependiendo que parte de la Web se esté explotando, como: Minería del contenido (Content Mining), Minería de Estructura (Structure Mining) o Minería de Uso (Usage Mining)[5].

2. Minería de uso web

El análisis de los hábitos de uso, de los visitantes de la Web, puede dar pistas importantes sobre tendencias de mercado y organizaciones actuales de la ayuda de predecir las tendencias futuras de clientes potenciales. Analizar las largas trayectorias de visitas (camino de visitas) pueden indicar la necesidad de reestructurar el sitio Web, para ayudar a los visitantes a encontrar la información rápidamente. También, se puede utilizar para ofrecer el contenido preferido de los visitantes. Mucho del trabajo de investigación conduce a la personalización Web. Sitios adaptativos, usan la información acerca de los caminos que utilizaron los usuarios para acceder al sitio y así mejorar su organización y presentación para cada uno de los diferentes usuarios. Algunas de las técnicas que se han utilizado son las siguientes:

- Capturar los perfiles comunes de los visitantes con una regla de asociación de descubrimiento y un clustering basado en el uso de las URLs.
- Descubrir por medio de los datos que se utilizaron una regla de asociación.
- Descubrimiento y comparación de los patrones de la navegación de clientes y no clientes para determinar la calidad de un sitio web.
- Usar frecuencias de acceso a la página y la clasificación de páginas.
- Desarrollar una herramienta para modificar el sitio web, y adaptarlo a requisitos particulares dinámicamente.

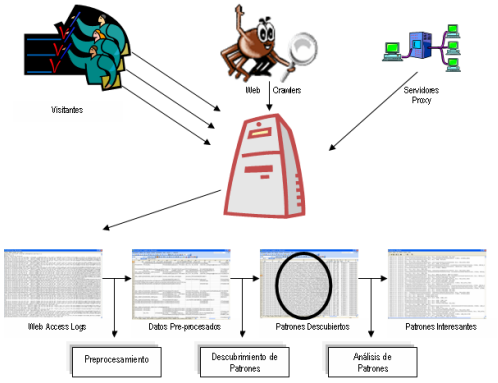

Figura 1: El Proceso de la Minería de Uso Web.

- Descubrir patrones del comportamiento interesantes de los usuarios móviles del dispositivo.
- Encontrar las localizaciones de la página que son diferentes a lo que los visitantes esperan que sea, mientras visitan la web, también ayuda a reestructurar la organización del sitio web.

El proceso completo de la Minería de Uso Web se muestra en la figura 1.

2.1. Minería de uso web y logs de acceso

El *Registro de Accesso Web* (en adelante Log) es el recurso más importante para la explotación de la Minería de uso Web porque almacena los datos que pertenecen a los accesos del sitio web. Los datos se pueden almacenar en un *Formato Común de Registro*, (CLF), o *Formato Extendido de Registro* (ELF). El registro guardado en CLF, contiene información de la dirección IP del visitante, la identificación de usuario (UserID), y la fecha y hora de la página que solicitó. El método significa la página que fue solicitada y puede ser determinado por los siguientes parámetros: *GET*, *PUT*, *POST* o *HEAD*. La URL es la pagina que fue solicitada y el protocolo significa como se estableció la comunicación. El estatus es el código de terminación, por ejemplo 200 es el código para determinar que fue satisfactorio.

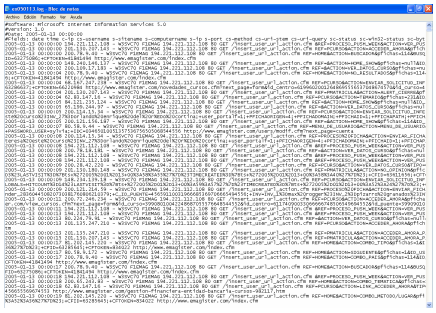

Figura 2: Datos registrados en un web log.

El tamaño de la celda muestra los bytes transferidos como resultado del acceso a la página. El formato de registro extendido, además de esta información almacena el *REFERRER* que muestra desde qué página estamos accediendo y el agente es el navegador web que utilizamos (ver figura 2).

Los patrones de descubrimiento implican la aplicación de varias técnicas y algoritmos para preprocesar datos. Con estos algoritmos se busca la identificación de patrones útiles fuera de los patrones descubiertos.

3. Preprocesamiento de logs en minería de uso web

El procesamiento que se hace en el Web Usage Mining, probablemente sea la fase mas difícil en todo lo que es el Web Mining, por la complejidad y la cantidad de tiempo que requiere. Esta fase esta integrada por las tareas de remover los datos que no necesitamos, identificar los visitantes, identificar las acciones de los visitantes y la extracción de características. Este proceso puede ser dividido en cuatro pasos:

3.1. Eliminación de robots de acceso web

Los Logs, por lo regular, contienen entradas de robots web, como los *crawlers*, *spiders*, índices y otros. La decisión de eliminarlos depende de la meta de la que tengamos por llevar a cabo. Los robots por lo general son creados por el administrador del sitio web, para generar los permisos de acceso. Y por lo regular se encuentran en un archivo .txt. Además, algunos robots pueden ser identificados si observamos el Host Name, de la dirección IP.

3.2. Filtrado de imágenes y datos ruidosos

Las paginas web, además cargan archivos de imágenes, sonidos o video con los datos. Por consiguiente, el servidor web graba las entradas que fueron solicitadas de imágenes, sonidos o videos, así como las que fueron enviadas. Generalmente estas entradas son descartadas cuando hacemos un análisis de Logs. Sin embargo, las rutas de acceso de estos archivos nos pueden dar pistas interesantes sobre la estructura, y el comportamiento del tráfico así como la motivación del usuario, en el sitio web. Algunas veces, los visitantes acceden a páginas que no existen, así como las veces que el servidor falló, o los visitantes no lograron identificarse bien. En todos estos casos el servidor web, hace un registro con un código apropiado dependiendo del caso que haya ocurrido, y entonces es decisión del analista si elimina o no estas entradas dependiendo del caso de estudio.

3.3. Extraer transacciones

Una vez que las entradas inaplicables se quiten del Log, el paso siguiente es extraer las transacciones que pertenecen a los usuarios individuales. No hay definición natural de una transacción en el panorama de la navegación del sitio. Una transacción se puede considerar como sola entrada en el registro o un sistema de entradas alcanzadas por un visitante de la misma máquina en un lapso de tiempo definido o sesión. La transacción deseada puede ser el sistema de entradas de registro de un visitante en una sola visita. El uso de los servidores Proxy establecen un límite de la sesión y puede ser el máximo de tiempo para navegar en la web en una sola visita. El registro *Referrer* está almacenado en el Log y denota el conjunto de las URLs de las páginas previas visitadas por el usuario desde las cuales el hipervínculo fue seguido. Por ejemplo, si el visitante está viendo la página */courses* y solicita actual-

mente la página */courses/postgraduate* entonces una entrada en el registro del *referrer* se puede hacer con */courses* como la página referida. El uso del registro del *Referrer* junto con la estructura del sitio web puede ayudar a identificar a visitantes únicos. El concepto de la longitud de la referencia se puede también utilizar para identificar transacciones. El módulo de la longitud de la referencia se basa en la creencia de que los visitantes pasan menos tiempo en las páginas de navegación y más tiempo en las páginas de contenido. Las transacciones se pueden también extraer usando las referencias delanteras máximas. Una referencia delantera máxima es una página que fue accesada antes de que se regrese a la anterior. Por ejemplo, en una sola visita, si un visitante ha visitado las páginas I-J-K-I-L-J entonces las referencias delanteras máximas son K y L.

3.4. Extracción y formato de las características

El paso final es extraer características de las transacciones disponibles. La extracción de las características implica el identificar las cualidades relevantes y el reducir la dimensionalidad de los datos excluyendo cualidades inaplicables. La tarea es convertir transacciones de longitud variable en vectores de longitud fija de la característica seleccionada. Los visitantes de la Web pasan generalmente más tiempo en las páginas que más les interesan. El tamaño de las páginas y la velocidad de la red se toman en consideración para definir esta característica mientras que estos dos factores afectan el tiempo transcurrido en visitar las páginas. Finalmente, los vectores de klongitud fija de cada característica se convierten en el formato requerido por la herramienta para realizar la preparación y de los datos para ser utilizados. Uno de los problemas en conseguir un cuadro exacto del acceso del sitio web es causado por los navegadores web y los servidores Proxy. Los navegadores Web, almacenan las páginas que se han visitado y las veces que se solicita la misma página. Los servidores Proxy depositan las páginas que con frecuencia fueron visitadas localmente para reducir tráfico en la red y para mejorar funcionamiento del servidor. Este problema se puede solucionar por medio de las cookies y los agentes remotos.

Es muy común que los visitantes de la web, visiten un sitio web más de una vez. También, es posible que un usuario termine su visita y otro la comience en la el mismo sitio web, en la misma computadora. Para poder identificar las transacciones individuales es que se utiliza un *time out*, es decir si el periodo entre dos visitas consecutivas es mas que el tiempo prefijado entonces podríamos considerar el fin y el comienzo de una transacción. El time out más común es de treinta minutos.

Después de haber aplicado todo este proceso, podemos decir que tenemos unos datos limpios para ser preparados con técnicas de data mining, y es así como empezaremos la preparación de los datos.

4. Preparación y procesamiento de datos

Esta etapa es conocida la la fase de preparación de datos del proceso CRISP-DM [10]. La preparación de datos genera *datos de calidad*, los cuales pueden conducir a patrones o reglas de calidad [4]. Una vez teniendo los datos de los logs preprocesados se requiere buscar y resolver las inconsistencias así como eliminar outliers[1] que pueden surgir de la consolidación e integración con las bases de datos de usuarios que contienen información personalizada de preferencias e intereses (información sociodemográfica). Los datos integrados son preparados con herramientas especializadas que se analizarán más adelante, para que sean usados como entradas fiables en la fase de descubrimiento de patrones evaluando e interpretando tendencias.

En realidad el trabajo que se hace en la parte de preparación de datos es evaluar que tipo de datos son, para que nos sirven, que podemos obtener de ellos. Por esta razón existen otras

[1] Son datos que no parecen seguir la distribución característica del resto de datos. Estos datos podrían reflejar propiedades genuinas de un fenómeno subyacente al que se está analizando o ser debidos a errores u otras anomalías que no deberían ser modeladas

sub-etapas dentro de esta etapa de preparación de datos:

- Recolección de datos
- Limpieza de datos
- Transformación de datos
- Reducción de datos

Estas sub-etapas permiten reunir una gran variedad de técnicas de análisis de datos, que permiten mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento puedan obtener mayor y mejor información [9]. A pesar de ello, no se puede asegurar una lista única y consecutiva de tareas a realizar [4] y es aquí donde la experiencia del analista juega un papel decisivo en la totalidad del proceso. No obstante, hemos creado una taxonomía posible de las técnicas y algoritmos empleados en esta fase que serán analizados a partir de las herramientas descritas en la siguiente sección.

5. Análisis de herramientas y técnicas: caso de estudio

La personalización y fidelización de usuarios en entornos abiertos como Internet es un reto que no termina de generar innovaciones y nuevas alternativas de solución. Para ello el uso de herramientas y técnicas basadas en Inteligencia Artificial y aprendizaje automático es imperativo como base para construir modelos de usuario que reflejen las preferencias y gustos de quienes representan a partir de las acciones realizadas en línea. Nuestro caso de análisis es el Proyecto Mining+Lab[2] donde disponemos de dos tipos de datos: Bases de datos de características socio-demográficas de 2.161.159 usuarios y logs mensuales del orden de 50 Gb de los hábitos de navegación del usuario. El objetivo es preprocesar automáticamente los logs que son altamente dimensionales para construir patrones de comportamiento que permitan construir modelos

[2]Mining + Lab: Increase of On-line Transactions through Smart User Models in Push and Newsletters Communications of www.emagister.com.

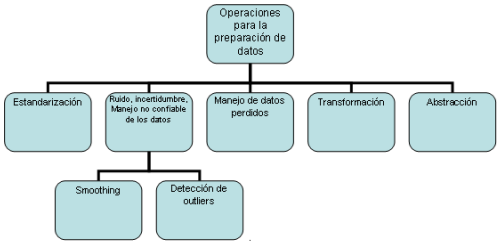

Figura 3: Esquema General de Procesamiento de Datos de las herramientas de Data Mining.

de usuario más adecuados a las preferencias de los usuarios. Estos modelos de usuario a su vez pueden estar conformados hasta por 984 atributos objetivos, subjetivos y emocionales [3], [2] y son extraidos, enriquecidos y getionados por agentes inteligentes que operan distribuidamente y descubren la sensibilidades de los usuarios a partir de sus interacciones en la web de emagister.com de manera dinámica y no-intrusiva para mejorar las recomendaciones relacionadas con los cursos de formación ofrecidos.

La mayoría de las herramientas requiere de datos en un formato pre-establecido para realizar las operaciones del esquema de la figura 3. Suponemos que la primera línea de la tabla contiene el nombre del atributo y la clase, las siguientes líneas contienen los datos que describen el número de casos y el número de valores de cada atributo por clase.

5.1. Agente preprocesador de logs para minería de uso web

Recordemos que no hay ningún método preestablecido para preprocesar logs, por esta razón se habla de buenas prácticas más que de métodos a realizar. En esta sección describimos buenas prácticas para preprocesar logs multidimensionales.

Para determinar cuáles son las acciones del usuario, se ha localizado la variable *REF*, que corresponde a las acciones que el usuario ejecuta dentro del tiempo de visita (sesión) [12], dentro del log original. Para ello, se debe separar este archivo en un formato *.csv, *.C4.5 o

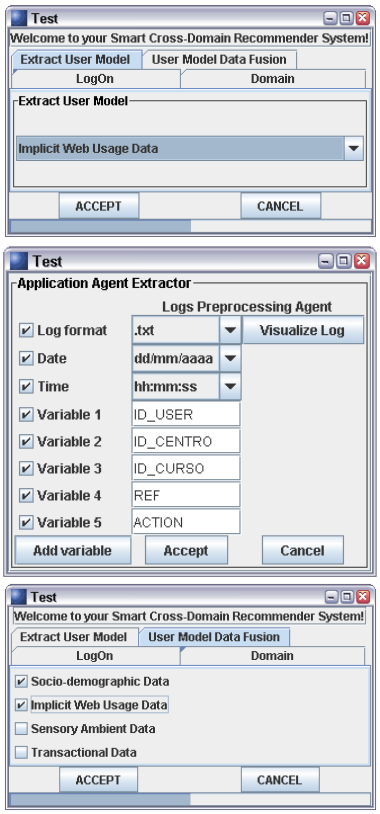

Figura 4: Diferentes Funcionalidades del Agente Preprocesador de Logs.

delimitado por tabulaciones, que generalmente son los formatos de archivos que leen las herramientas que nos sirven para descubrir las tendencias.

Sin embargo dentro del archivo *REF*, existen 256 atributos de los cuales sólo usamos 5 atributos para determinar tendencias: *IDUser*, *IDCentro*, *IDCurso*, *ACTION* y *REF*. Adicionalmente se requiere la fecha y hora de la sesión.

En razón a que cada herramienta necesita un tipo de archivo preprocesado en formatos diferentes, se ha construido un agente capaz de buscar las acciones identificadas por las variables anteriormente descritas, extraerlas y escribirlas en un archivo con formato .csv, .dat,

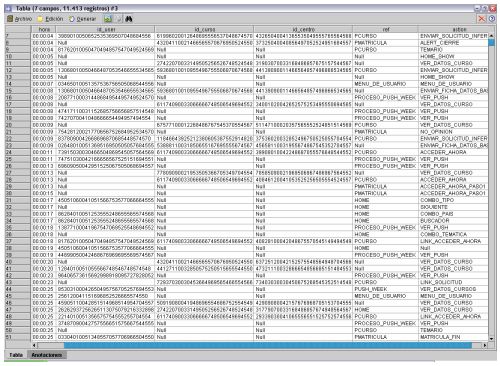

Figura 5: Log Preprocesado por el Agente.

txt, .arff, o cualquier otro. El agente opera en en un entorno distribuido y puede acceder al archivo log original en tiempo real. Fue desarrollado en Java y Perl y puede solicitar como entradas las variables que se desean extraer (ver figura 4) y el formato del archivo log en el que encontrará los atributos. Para este caso de estudio el conjunto de logs es de seis meses y cada uno con más de 10 millones de líneas. El agente entregará un log preprocesado como el que se muestra en la figura 5.

5.2. Herramientas

Hay numerosas herramientas tanto comerciales como de código abierto para procesamiento de datos. Sin embargo reduciremos el estudio a las cinco que hemos considerado más importantes: WebTrends Professional V7.0, Orange V0.9.5, MiningMart System V0.22, WEKA V3.4.4. y Clementine V8.1.

WebTrends professional V7.0

Es una herramienta comercial muy completa y trabaja directamente y en línea con el servidor de logs. Por tanto el preprocesamiento no es necesario. Permite análisis de tráfico del sitio web, análisis de ventas cruzadas, visualización jerárquica de campañas marketing y de cestas de productos con utilidades de visualización, análisis de recorridos complejos en la web con caminos visuales sencillos e intuitivos,

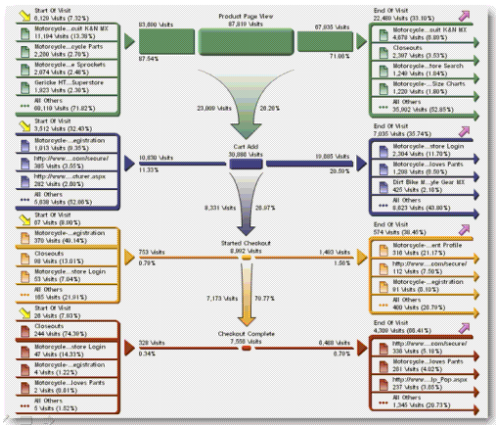

Figura 6: Análisis directo de logs desde WebTrends.

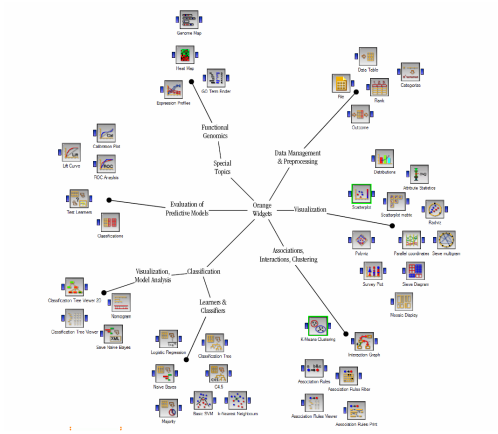

Figura 7: Características funcionales de Orange V.0.9.5.

visualización de escenarios y reportes intuitivos entre otras (ver figura 6). No dispone de de algoritmos de aprendizaje automático para realizar minería de datos [8].

Orange V0.9.5

Esta herramienta contiene técnicas de clasificación y evaluación automatizadas, como se muestra en la figura 7 [7]. Se pueden cargar logs preprocesados unicamente pero solo soporta archivos de hasta 3 Mb. Funciona con varias tipos de archivos: *.tab, *.txt, *.data, *.dat, *.rda, *.rdo. Es una herramienta completa aunque falla a la hora de cargar archivos de más de 3 Mb.

Miningmart System V0.22

Esta herramienta carga los archivos log en línea, y nos muestra estadísticas como cualquier analizador de logs. Entrega estadísticas generales, páginas más y menos visitadas, directorios más y menos visitados, visitantes con mayor actividad, ranking de referrers y páginas, entre otras variables [6]

Weka V3.4.4

Es quizá una de las mejores herramientas de código abierto para realizar minería de datos. Una de sus ventajas es que permite guardar los datos en diferentes formatos, como binary serialized instances, C4.5, *.csv y *.arff para después exportarlos a otra herramienta que disponga de alguna técnica de aprendizaje automático que Weka no posea. Adicionalmente cuenta con una gran librería de técnicas de aprendizaje automático tanto supervisadas como no supervisadas (ver figura 8) [1].

Clementine V8.1

Es una herramienta muy completa e intuitiva de la casa SPSS [11] (ver figura 9) que permite crear rutas y modelos de minería de datos de acuerdo con el proceso CRISP-DM. La versión estándar admite únicamente logs preprocesados y se puede integrar en los manejadores de bases de datos comerciales existentes. Sin embargo, se puede adquirir el módulo *Web Mining Clementine Application Template* para extraer logs directamente del servidor y trabajar con ellos en línea. Posee potentes herramientas de visualización y una gran variedad de técnicas de aprendizaje automático para clasificación, regresión, clustering y dis-

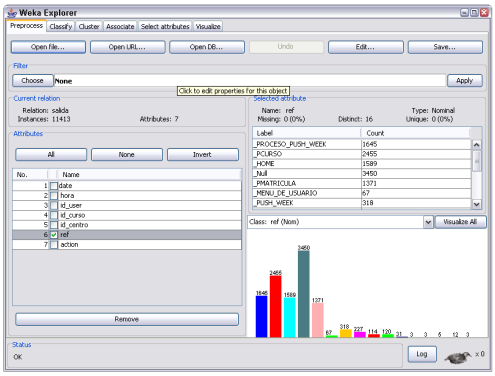

Figura 8: Análisis de Logs Preprocesados en Weka V3.4.4.

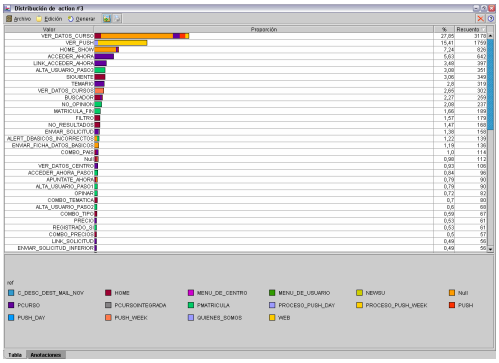

Figura 9: Log con las Acciones más Representativas del Usuario en Clementine V8.1.

cretización entre otras.

5.3. Comparación de herramientas

Al analizar las herramientas con base en características comunes para el preprocesamiento de weblogs, encontramos dos grandes ramas: las que analizan logs de forma meramente estadística y las que usan logs preprocesados para aplicar técnicas de aprendizaje automático y de minería de datos y descubrir tendencias de las preferencias de los usuarios. A continuación desarrollamos una taxonomía no exahustiva de las herramientas usadas en el caso de estudio en donde se observa claramente las utilidades,

	CARACTERISTICAS / HERRAMIENTAS	CARGA ARCHIVOS	EDITA ARCHIVOS	CONVERSION	TECNICAS Y ALGORITMOS	VISUALIZACION	FILTROS	COSTO COMPUTACIONAL
APRENDIZAJE AUTOMÁTICO	CLEMENTINE 8.1	SPSS, SAS, DB, DATOS DE USUARIO, CSV, LONGITUD FIJA	SI	*.CSV, *.TAB, *.DAT, *.DATA, SQL	REGRESION, ALGORITMOS A PRIORI K-MEDIAS, K-ONONEN	SI	SI	MEDIO
	OI ORANGE Canvas	TAB-DELIMITED (*.TAB; *.TXT) C4.5 (*.DATA), ASSISTANT FILES (*.DAT), RETIS (*.RDA, *.RDO)	NO	NO	CLASIFIERS, SELECT ATRIBUTES, ASSOCIATION, CLUSTERS	SI	NO	ALTO
	WEKA 3-4	CSV, ARFF, DB, URL	NO	C4.5, BINARY, CSV, ARFF	CLASIFIERS, CLUSTERS ASSOCIATION RULES, SELECT ATTRIBUTES	SI	SUPERVISADOS NO SUPERVISADOS	ALTO
ESTADISTICAS	MINING MART V0.22	LOG, URL	NO	NO	ESTADISTICA	SI	NO	BAJO
	WEBTRENDS Professional V.7.0	URL, *.TXT, *.XLS *.LOG	NO	NO	ESTADISTICA	SI	NO	ALTO

Figura 10: Comparación de Herramientas.

ventajas y desventajas de cada una (ver figura 10).
La figura 11 muestra las herramientas más usadas en minería de datos con sus funcionalidades de aprendizaje automático.
Dentro de las herramientas que contienen algoritmos de aprendizaje automático y técnicas de preparación de datos hemos desarrollado una sub-taxonomía con base en la aplicación de filtros tanto supervisados como no supervisados para atributos e instancias. No existe una herramienta que contenga todas las funcionalidades pero definitivamente Weka y Clementine son las más completas.

6. Conclusiones

El preprocesamiento de datos no tiene una secuencia programada ni establecida por un método único. El papel que juega la experiencia del analista de datos es relevante a la hora de extraer las acciones que mejor describen las preferencias y gustos de los usuarios en sistemas de recomendación basados en la web. Las características óptimas para realizar minería de datos para modelado de usuarios y extraer patrones de comportamiento en línea constituyen un proceso que combina la sinergia creada por las herramientas, los métodos de aprendizaje automático y los algoritmos. Puede decirse que el preprocesamiento de logs es un *arte* que se vale de herramientas para crear una téc-

TECNICAS Y ALGORITMOS			
HERRAMIENTAS / TECNICAS Y ALGORITMOS	CLEMENTINE 8.1	Qt ORANGE Canvas	WEKA 3-4
RED NEURONAL	X		X
C5.0	X		
ARBOL C&R	X	X	X
KOHONEN	X		
K-MEDIAS	X		X
BIETAPICO	X		
A PRIORI	X		
GRI	X		
SECUENCIA	X		X
PCA/FACTORIAL	X		
REGRESION	X	X	X
LOGISTICA	X	X	X
REDES BAYESIANAS		X	X
MAYORIA		X	
K NEAREST NEIGHBOURS		X	X
C4.5		X	X
SVM		X	X
PERCEPTION			X
SMO			X
IB1			X
K STAR			X
BOOST			X

Figura 11: Algunas Herramientas usadas en la Prepaparación de Datos con Técnicas y Algoritmos de Aprendizaje Automático.

nica a partir de buenas prácticas.

Los archivos Logs, están diseñados para mostrarnos comportamientos de usuarios en la web, a partir de ellos podemos tener datos de calidad preprocesándolos previamente para construir modelos de usuario. Las herramientas comerciales para gestionar este tipo de archivos por lo general son de costo elevado tanto por valor de sus licencias como por los recursos de hardaware requeridos. En este artículo hemos presentado un agente preprocesador de logs para extraer modelos de usuario como parte de un sistema de recomendación para múltiples dominios de aplicación.

El caso de estudio real ha permitido crear una taxonomía de uso de las herramientas de extracción, preprocesamiento y preparación de datos con grandes bases de datos multidimensionales. Con base en esa taxonomía podemos concluir que la herramienata más completa es Weka. Clementine es la herramienta más robusta e intuitiva y nos permite tener casi las mismas características que Weka, agregándole que tiene un costo computacional menor. Por último Webtrends es la herramienta líder en el mercado del marketing en línea pudiendo obtener seguimiento de campañas en tiempo real.

Como trabajo futuro nuestra investigación se orientará al aprovechamiento de dichos modelos de usuario en redes sociales usando conceptos y métricas de Minería de Uso Web.

Reconocimientos

Esta investigación ha sido financiada por el Proyecto *Mining + Lab: Increase of Online Transactions through Smart User Models in Push and Newsletters Communications of www.emagister.com* a través del Centre d'Innovació en Informàtica i Electrónica Industrial i Sistemes Intelligents (EASY) de la Univeristat de Girona, Agents Inspired Technologies S.A. y emagister.com.

Referencias

[1] Srikant R. Dai, H. and C. Zhang, editors. *Evaluating the replicability of significance tests for comparing learning algorithms.* Springer-Verlag., 2004.

[2] G. González, C. Angulo, B. López, and J.LL. de la Rosa. Smart user models: Modelling the humans in ambient recommender systems. In *Proceedings of Workshop on Decentralized, Agent Based and Social Approaches to User Modelling (DASUM 2005). In conjunction with 10th International Conference on User Modelling (UM'05) Edinburgh, Scotland.*, 25-29 July 2005.

[3] G. González, B. López, and J.LL. de la Rosa. A multi-agent smart user model for cross-domain recommender systems. In *Procceedings of Beyond Personalization 2005: The Next Stage of Recommender Systems Research. International Conference on Intelligent User Interfaces IUI'05. San Diego, California, USA.*, January 9 2005,.

[4] F. Herrera, J. Riquelme, and F. Ruiz. Preprocesamiento de datos. ii reunión red nacional de data mining y machine learning. Technical report, mayo 2004.

[5] R. Kosala and H. Blockeel. Web mining research: A survey. In *SIGKDD Explorations*, pages 1–15. SIGKDD, 2000.

[6] Ingo Mierswa and Katharina. Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:127–149, 2005.

[7] M. Mozina, J. Demsar, M. Kattan, and B. Zupan. Nomograms for visualization of naive bayesian classifier. In *Principles and Practice of Knowledge Discovery in Databases (PKDD-2004). Pisa, Italy.*, pages 337–348, 2004.

[8] Jim Novo. Increase customer retention by analyzing visitor segments, December 2004.

[9] Zhang S., Zhang C., and Yang Q. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):2003, 375-381.

[10] C. Shearer and et.al. The cross industry standard process for data mining crisp-dm model: The new blueprint for data mining. *Journal of Datawarehousing*, 5(4):13–22, Fall 2000.

[11] SPSS. http://www.spss.com/clementine/.

[12] J.R. Velasco and L. Magdalena. Minería de datos para análisis del uso de sitio web. In *I Workshop Aprendizaje y Minería de Datos. Conferencia Iberoamericana de Inteligencia Artificial Iberamia 2002. Sevilla.*, 12-15 Noviembre 2002.

[13] O. Zaiane and et.al. Querying the World-Wide Web for Resources and Knowledge. In *Workshop on Web Information and Data Management (WIDM'98)*, pages 9–12, Washington D.C:, November 1998.