INTEGRACIÓN de HERMINWEB CON LA ETL SPOON DE PENTAHO

Julio Antonio Hernández Pérez
Universidad de la Ciencias Informáticas, La Habana,Cuba, jhperez@uci.cu

Husseyn Despaigne Reyes
Universidad de la Ciencias Informáticas, La Habana,Cuba, hdespaigne@uci.cu

ABSTRACT

Data mining is responsible for finding variations of behavior in a high volume of information. This paper presents a software tool that can discover patterns in the navigational records (logs), generated by the proxy server that provides Internet access to users from the University of Informatics Sciences (UCI). This tool is useful for the Management of Networks and Information Security (DRSI), it provides information necessary for decision-making. Develop a process of Knowledge Discovery in Databases (KDD) in which to apply the task of bringing together, in order to find classes of users in the use of Internet browsing and task association rules to find relationships between The attributes of these classes. To prepare the data is communicated with the ETL Pentaho Spoon and performs a distributed processing of the logs, reducing the time it takes for this phase and the complexity in programming the same. Integrates two technologies with different characteristics to implement the clustering algorithm and the Rules of Association (JAVA and Python), with a design extensible libraries of algorithms of different technologies. The design hierarchical storage streamlines information data query. Also manages the integration of navigational records user data from other information systems.

Keywords: Knowledge Discovery in Databases, Rules of Association, Pentaho

RESUMEN

La Minería de Datos se encarga de encontrar variaciones de comportamiento en un alto volumen de información. En este trabajo se presenta una herramienta informática que permite descubrir patrones de comportamiento en los registros de navegación (logs), generados por el servidor proxy que da acceso a Internet a los usuarios de la Universidad de las Ciencias Informáticas (UCI). Esta herramienta es de utilidad para la Dirección de Redes y Seguridad Informática (DRSI), pues le brinda información necesaria para la toma de decisiones. Desarrolla un proceso de Descubrimiento de Conocimientos en Base de Datos (KDD) en el cual se aplican la tarea de Agrupamiento, con el fin de encontrar clases de usuarios en el uso de la navegación por Internet y la tarea Reglas de Asociación para encontrar relaciones entre los atributos de estas clases. Para preparar los datos se comunica con la herramienta ETL Spoon de Pentaho y realiza un procesamiento distribuido de los logs, reduciendo el tiempo que demora esta fase y la complejidad en cuanto a la programación de la misma. Integra dos tecnologías con características diferentes para la ejecución del algoritmo de Agrupamiento y el de Reglas de Asociación (JAVA y Python), con un diseño extensible a bibliotecas de algoritmos de diferentes tecnologías. El diseño jerárquico en el almacenamiento de la información agiliza las consultas de datos. Además logra integrar a los registros de navegación datos de los usuarios de otros sistemas de información.

Palabras claves: integración de tecnologías, minería de datos, programación distribuida, toma de decisiones.

1. INTRODUCCIÓN

El cursar de los años y el desarrollo de las tecnologías han generado un crecimiento considerable de datos (Hernández Orallo, y otros, 2004). El uso de esta información para el perfeccionamiento de los procesos de las empresas e instituciones crea la necesidad de desarrollar nuevas técnicas y herramientas para analizar esta enorme cantidad de datos (Olmos, y otros, 2007). El análisis de datos es una tarea que consiste en buscar o encontrar tendencias o variaciones de comportamiento en los mismos, de tal manera que esta información resulte de utilidad para los usuarios finales. A estas tendencias o variaciones se le conocen como patrón, los cuales si son de importancia y útiles para el dominio en cuestión se le denomina conocimiento (Olmos, y otros, 2007).

Una de las técnicas para el análisis de datos es la Minería de Datos. Larose plantea que la Minería de Datos: "... es el proceso de descubrir nuevas correlaciones significativas, patrones y tendencias ocultas a través de grandes cantidades de datos almacenados en los repositorios, utilizando tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas" (Larose, 2005).

Uno de los escenarios donde se ha aplicado la Minería de Datos es la World Wide Web (WWW), su acelerado crecimiento y la competencia entre las organizaciones ha traído la necesidad de mejorar la calidad de los sitios web, utilizando como base el comportamiento de los usuarios que lo usan. Se le ha denominado Minería Web (MW) al descubrimiento de información útil en la WWW. La MW posee varias clasificaciones atendiendo al contenido que se analiza, una de ellas es la Minería de Uso de la Web (Web Usage Mining: WUM) centrada en el análisis de los logs o registros de navegación.

Al navegar por la Web las computadoras dejan rastros o registros de navegación en los servidores donde están hospedados los sitios web y en los que brindan el acceso a Internet. Los servidores proxy encargados de gestionar el acceso a Internet de algunas instituciones generan un alto volumen de registros de navegación, que archivan la navegación de un grupo de usuarios de una determinada organización.

La Universidad de las Ciencias Informáticas (UCI) cuenta con un servicio de navegación por Internet para miles de usuarios y posee un insuficiente ancho de banda para brindar un buen servicio. Esta situación provoca que la navegación sea lenta, además no todos los usuarios hacen un uso de Internet acorde a los intereses de la Universidad, no existiendo ningún mecanismo que diferencie en cuestiones de calidad de servicio a los usuarios que realizan una navegación enmarcada en los intereses de la institución. La Dirección de Redes y Seguridad Informática (DRSI) encargada de garantizar el correcto funcionamiento de la red de computadoras en la UCI, actualmente no posee herramienta alguna capaz de analizar el alto volumen de información presente en los logs del proxy; enfocada en la búsqueda de patrones que describan el uso de la navegación por parte de los usuarios de la institución, dificultando así la toma de decisiones para mejorar los problemas planteados anteriormente.

Existe un proceso definido para el análisis de datos con el fin de encontrar patrones en grandes cantidades de datos llamado Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases: KDD). El análisis de estos grandes volúmenes de datos es engorroso realizarlo manualmente, siendo necesaria su automatización. En tal sentido se desarrolló una Herramienta de Minería de Uso de la Web aplicada a los registros de navegación del proxy (HERMINWEB), encargada de automatizar el proceso KDD para el escenario de los registros de navegación por Internet de la UCI. La herramienta HERMINWEB fue desarrollada por los autores de este artículo. HERMINWEB que automatiza un proceso de Extracción de Conocimientos en Base de Datos (KDD), utilizando tecnologías libres, apoyado en la Minería de Datos, aplicado a los registros de navegación por Internet almacenados por el servidor proxy.

En KDD, la fase de Preparación de los Datos requiere alrededor del 60% del esfuerzo a realizar durante todo el desarrollo del proceso (Molina López, y otros, 2006). En la primera versión de HERMINWEB, para llevar a cabo esta fase fue probada en varias circunstancias de trabajo la herramienta ETL Spoon, perteneciente a la suite de aplicaciones para Inteligencia de Negocio (BI): Pentaho BI Suite Enterprise Edition, comportándose correctamente salvo en la ejecución por líneas de comandos, donde tuvo algunas fallas. Por tanto se llevó a cabo la implementación de varias herramientas para la integración de datos, centradas en el ámbito y características de la UCI. En una segunda iteración, se probó la versión más reciente de Pentaho, la 4.0.1, la cual no presentó problemas para ejecutarla por líneas de comandos, posibilitando de esta forma su integración con HERMINWEB,

con lo cual se disminuye considerablemente el esfuerzo durante la fase de Preparación de los Datos en cuanto a tiempo y complejidad en la programación del sistema .

De acuerdo a la situación descrita se tiene el siguiente problema científico ¿Cómo mejorar la fase de Preparación de los Datos en la herramienta HERMINWEB en cuanto a tiempo y complejidad en la implementación del sistema? El objetivo de la presente investigación es integrar la ETL Spoon de Pentaho con la herramienta HERMINWEB para mejorar la fase de Preparación de los Datos que se realiza en esta última como parte de un Proceso de Descubrimiento de Conocimiento en Base de Datos apoyado en las tareas de Agrupamiento y Reglas de Asociación (Hernández Orallo, y otros, 2004) que permite obtener patrones descriptivos del uso de la navegación por Internet de los usuarios de la UCI.

El presente trabajo muestra la integración de un sistema encargado de automatizar el proceso KDD para el escenario de los registros de navegación por Internet de la UCI con la herramienta Pentaho. Dicho proceso es guiado por la metodología CRISP-DM (Chapman, y otros, 2000) mientras que el desarrollo de la aplicación informática es orientado por la metodología RUP (Jacobson, y otros, 2000).

2. DESARROLLO

El término Minería de Datos en muchas ocasiones se utiliza como sinónimo de Descubrimiento de Conocimiento en Bases de Datos, siendo en realidad la Minería de Datos una de las fases en las que está compuesto el proceso KDD, según la metodología CRISP-DM es la llamada Modelado. La Figura 1 muestra las fases definidas por la metodología CRISP-DM para un proceso KDD (Chapman, y otros, 2000):

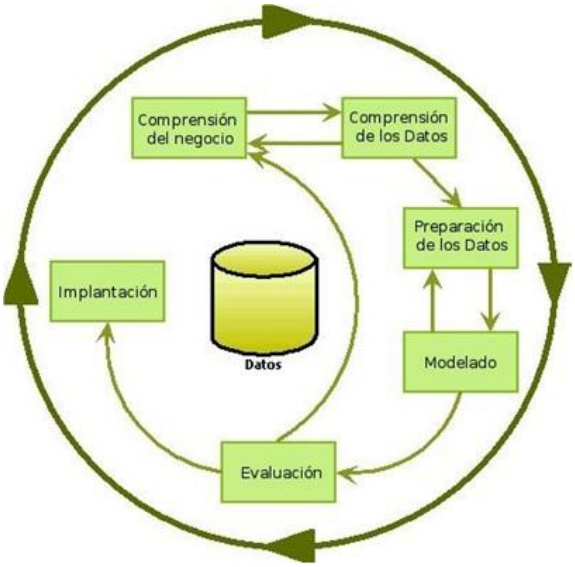

Figura 1 - Fases del proceso KDD según la metodología CRISP-DM (Ordoñes Leyva, y otros, 2010)

Primeramente se debe estudiar el problema que se desea resolver, es imprescindible comprender a profundidad el dominio donde se desenvuelve la investigación, centrando el análisis en las necesidades de la organización, así como para definir y priorizar los objetivos del negocio. En la Preparación de Datos se determinan las fuentes de información que pueden ser útiles; se transforman los datos a un formato común, además se detectan y se resuelven las inconsistencias presentes en los mismos. Posteriormente se eliminan o corrigen los datos incorrectos, decidiéndose la estrategia a seguir con los datos incompletos; además, se consideran únicamente aquellos atributos que van a ser relevantes (Hernández Orallo, y otros, 2004).

En la fase Modelado se aplica el modelo, la tarea, la técnica y el algoritmo seleccionado para la obtención de reglas y patrones. Luego en la fase de Evaluación se comprueban los patrones y se analizan por expertos, se puede regresar a la fase Análisis del Problema en caso de querer perfeccionar los resultados. Finalmente, en la fase de

Explotación se comparte el nuevo conocimiento con los interesados. Las fases que componen el KDD hacen que su desarrollo sea un proceso iterativo e interactivo con el usuario (Hernández Orallo, y otros, 2004).

Minería de Datos aplicada a los registros de navegación en la UCI

La UCI cuenta con un servicio de navegación por Internet para miles de usuarios. Para ello cuenta con servidores proxy que gestionan todo el flujo de peticiones realizadas. Los sistemas actualmente instalados y en explotación en la DRSI (Martín Álvarez, y otros, 2007) no cubren todo el conocimiento implícito en los registros de navegación, dificultando la toma de decisiones a la DRSI. Por otra parte se cuenta con sistemas de gestión de información de los trabajadores y gestión académica de los estudiantes. Esta información en conjunto con los registros de navegación es de mucha utilidad e interés para la DRSI, con ella se pueden encontrar patrones que describan el uso de la navegación de los diferentes usuarios de la institución.

Para extraer esta información es necesario desarrollar una herramienta que permita: obtener y mezclar los datos registrados por el servidor proxy con la información de los usuarios almacenada en los sistemas de los trabajadores y estudiantes; realizar la fase de Preparación de los Datos en el menor tiempo posible; extraer patrones descriptivos presentes en los datos, enfocados en las tareas de agrupamiento y reglas de asociación, con el fin de encontrar clases de usuarios que se comporten de manera similar en el uso de la navegación por Internet y relaciones entre los atributos de estas clases, ayudando a la DRSI en la toma de decisiones. La herramienta forma parte de una Plataforma de Gestión de Servicios Telemáticos (PGST) escrita en Python y desarrollada en el Centro de Telemática (TLM) de la UCI (Pérez Hurtado, y otros, 2010).

La aplicación posee una arquitectura de 4 Capas. Entre las que se encuentran la interfaz de usuario, GUI; Servicios, donde se realiza el negocio del sistema; la encargada de las acciones para comunicar el negocio con la interfaz de usuario llamada Entorno de Ejecución; y el Acceso a Datos para obtener y almacenar los datos. La configuración de todos los componentes de cada una de las capas a utilizar se encuentran en un fichero XML donde se definen para cada una de las interfaces, su implementación. La Figura 2 muestra el diseño de la arquitectura del sistema.

Figura 2 - Arquitectura de HERMINWEB

En Dominio estarán las clases pertenecientes a la información que se compartirá en las capas de la herramienta. La PGST posee un marco de trabajo o Framework para homogeneizar los sistemas que forman parte de la misma. Este se especializa en la gestión de servicios telemáticos. Brinda además, soporte para la creación de interfaces gráficas de usuario y la comunicación con servicios web y bases de datos. Las interfaces gráficas de usuario son creadas utilizando el Framework Qt mediante la implementación PyQt para Python.

Para el desarrollo de HERMINWEB fue de gran importancia la utilización de dos herramientas fundamentales, la ETL Spoon de Pentaho y RapidMiner, empleadas en las fases de Preparación de los Datos y Minería respectivamente.

RapidMiner es una herramienta creada en la Universidad de Dortmund para el descubrimiento del conocimiento y la Minería de Datos. Es un entorno con muchos algoritmos de aprendizaje y otras utilidades añadidas, está desarrollada sobre el lenguaje Java y funciona en los sistemas operativos más conocidos, constituyendo un software de código abierto y de libre distribución. Desde la perspectiva de la visualización ofrece representaciones de datos en dispersión en 2D y 3D; coordenadas paralelas y grandes posibilidades de transformar las visualizaciones de los datos (Sitio Oficial de RapidMiner, 2011).

La compañía Pentaho es una alternativa de código abierto para la Inteligencia de Negocio. Desarrolla varias herramientas entre las que se encuentra la ETL Spoon, utilizada para la transformación y limpieza de datos. Una de estas herramientas engloba a las demás, la cual se denomina Pentaho BI Suite Enterprise Edition que provee reportes, integración de datos, Minería de Datos y una plataforma de BI. Pentaho representa una solución completa de herramientas para la integración de datos (Pentaho Company, 2011).

Fases de KDD en la aplicación

La herramienta realiza cada una de las fases descritas en la metodología CRISP-DM, desarrollando un flujo de procesos, entre las que se encuentran las correspondientes a las fases de: Preparación de Datos, Modelado, Evaluación y Explotación de los resultados. Posee dos modos de ejecución: líneas de comandos e interfaz gráfica. En lo que sigue del documento se hará énfasis sólo en la fase de Preparación de los Datos.

Preparación de Datos

Para realizar la Preparación de los Datos se utilizaron varias fuentes de información con el objetivo de mezclar los datos de la navegación con las características de los usuarios, para así obtener patrones descriptivos del uso de la navegación. Dichas fuentes son las siguientes:

Los registros del servidor proxy.

Los sistemas de recursos humanos pertenecientes a los trabajadores (son dos servicios web, uno para los trabajadores de la Universidad y otro para los que brindan servicios en ella, en este documento se tratarán como trabajadores internos y externos respectivamente).

El sistema de gestión académica estudiantil llamado Akademos.

Las clasificaciones de sitios web de Internet usando BlackLists.

Los datos de los trabajadores son obtenidos mediante servicios web y los pertenecientes a los estudiantes accediendo directamente a la base de datos de Akademos, en este caso el gestor es Microsoft SQLServer. Al terminar este proceso se tienen los datos listos para ser analizados en la fase de Minería. La preparación de los datos es una de las tareas que más procesamiento computacional necesita, pues analiza un volumen elevado de información (la media de logs generados en un mes es 30 GigaBytes). En esta fase se procesan los logs del servidor proxy y se mezclan con los datos de los usuarios. Las computadoras que intervienen en la investigación cuentan con un CPU Intel Dual Core con 2.1 GHz de velocidad y 1.0 GB de RAM, situación que requiere ser analizada para minimizar el costo en cuanto a recursos y tiempo.

Una de las cuestiones importantes en el desarrollo de aplicaciones es el tiempo de ejecución en la realización de las tareas. En tal sentido, la utilización de la ETL Spoon de la suite de aplicaciones Pentaho fue de gran importancia para la preparación de los datos. HERMINWEB sólo se encarga de solicitar al usuario las clasificaciones que desee realizar de cada uno de los atributos necesarios en las vistas minables para el estudio del uso de las cuotas de navegación por Internet por parte de los usuarios de la UCI. Estas clasificaciones son procesadas para que se creen dinámicamente las consultas y transformaciones a procesar en Pentaho a través de un parser que construye el XML de dichas transformaciones, que se ejecutan por líneas de comandos, aumentando el rendimiento de la aplicación, debido a que Pentaho ejecuta las transformaciones con mayor rapidez sin necesidad de levantar su interfaz gráfica. En la Figura 3 se muestra cómo ocurre el flujo.

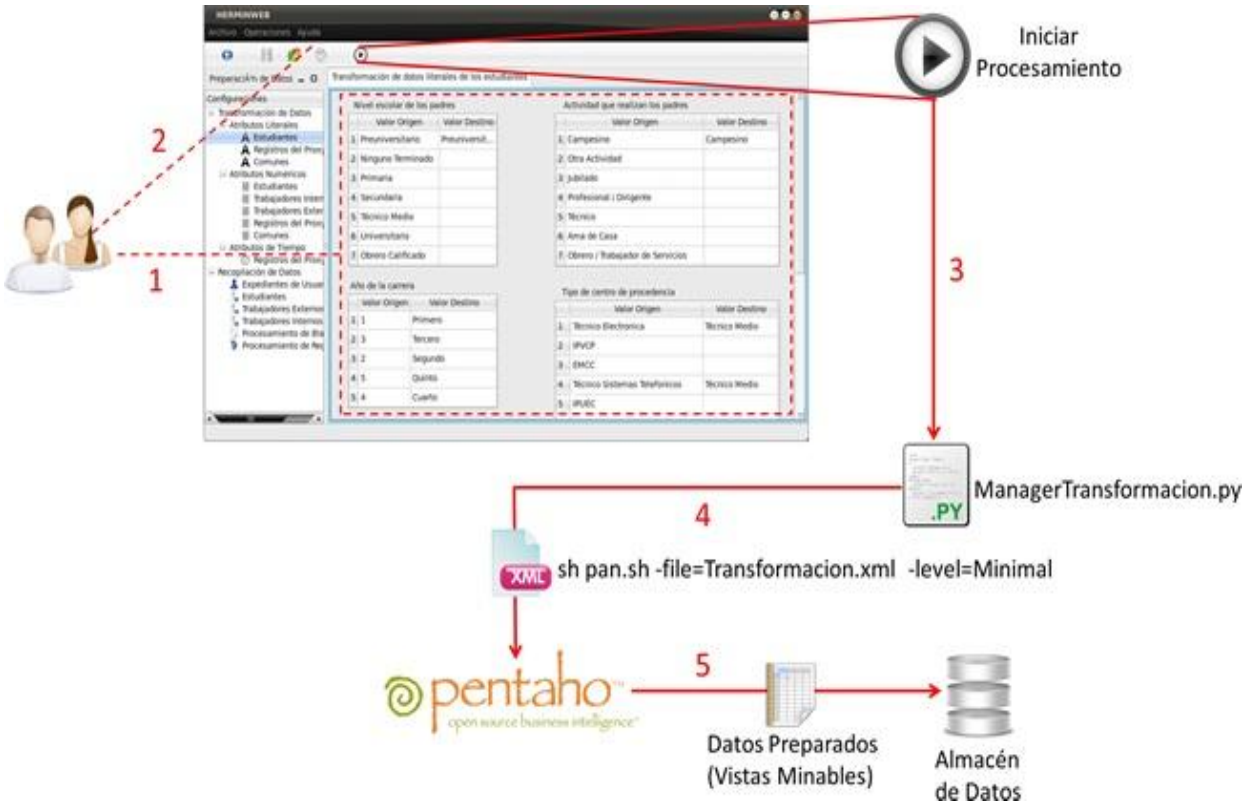

Figura 3 - Integración de HERMINWEB con la ETL Spoon de Pentaho.

La integración de HERMINWEB con la ETL Spoon de Pentaho se diseñó tal como se muestra en la Figura 4:

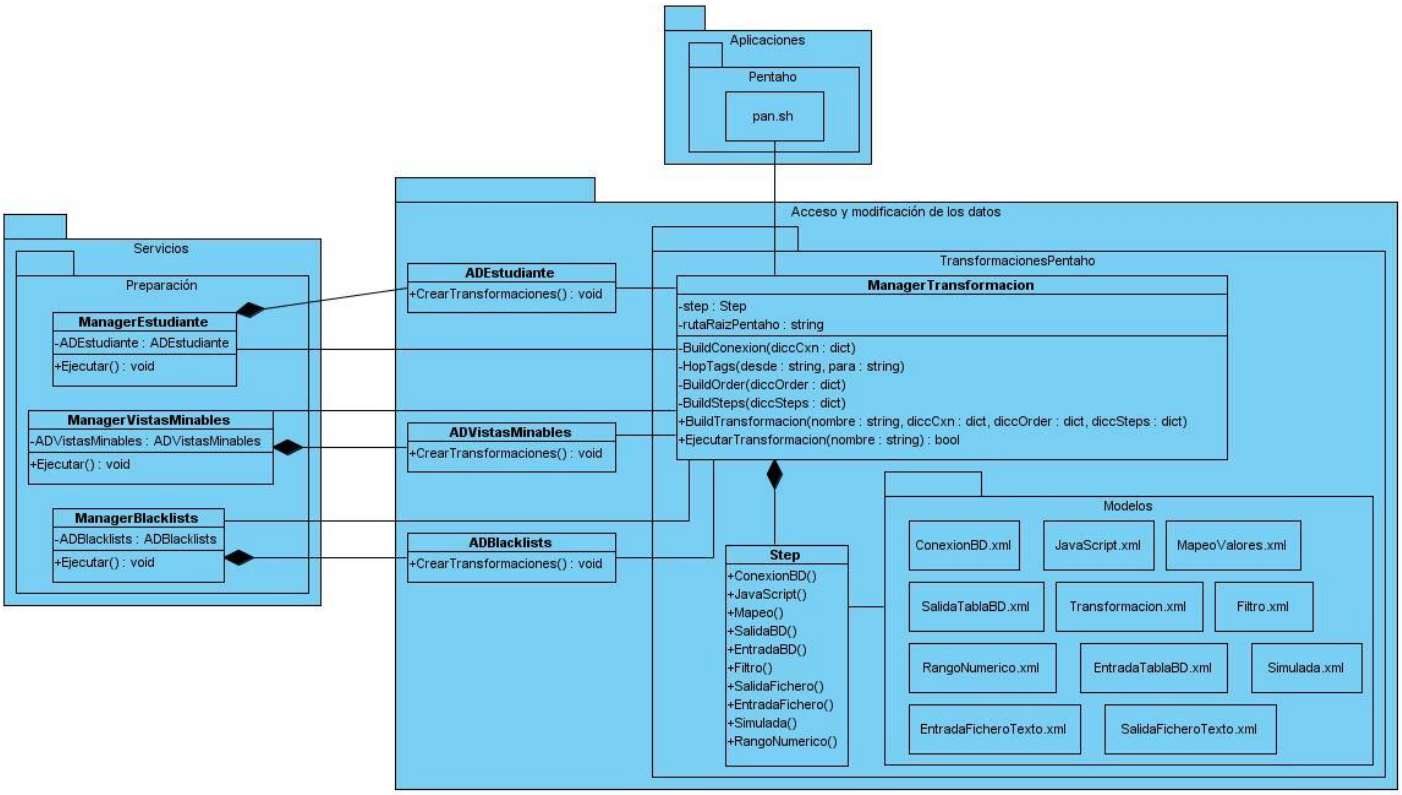

Figura 4 - Diseño de la integración de HERMINWEB con la ETL Spoon de Pentaho.

Para cada paso (step) de la ETL Spoon se creó un fichero XML (modelo) con la estructura correspondiente al mismo, tal y como lo genera Pentaho. Estos ficheros se encuentran en el paquete Modelos, el cual se ubica dentro del paquete TransformacionesPentaho en la capa de Acceso y modificación de los datos (CapaAD). Cada modelo es utilizado por la clase Step, que es la responsable de cargar en memoria el contenido de los modelos según sean los componentes que serán utilizados en la transformación y agregar en tiempo de ejecución los parámetros que sean necesarios para que la clase ManagerTransformacion construya la misma. Los parámetros antes mencionados pueden ser conexiones a bases de datos, consultas SQL que HERMINWEB construye dinámicamente a partir de los atributos seleccionados por el usuario, nombre del step, valores que se deben transformar, valores a utilizar para aplicar filtros, etc. La clase ManagerTransformacion cuenta con las funcionalidades necesarias para crear las conexiones a bases de datos (BuildConexion), los steps utilizando la clase Step como se explicó anteriormente (BuildSteps), el orden en que debe ejecutarse la tarea que realiza cada step (BuildOrder) y finalmente construye la transformación a partir de la selección realizada por el usuario utilizando la función BuildTransformacion. Con la función EjecutarTransformacion, se ejecuta la transformación por línea de comandos en la ETL Spoon.

Las clases de la CapaAD en las que se necesita crear alguna transformación utilizan una instancia de ManagerTransformacion para realizar esta tarea, usando para ello el método CrearTransformaciones que invoca a BuildTransformacion. Estas transformaciones son ejecutadas en el momento que se precise durante la preparación de los datos desde la capa de Servicios, donde cada clase del negocio que maneja su clase correspondiente en la CapaAD, posee un método denominado Ejecutar que realiza una llamada a la función EjecutarTransformacion de ManagerTransformacion.

La ETL Spoon no pudo utilizarse para obtener los datos de los trabajadores provenientes de los servicios web debido a que presentó problemas al ejecutar el componente que gestiona el acceso a este tipo de servicios. Esta tarea siguió siendo responsabilidad de HERMINWEB, pero no hizo que la utilización de Pentaho fuera descartable.

Como HERMINWEB tiene que procesar un alto volumen de información proveniente de los registros de navegación fue necesario buscar alguna alternativa para realizar esta tarea en el menor tiempo posible. Se creó una transformación con este propósito para ser ejecutada en Pentaho, la cual tenía una gran complejidad debido a la estructura que poseen los logs del proxy. En la mencionada transformación era necesario utilizar un componente denominado Agrupar por, para agrupar las peticiones según los atributos de los logs que se hubiesen seleccionado para crear las vistas minables, pero esto imponía una restricción, y es que para que este componente arroje los mismos resultados que la sentencia GroupBy de SQL, los datos de entrada al componente deben estar ordenados, haciendo de carácter obligatorio primeramente bloquear el paso con el componente Paso de Bloqueo y luego ordenarlo con el step Ordenar filas antes de utilizar el Agrupar por. Esto evidentemente ralentizaba el procesamiento de los datos, ya que la rapidez de Pentaho radica en realizar el flujo completo de la transformación paralelamente, es decir, sin esperar que termine una fila para procesar la otra. No obstante, esta transformación fue probada aplicando otras dos alternativas que proporciona la ETL Spoon: la primera utilizando un servidor esclavo y la segunda usando un grupo o cluster de servidores para realizar el procesamiento en paralelo. En ambos casos el resultado no fue mejor que la variante de realizar esta tarea desde HERMINWEB como se explica más adelante. Una posible solución a este problema era almacenar los resultados sin agrupar en una tabla temporal en la base de datos y luego agruparlos y almacenarlos en la tabla final, pero dado el alto número de filas que se generan en la creación de las sesiones y la gran cantidad de usuarios a procesar, evidentemente esta solución es ineficiente.

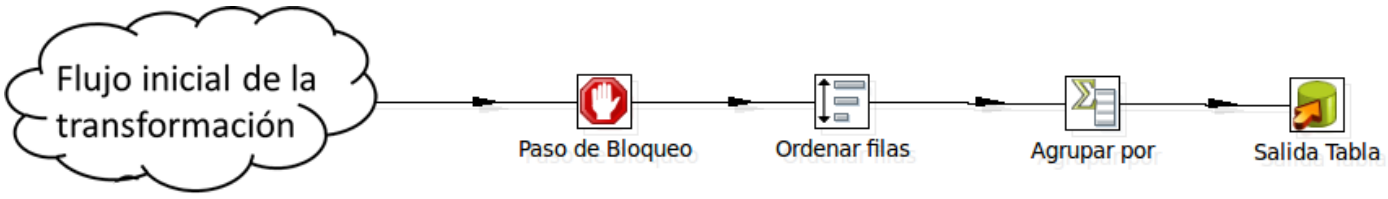

Figura 5 - Ejemplo de uso del componente Agrupar por de la ETL Spoon.

Se analizó la posibilidad de utilizar la tecnología de código abierto Apache Hadoop (Holt, 2011) (Venner, 2009), que ha venido a convertirse en una de las tecnologías de preferencia para empresas que necesitan reunir, almacenar y procesar efectivamente grandes cantidades de información estructurada y compleja, reduciendo los costos para ello. Pero, si bien Hadoop es muy poderoso, presenta grandes desafíos a los usuarios: en su forma cruda carece de interfaces fáciles de usar para análisis efectivos en tiempo y costo, presenta una empinada curva de aprendizaje técnico, no hay suficiente personal técnico cualificado y existe poca disponibilidad de aplicaciones de desarrollo e implementación para la integración de datos y la realización de inteligencia de negocios con Hadoop. Pentaho rompe estas barreras y hace que Hadoop sea fácil de usar, disminuyendo la complejidad y la curva de aprendizaje para los usuarios Hadoop y permitiendo integrar fácil y rápidamente Hadoop dentro de arquitecturas de datos existentes en bases de datos y almacenes de datos empresariales (Pentaho Company, 2011).

Sin embargo, no fue posible aprovechar la integración de estas tecnologías producto de que el equipo de desarrollo no dispone de una versión de Pentaho Hadoop para Linux. Sólo se pudo adquirir una versión de prueba por 30 días para Windows. Hay que tener en cuenta que HERMINWEB se desarrolla para ser usado en sistemas operativos GNU/Linux, en concordancia con la política migratoria que lleva a cabo el país en busca de la utilización de soluciones libres que fomenten la soberanía tecnológica.

El procesamiento secuencial en la preparación de los datos fue descartado. Una de las soluciones aplicadas fue el procesamiento basado en hilos, la cual mejoró el tiempo de ejecución pero seguía siendo alto, para 30 GB de registros del proxy consumió 2 horas aproximadamente en realizar el procesamiento. Se tomó como alternativa la programación distribuida, siendo la definitiva a utilizar.

3. CONCLUSIONES

En la investigación se integró Pentaho con HERMINWEB para mejorar la fase de preparación de los datos que realiza esta última como parte de un proceso KDD para descubrir patrones en la navegación por Internet de los usuarios de la UCI. La arquitectura en 4 capas facilita la reutilización por el desacople de las funcionalidades que implica, pudiéndose utilizar en otras investigaciones. La guía de CRISP-DM permitió que se desarrollara el proceso KDD ágilmente, proporcionando una estructura comprensible para su aplicación en otros escenarios.

La herramienta logró integrar satisfactoriamente varias fuentes de información con formatos diferentes como: ficheros de logs, base de datos relacionales (SQLServer y PostgreSQL), LDAP y servicios web, siendo de gran importancia en este sentido la utilización de la ETL Spoon de la suite de aplicaciones Pentaho, que permitió acelerar el proceso de preparación de los datos, disminuir la complejidad en la implementación del sistema y liberar de responsabilidades HERMINWEB. No obstante, es necesario seguir profundizando en el trabajo con Pentaho para lograr obtener los datos de los trabajadores provenientes de los servicios web mediante la utilización de esta herramienta y seguir de cerca el avance de Pentaho Hadoop.

La integración de HERMINWEB con la ETL Spoon de Pentaho y la programación distribuida para el procesamiento de grandes volúmenes de datos constituyen un punto de referencia en el TLM, ya que en el mismo existe poca experiencia en el desarrollo de aplicaciones de minería de datos, en el trabajo con herramientas para la recopilación, limpieza y trasformación de datos, en la programación distribuida, así como en la integración de varias aplicaciones en un solo sistema, que gestionan información paralelamente sobre las mismas fuentes de datos.

REFERENCES

Chapman, Pete, y otros. 2000. CRISP-DM 1.0: Step-by-step data mining guide. s.l. : SPSS Inc., 2000.

Hernández Orallo, José, Ramírez Quintana, María José y Ferri Ramírez, César. 2004. Introducción a la Minería de Datos. Madrid : Pearson Prentice Hall, 2004.

Holt, Bradley. 2011. Writing and Querying MapReduce. s.l. : O'Reilly Media, Inc., 2011.

Jacobson, Ivar, Booch, Grady y Rumbaugh, James. 2000. El proceso unificado de desarrollo de software. s.l. : Pearson Adisson-Wesley, 2000.

Jain, Anil K. 2009. Data Clustering: 50 Years Beyond K-Means. Universidad del Estado de Michigan. Michigan : s.n., 2009.

Larose, Daniel T. 2005. Discovery Knowledge in Data: An Introduction to Data Mining. New Jersey : Wiley, 2005.

Martín Álvarez, Luis Orlando y García Martínez, Yassier. 2007. Sistema de Reportes de Navegación por Internet. Unicersidad de las Ciencias Informáticas. 2007.

13. Molina López, José Manuel y García Herrero, Jesús. Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA. Universidad Carlos III de Madrid. Madrid : s.n., 2006.

Olmos, I. y J, González. 2007. Minería de Datos. Puebla : s.n., 2007.

Ordoñes Leyva, Yoanni y Avilés Vázquez, Ernesto. Herramienta informática de Minería de Uso de la Web sobre los registros de navegación por Internet. Universidad de las Ciencias Informáticas. Ciudad de La Habana : s.n., 2010.

Pentaho Company. 2011. Sitio Oficial Pentaho. [En línea] 2011. [Citado el: 4 de mayo de 2011.] [Disponible en http://www.pentaho.com/].

Pérez Hurtado, Alexei y Padilla Moya, Álvaro. 2010. Plataforma de Gestión de Servicios Telemáticos en GNU/Linux. Módulo DNS v2.0. Universidad de las Ciencias Informáticas. La Habana : s.n., 2010.

Sitio Oficial de RapidMiner. 2011. Sitio Oficial de RapidMiner. [En línea] 2011. [Citado el: 20 de abril de 2011.] [Disponible en: http://www.rapidminer.com].

Venner, Jason. 2009. Pro Hadoop. s.l. : Apress, 2009.