Minería de Datos para los Sistemas Gestores de Bases de Datos

Brenda Mariza Quintero Beltrán,

Resumen—Muchas de las decisiones importantes que se toman alrededor del mundo se basan en observaciones y/o eventos que han sido previamente registrados de alguna forma en una base o modelo de datos. Esta información puede llevar a analistas de mercado a tomar decisiones en cuanto a la compra o venta de acciones, a médicos que trabajan en la obtención de técnicas para detectar enfermedades a tiempo, etc.

Palabras Clave—Minería de Datos, Técnicas de Minería de Datos, Almacén de Datos, Algoritmos de Minería de Datos.

1 INTRODUCCIÓN

En la actualidad es un gran reto para las organizaciones manejar grandes volúmenes de información, ya que los datos que se llegan a almacenar pueden contener demasiadas propiedades o atributos que causan que la información sea complicada de visualizar a primera instancia así también las bases de datos pueden llegar a almacenar miles o millones de instancias de datos, las cuales pueden llegar a variar después de cientos o miles de muestras. Esto hace que en ocasiones las organizaciones no sean capaces de utilizar al máximo esta información, pues no la tienen organizada adecuadamente y carecen de los métodos necesarios para procesarla y analizarla de la mejor manera. Debido a lo importante que es extraer el conocimiento guardado en estos datos, ha surgido lo que se conoce como Minería de Datos.

Esta investigación tratará los conceptos y aplicaciones de Minería de Datos, así también se abordarán temas relacionados a los Sistemas Gestores de Bases de Datos comerciales y Libres que cuentan con técnicas de Minería de Datos para el tratamiento de la información.

Los SGBD que analizaremos son: SQL Server, Oracle, MySQL y PostgreSQL.

• *B.M. Quintero Belrán está con la Escuela de Informática, Universidad Autónoma de Sinaloa, Josefa Ortiz de S/N, Ciudad Universitaria, Culiacán, Sinaloa, 80010, México.*
E-mail: lmarizita@hotmail.com

2 CONCEPTOS BASICOS

Para entender mejor de lo que se trata esta investigación, describiremos los conceptos de Minería de Datos, Técnicas de Minería de Datos, Almacen de Datos.

2.1 ¿Que es un Almacen de Datos?

El almacenamiento de datos se define como un proceso de organización de grandes cantidades de datos de diversos tipos ”guardados” en la organización con el objetivo de facilitar la recuperación de la misma con fines analíticos.

El almacenamiento de datos tiene un gran importancia en el proceso de minería de datos pues en cierta medida, permite la recuperación o al menos la referencia a determinados conjuntos de datos de importancia para un proceso de toma de decisión dado. En la actualidad existe gran variedad de sistemas comerciales y libres para el almacenamiento de datos entre los que se destacan Oracle, MS SQL Server, PostgreSQL, MySQL, entre otros.

2.2 Minería de Datos

La minería de datos es la extracción de información implícita, desconocida o previamente ignorada, que puede ser potencialmente útil, de un conjunto de datos. Se puede considerar a la minería de datos como una colección de diferentes técnicas que sirven para inducir el conocimiento e información de una manera estructurada de un gran conjunto de datos.

La minería de datos ayuda a las organizaciones a encontrar información que no es perceptible de forma directa, como por ejemplo patrones de comportamiento, relaciones, asociaciones, etc., que nos permitan tomar mejores decisiones. A través del análisis del pasado, y aplicando algoritmos, se construyen predicciones que nos permiten mejorar nuestra eficiencia y conseguir así una mayor rentabilidad de la actividad de negocio, y también se le relaciona con el descubrimiento del conocimiento en bases de datos conocido como Knowledge Data Discovery (KDD).

2.3 Técnicas de Minería de Datos

Las técnicas de minería de datos se emplean para mejorar el rendimiento de procesos de negocio o industriales en los que se manejan grandes volúmenes de información estructurada y almacenada en bases de datos. Por ejemplo, se usan con éxito en aplicaciones de control de procesos productivos, como herramienta de ayuda a la planificación y a la decisión en marketing, finanzas, etc.

La minería de datos tiene una incidencia en diferentes disciplinas como la estadística, la inteligencia artificial, los aprendizajes de máquina, el reconocimiento de patrones, etc. Ésta se basa en diferentes tipos de técnicas como redes neuronales artificiales, árboles de decisión, algoritmos genéticos, el método del vecino más cercano y las reglas de inducción, entre otras.

3 Minería de Datos SQL Server

SQL Server una plataforma global de base de datos que ofrece administración de datos empresariales con herramientas integradas de inteligencia empresarial (BI). El motor de la base de datos SQL Server es un almacenamiento seguro y confiable tanto para datos relacionales como estructurados, lo que permite crear y administrar aplicaciones de datos altamente disponibles y con mayor rendimiento para utilizarse en diferentes organizaciones.

3.1 Minería de Datos SQL Server 2005

Microsoft SQL Server 2005 ofrece un entorno integrado para crear modelos de minería de datos y trabajar con ellos, este entorno es la tecnología Business Intelligence que permite construir modelos analíticos complejos e integrar esos modelos con las operaciones comerciales en diferentes tipos de negocios, proporcionando acceso continuo a aplicaciones de amplia difusión e informes, dando cobertura a todos los aspectos del proceso de toma de decisiones.

Microsoft SQL Server 2005 incorpora la herramienta SQL Analysis Server (SSAS), la cual facilita la creación de sofisticadas soluciones de procesamiento analítico en línea (OLAP) y minería de datos. Las herramientas de Analysis Services proporcionan la capacidad de diseñar, crear y administrar cubos y modelos de minería de datos de los almacenes de datos, permiten que el cliente pueda obtener acceso a los datos de la minería de datos, así como identificar reglas y patrones en los datos, y así determinar las razones por las que suceden las cosas y predecir lo que puede pasar en el futuro.

Cuando se crea una solución de minería de datos en Analysis Services, primero se crea un modelo que describe el problema y después se procesan los datos mediante un algoritmo que genera un modelo matemático de ellos, un proceso que se conoce como entrenamiento del modelo. A continuación, puede explorar visualmente el modelo de minería de datos o crear consultas de predicción en él. Analysis

Services puede utilizar conjuntos de datos a partir de bases de datos relacionales u OLAP, e incluye una variedad de algoritmos que se pueden usar para analizar estos datos a través de un modelo UDM o directamente a partir de un almacén de datos físico.

Entre las facilidades para realizar Minería de Datos se cuentan:

- El procesamiento de los modelos de una misma estructura de minería ocurre en paralelo, en una sola lectura de los datos.
- Proporciona más de 12 visores de resultados para los algoritmos que ayudarán a comprender mejor los patrones encontrados en el proceso de minería.
- Proporciona gráficos de elevación, de beneficios y una matriz de clasificación que permite establecer una comparación de lo real con lo previsto; para contrastar y comparar la calidad de los modelos.
- Posee un lenguaje para la creación de consultas de minería (DMX) similar al SQL que facilita la tarea de creación de aplicaciones de minería de datos.
- Posee una interfaz gráfica para generar las consultas DMX.
- Cuenta con los algoritmos de minería más avanzados: Naive Bayes, Clustering, Clústeres de Secuencia, Árboles de Decisión, Redes Neuronales, Series Temporales, Reglas de Asociación, Regresión Logística, y Regresión Lineal y minería de textos.

Se pueden usar varias de estas características y herramientas a la vez para detectar las tendencias y los patrones existentes en los datos; después, se pueden usar las tendencias y los patrones para tomar decisiones informadas sobre los problemas empresariales más complicados.

3.2 Algoritmos de minería de datos de Analysis Services

El algoritmo de minería de datos es el mecanismo que crea un modelo de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos y luego busca patrones y tendencias específicos. El algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El modelo de minería de datos que crea un algoritmo puede tomar diversas formas, ncluyendo:

- El procesamiento de los modelos de una misma estructura de minería ocurre en paralelo, en una sola lectura de los datos.
- Proporciona más de 12 visores de resultados para los algoritmos que ayudarán a comprender mejor los patrones encontrados en el proceso de minería.
- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción.
- Un árbol de decisión que predice si un cliente determinado comprará un producto.
- Un modelo matemático que predice las ventas.
- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.

Microsoft SQL Server Microsoft SQL Server Analysis Services proporciona varios algoritmos que puede usar en las soluciones de minería de datos. Estos algoritmos son un subconjunto de todos los algoritmos que pueden utilizarse en la minería de datos.

Tipos de algoritmos de minería de datos

Analysis Services incluye los siguientes tipos de algoritmos:

- Algoritmos de clasificación, que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos. Un ejemplo de algoritmo de clasificación es el Algoritmo de árboles de decisión de Microsoft. El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporciona por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos.

  Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

- Algoritmos de regresión, que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos. Un ejemplo de algoritmo de regresión es el Algoritmo de serie temporal de Microsoft (Analysis Services - Minería de datos).

  El algoritmo de serie temporal de Microsoft proporciona los algoritmos de regresión que se optimizan para la previsión en el tiempo de valores continuos tales como las ventas de productos. Mientras que otros algoritmos de Microsoft, como por ejemplo los árboles de decisión, requieren columnas adicionales de nueva información como entrada para predecir una tendencia, los modelos de serie temporal no las necesitan. Un modelo de serie temporal puede predecir tendencias basadas únicamente en el conjunto de datos original utilizado para crear el modelo. Es posible también agregar nuevos datos al modelo al realizar una predicción e incorporar automáticamente los nuevos datos en el análisis de tendencias.

- Algoritmos de segmentación, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares. Un ejemplo de algoritmo de segmentación es el Algoritmo de clústeres de Microsoft (Analysis Services - Minería de datos).

  El algoritmo de clústeres de Microsoft es un algoritmo de segmentación suministrado por SQL Server 2008 Analysis Services (SSAS). El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

  Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual.

- Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden utilizarse en un análisis de la cesta de compra. Un ejemplo de algoritmo de asociación es el Algoritmo de asociación de Microsoft.

  Este algoritmo de Microsoft es un algoritmo de asociación suministrado por Analysis Services, útil para los motores de recomendación. Un motor de recomendación recomienda productos a los clientes basándose en los elementos que ya han adquirido o en los que tienen

interés. El algoritmo de asociación de Microsoft también resulta útil para el análisis de la cesta de compra.

Los modelos de asociación se generan basándose en conjuntos de datos que contienen identificadores para casos individuales y para los elementos que contienen los casos. Un grupo de elementos de un caso se denomina un conjunto de elementos. Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen cómo estos elementos se agrupan dentro de los casos. Las reglas que el algoritmo identifica pueden utilizarse para predecir las probables compras de un cliente en el futuro, basándose en los elementos existentes en la cesta de compra actual del cliente.

- Algoritmos de análisis de secuencias, que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas Web. Un ejemplo de algoritmo de análisis de secuencias es el Algoritmo de agrupación en clústeres de secuencia de Microsoft.

  El algoritmo de clústeres de secuencia de Microsoft es un algoritmo de análisis de secuencias que proporciona Microsoft SQL Server Analysis Services. Puede utilizar este algoritmo para explorar los datos que contienen eventos que pueden vincularse mediante rutas o secuencias. El algoritmo encuentra las secuencias más comunes mediante la agrupación, o agrupación en clústeres, de las secuencias que son idénticas.

Además de los algoritmos anteriormente mencionados existen los algoritmos proporcionados por terceras partes.

SQL Server Data Mining permite la agregación de algoritmos. Sin embargo se restringe a que los algoritmos de terceras partes desarrolladoras puedan soportarse en términos de lenguaje y tipos de datos, deberá ser de manera gratuita, también deberán permitir la integración con la herramienta Analysis Services, incluyendo la capacidad de construir Mineria OLAP y dimensiones de minería de datos. Se usa el término Plug-in Algoritmos para describir los algoritmos de terceras partes, que están incluidas en Microsoft Analysis Services.

4 Minería de Datos Oracle (Oracle Data Mining)

Oracle Data Mining es una opción de Oracle Corporation 's base de datos relacional Management System (RDBMS) Enterprise Edition (EE). Contiene una serie de minería de datos de análisis de datos y algoritmos de clasificación, predicción, regresión, las agrupaciones, asociaciones, característica de la selección, la detección de anomalías, extracción de características, y análisis especializados. Proporciona los medios para la creación, gestión y despliegue operacional de los modelos de minería de datos en el interior de la base de datos de medio ambiente.

La base de datos Oracle incluye funcionalidad para la minería de datos en la edición Enterprise. Esta funcionalidad está totalmente integrada y bajo el mismo motor que la parte relacional de la misma. Se puede acceder a toda la funcionalidad Data Mining a través de la API Java que incluye la base de datos, de manera que las aplicaciones puedan sacar el máximo partido de las funciones disponibles.

Al estar integrado en la base de datos, Oracle Data Mining simplifica el proceso de extracción de conclusiones basadas en grandes cantidades de datos, ya que se elimina la necesidad de movimientos de datos para el proceso de análisis. Todas las operaciones de preparación, creación de modelos y análisis permanecen en la base de datos lo que resulta en una mejora de la productividad, automatización e integración.

Oracle Data Mining acepta tablas transaccionales y no transaccionales

(resúmenes, registros únicos). Oracle Data Mining hace todas las transformaciones necesarias automáticamente de forma interna, liberando así de este trabajo a los usuarios o desarrolladores.

Oracle Data Mining soporta la clasificación de valores dentro de un campo en grupos que tengan sentido, por ejemplo, el campo edad puede ser clasificado en rangos como: 0-16,17-21, etc.

4.1 Algoritmos Oracle Data Mining

Oracle proporciona dos algoritmos para dos tipos diferentes de análisis: 1) Naive Bayes para clasificaciones y predicciones 2) Reglas de asociación para encontrar patrones.

- Naive Bayes: Naive Bayes es una técnica de clasificación y predicción que construye modelos que predice la probabilidad de posibles resultados. Naive Bayes utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones. Este algoritmo predice resultados binarios o multiclase. En los problemas binarios, cada registro cumplirá o no el comportamiento modelado. Por ejemplo, se puede construir un modelo para averiguar si un cliente será fiel o cambiará de proveedor.

  Naive Bayes puede hacer predicciones para problemas multiclase, en los cuales hay varios resultados posibles. Por ejemplo, se puede construir un modelo para predecir qué clase de servicio prefiere cada cliente.

- Reglas de Asociación.-Las reglas de asociación detectan eventos asociados que se ocultan en las bases de datos. Este tipo de análisis a menudo se utiliza para encontrar combinaciones populares de productos, tales como cereales y leche asociados con plátanos. Las reglas de asociación generan un conjunto de pares A-B con una probabilidad n

  Una vez creados los modelos Naive Bayes, los registros de datos pueden ser puntuados. La puntuación es el proceso de predicción de resultados, y puede hacerse en modo batch o bajo demanda. En modo batch el algoritmo recorre una tabla y va almacenando las predicciones en otra tabla, bajo demanda el algoritmo puntúa un solo registro y d devuelve la predicción, que puede utilizarse directamente en la aplicación que haya pedido esta puntuación.

4.2 Herramientas Open Source para Oracle.

Las nuevas herramientas de terceros, optimizadas para Oracle data warehouses, hacen que sea más efectiva la gestión de grandes volúmenes de datos, la realización de análisis complejos, la minería de datos y visualización, así como la realización de múltiples trabajos simultáneos y el establecimiento de crecientes comunidades de usuarios. Las más destacadas son:

- Visual data mining. Mineset 3.2 de Vero Insight, un conjunto integrado de herramientas visuales de minería de datos que revela el valor escondido que se encuentra en los datos, tendencias, patrones y relaciones. Mineset proporciona información compleja en gráficos fácilmente comprensibles, facilitando el análisis en tiempo real con parámetros que pueden ser ajustados de manera improvisada.

- Open source Business Intelligence. Pentaho Open BI Suite de Pentaho Corporation proporciona análisis OLAP, paneles de control, minería de datos e integración de datos. Pentaho ofrece integración de datos en tiempo real para mandos de control operacionales, así como envío de información BI a través del iPhone de Apple.

5 Herramientas y Suites de Minería de Datos Para PostgreSQL y MySQL

En la actualidad se han desarrollado diversos sistemas que ofrecen soluciones a los problemas de Minería de Datos, y que implementan todas o algunas de sus tecnologías.

5.1 Herramientas simples

Las herramientas simples son aquellas que dan soporte a las tecnologías de la inteligencia empresarial individualmente. Estas herramientas han sido clasificadas a su vez de acuerdo a las tecnologías de inteligencia empresarial a la que dan soporte, en seguida mencionaremos las más importantes y populares para cada una.

TariyKDD

Puede definirse como una herramienta genérica de tareas múltiples débilmente acoplada a un SGBD. TariyKDD comprende cuatro módulos que cubren el módulo de conexión, tanto a archivos planos como a bases de datos, un módulo de utilidades con clases y librerías comunes a toda la aplicación, un módulo kernel que incluye las tareas de preprocesamiento, minería y visualización y el módulo de la interfaz gráfica de usuario. Dentro del kernel de minería se implementaron los algoritmos EquipAsso, un algoritmo para el cálculo de conjuntos de ítems frecuentes, y Mate, un algoritmo para la construcción de arboles de clasificación, basados en nuevos operadores del álgebra relacional, propuestos por Ricardo Timarán Pereira, PhD, director del Grupo de Investigación Aplicado a Sistemas (GRiAS) de la Universidad de Nariño.En este proyecto se implementaron además los algoritmos Apriori y FPGrowth para asociación y C4.5 para clasificación.

En TariyKDD se trabajó una conexión a bases de datos relacionales a través de drivers JDBC tipo 4 y se han hecho pruebas con PostgreSQL (principalmente), MySQL y Oracle 10g. Los Drivers JDBC hacia estos gestores se distribuyen con la herramienta. Se aprovechó las nuevas características de estos drivers para desplegar una interfaz visual para la selección de atributos y facilitar la construcción del conjunto de datos que será el objeto de análisis de los algoritmos de minería.

Igualmente se incluyeron rutinas para la predicción de nuevos registros a partir de los modelos construidos con los algoritmos de clasificación.

Weka

Del inglés Waikato Environment for Knowledge Analysis: Es una colección de algoritmos de aprendizaje por computadora o ML (del inglés Machine Learning) para realizar tareas de Minería de Datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o dataset desde la interfaz gráfica del programa (Java Swing), mandándolos llamar desde el shell o utilizar los códigos independientes que se proporcionan mandándolos llamar desde nuestro File Format), de instancias binarias serializadas, de archivos C45, de archivos separados por comas (*.csv), de archivos separados por tabulaciones utilizando el conversor que nos ofrece, de una URL o de una base de datos PostgreSQL, MySQL. Su última versión, la 3.4.10, es del 25 de enero de 2007. Weka es el proyecto de este tipo más antiguo (se inició alrededor del año 1993) y es de los más difundidosa la fecha.

Este paquete se distribuye como un archivo comprimido, en donde se incluye la documentación, un tutorial, el log de cambios, el icono, los códigos fuente, entre otras muchas cosas.

YALE

Del inglés Yet Another Learning Environment: Es una colección de operadores de Miner-ía de Datos (más de 400) desarrollada en Java que integra completamente los códigos de Weka y que nos permite realizar ML y

Minería de Datos. Cuenta con una interfaz gráfica fácil de utilizar, pero también puede ser utilizado desde el shell o como una librería dentro de tus propios programas en Java.

YALE cuenta con un mecanismo sencillo para desarrollar extensiones y plugins que hace posible integrar nuevos operadores y con ello adaptar el paquete para los requerimientos personales. Existen plugins ya desarrollados disponibles en la página de YALE: Clustering Plugin, Word Vector Tool Plugin, Value Series Plugin, Distributed Data Mining Plugin, Data Stream Plugin y Selurtib8 Plugin (un juego de los 80s).

Este paquete tiene dos tipos de licencia, la gratuita bajo la licencia GPL y la propietaria, para cuando se desea modificar YALE y distribuirlo sin proporcionar los códigos fuente, desarrollar módulos comerciales extra, proporcionar servicio al cliente, soluciones individuales o soporte profesional.

YALE puede importar información a partir de distintos formatos de archivos como ARFF, csv, C45, SVMLight, mySVM, Excel, SPSS, etc., archivos de texto (a través del Word Vector Plugin), archivos de audio (a través del Value Series Plugin) o a partir de sistemas de bases de datos como Oracle, MySQL, PostgreSQL, Microsoft SQL Server, etc., e inclusive es capaz de trabajar directamente sobre una base de datos.

Intelligent Miner (IBM)

Intelligent Miner convierte información desestructurada en business knowledge para empresas de cualquier tamaño. Incluye componentes para crear aplicaciones avanzadas de text-mining y de text-search. Ofrece a integradores de sistema, proveedores de soluciones y desarrolladores de aplicaciones una gran cantidad de herramientas de análisis de texto, de componentes de recuperación de texto y de acceso a la web para aumentar en capacidades las herramientas de Business Intelligence y de gestión del conocimiento. Con Intelligent Miner se puede desbloquear la información atrapada en mensajes electrónicos, reclamaciones, noticias, Lotus Notes; analizar portafolios de patentes, cartas de reclamaciones de clientes e incluso páginas web de competidores.

DAT-A: Minería de Datos y OLAP en MySQL

DAT-A es una aplicación de código libre que se construye para permitir la minería de datos inteligente. Para la Minería de Datos Inteligente DAT-A ,los arquitectos de software están creando una aplicación que se centra en la atención del usuario sobre los resultados de minería de datos y no a la extracción de datos o proceso de modelado de datos. Todos los intercambios de datos se encuentran en XML y SOAP para garantizar la interoperabilidad.

Actualmente, MySQL no tiene módulos para construir minería de datos. DAT-A aplica una capa de abstracción de Minería de Datos sobre MySQL. La lógica de negocio para el control del modelo de Minería de Datos y el modelo de formación están escritos en el framawork J2EE.

Para la edición personal de DAT-A, el servidor de aplicaciones de minería de datos de MySQL está contenido dentro de la lógica de negocio desarrollado en la capa de framework J2EE. En versión empresarial, la lógica de negocio y la extracción de datos de control se encuentran en el servidor BEA WebLogic Application Server.

Clementine

Es uno de los sistemas de data mining más populares de mercado. Se trata de una herramienta visual inicialmente desarrollada por ISL (Integral Solutions Limited). En la actualidad esta herramienta, comercializada por SPSS, posee una arquitectura distribuida (cliente/servidor).

Sus características principales son:

- Acceso a datos: fuentes de datos ODBC, tablas Excel, archivos planos ASCII y

archivos SPSS.

- Preprocesado de datos: pick y mix, muestreo, particiones, reordenación de campos, nuevas estrategias para la fusión de tablas, etc.

- Técnicas de aprendizaje: árboles de decisión, redes neuronales, agrupamiento, reglas de asociación, regresión lineal y logística, combinación de modelos.

- Técnicas para la evaluación de modelos guiadas por las condiciones especificadas por el experto.
- Visualización de resultados: ofrece un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen final del modelo aprendido.

- Exportación: generación automática de informes (HTML y texto), volcado de los resultados del ejercicio de data mining en bases de datos, exportación de los modelos a distintos lenguajes (C, SPSS, HTML, estándar PMML, SQL para árboles de decisión y reglas).

5.2 Suites

Pentaho BI

Pentaho BI [Pentaho, 2007] es la única suite que existe de código abierto, ya que ha adoptado las herramientas de Kettle, Mondrian y WEKA para conformar una sola plataforma de inteligencia empresarial.

Tiene las funcionalidades de reportes, análisis, data mining e integración de datos. Los proyectos involucrados en Pentaho BI son:

- Mondrian: servidor de OLAP

- JFreeReport: reporteador.

- Kettle: integración de datos (ETL).

- Pentaho: plataforma de inteligencia empresarial.

- WEKA: data mining.

  Permite a los desarrolladores de Java diseñar componentes que pueden ser rápidamente ensamblados en soluciones de inteligencia empresarial y a los usuarios finales desplegar rápidamente las soluciones existentes de inteligencia empresarial.

  A pesar de que Pentaho BI se conforma ya como una suite que satisface las necesidades de inteligencia empresarial, sus componentes no se integran perfectamente ya que surgieron de manera individual. Sin embargo la compañía Pentaho ofrece soporte comercial para todas ellas como conjunto, por eso se considera una suite.

6 CONCLUSIONES

Para concluir, vale la pena mencionar que la minería de datos provee una gran utilidad a cualquier persona que tenga como fuente un conjunto de datos bien estructurado, organizado y que esté almacenado en una base de datos. Mientras mayor número de información se tenga para trabajar, mejores resultados proveerá la minería de datos.

En todo el proceso de Minería de Datos, el ser humano es el factor más importante, ya que solo el tiene la capacidad de analizar y decidir si los patrones encontrados tienen importancia para su empresa.

Las herramientas para Minería de Datos que existen actualmente en el mercado son muy variadas y excelentes en diversas aplicaciones, así también existen herramientas de licencia libre que también son muy útiles para realizar minería de datos en los distintos Sistemas Gestores de Bases de Datos.

REFERENCES

[1] Craig Utley, *(Introduction to SQL Server 2005 Data Mining) Microsoft SQL Server 2005 Aplica Microsoft Data mining* LaTeX, junio 2005

[2] Raman Iyer and Bogdan Crivat, *SQL Server Data Mining: Plug-In Algorithms* LaTeX, julio 2005 Applies Microsoft SQL Server 2005 Analysis Services

[3] Erika Vilches González e Iván A. Escobar Broitman, *Minería de Datos* LaTeX, septiembre 2007

[4] Mar Montalvo , *Oracle Business Intelligence* LaTeX, Marzo 2007

[5] José Echegaray, *Oracle Data Mining* LaTeX, Marzo 2007

[6] Olivier MOREL DES VALLONS Alexandre PERTRIAUX, *(DATA MINING TOOL FOR MYSQL DATABASES)*LaTeX, 20 Diciembre 2005