

Multiagentes Distribuidos Programables Vía Web para Minería de Datos

Lic. Karina Mino Polanco
kmino10@hotmail.com

M.C. Rogelio Ferreira Escutia
rferreir@itmorelia.edu.mx

Maestría en Ciencias en Ciencias de la Computación
Instituto Tecnológico de Morelia, Av. Tecnológico 1500, Col. Lomas de Santiaguito, C.P. 58120
Morelia, Michoacán, México

Resumen.- En este trabajo se muestra la fase de diseño de un proyecto para implementar multiagentes distribuidos que utilizan algoritmos de Minería de Datos Descriptiva para encontrar información útil en red y que son programables via web para permitir la visualización de resultados y la reconfiguración de los parámetros de minería de datos a través de Internet o cualquier dispositivo móvil (como celulares y asistentes personales), de tal manera que el proceso de minería de datos sea más automático, flexible y rápido que el proceso tradicional.

I. INTRODUCCIÓN

La automatización de los procesos manuales de las empresas ha propiciado la generación de grandes bases de datos difíciles de manejar por los sistemas manejadores de bases de datos actuales. Además, ya que se cuenta con éstas bases de datos ¿porqué no obtener nueva información a partir de la existente?.

La minería de datos es una tecnología reciente, debido al crecimiento de las Base de Datos, que permite el análisis de información para la toma de decisiones, además del planteamiento y descubrimiento automático de hechos e hipótesis, ya sean patrones, reglas, grupos, funciones, modelos, secuencias, relaciones, correlaciones, etc.

Esto permite a las empresas la posibilidad de obtener información para la toma de decisiones de sus grandes bases de datos, información que ordinariamente no se obtendría sin utilizar técnicas de minería de datos, haciéndolas más competitivas en el mercado.

En la actualidad, existe una gran cantidad de aplicaciones en las que se utiliza la minería de datos, en áreas tales como:

- a) Astronomía: clustering y clasificación de cuerpos celestes.
- b) Biología molecular: predicción de sustancias cancerígenas, genoma humano, etc.
- c) Aspectos climatológicos: predicción de tormentas, etc.
- d) Medicina: caracterización y predicción de enfermedades.
- e) Industria y manufactura: diagnóstico de fallas.
- f) Mercadotecnia: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, selección de sitios de tiendas, etc.
- g) Inversión en casas de bolsa y banca: análisis de clientes, aprobación de préstamos, etc.
- h) Detección de fraudes y comportamientos inusuales: telefónicos, seguros, electricidad, etc.
- i) Análisis de canastas: análisis de canastas de mercado para mejorar la organización de tiendas.

II. MINERÍA DE DATOS

La Minería de Datos, también conocido en inglés como **Data Mining**, es un proceso que, a través del descubrimiento y cuantificación de relaciones predictivas en los datos, permite transformar la información disponible en conocimiento útil de negocio. Esto es debido a que no es suficiente "navegar" por los datos para resolver los problemas de negocio, sino que se hace necesario seguir una metodología ordenada que permita obtener rendimientos tangibles de este conjunto de herramientas y técnicas de las que dispone el usuario.

Existen dos tipos de Minería de Datos:

1. *Minería Descriptiva*.- Describe los datos en forma concisa y presenta en forma general propiedades interesantes de los datos. Se aplica para descubrir patrones en datos existentes para guiar el proceso de toma de decisiones.
2. *Minería Predictiva*.- Construye un conjunto de modelos, aplicando inferencia sobre los datos disponibles, haciendo intentos para predecir el comportamiento de nuevos conjuntos de datos.

III. AGENTES

Los agentes son sistemas que cuentan con las siguientes propiedades [1]:

1. Movilidad.- Habilidad del moverse en forma autónoma.
2. Adaptabilidad.- Puede ajustarse a los métodos, hábitos de trabajo y preferencias del usuario.
3. Colaboración.- Permite corregir algunos errores causados por su usuario, como omisión de información, información ambigua, etc.

Los Agentes utilizados en la Minería de Datos se encargan de analizar la información para detectar patrones y relaciones, ya de forma automática o interactuando con el usuario analista [2]. Gracias a la flexibilidad, generalidad y posibilidad de distribución y modularidad que poseen los agentes, es posible pensar en construir diferentes tipos agentes que proporcionen una solución eficiente al proceso de minería de datos.

IV. CONSIDERACIONES BÁSICAS

Actualmente, la tendencia de las empresas es tener la mayor cantidad de información útil en bases de datos, por lo cual éstas pueden contener una gran cantidad de registros. La Minería de Datos se utiliza generalmente sobre bases de datos históricas gigantescas (contienen información de hace semanas, meses, años, etc.) para obtener información importante para la toma de decisiones que por otros métodos convencionales no se encuentra dado el gran volumen de información.

El problema de realizar la minería de datos sobre bases de datos históricas es que los resultados obtenidos reflejan relaciones pasadas, dado que son de datos de tiempo atrás. Los resultados de una minería de datos son más relevantes y oportunos para una empresa si se realiza la *minería sobre datos actuales* (de hoy, de ayer, de hace pocos días, etc.), de esta manera la toma de decisiones se puede basar sobre información que describe la situación actual de la empresa y no la pasada.

Existe una gran cantidad de herramientas para realizar minería de datos. Muchas de las cuales cubren la mayoría de las técnicas de minería, y otras sólo algunas. Estas herramientas en la mayoría de los casos no son de dominio público por lo que se deben de comprar a precios altos.

Otro problema consiste en que el proceso de Minería de datos se desarrolla en forma "manual", es decir el usuario tiene que iniciar el proceso y ver los resultados desde la máquina donde reside el minero, lo cual involucra una dependencia total del minero hacia el usuario.

Por lo tanto, se propone el diseño e implementación de multiagentes distribuidos para automatizar y agilizar el proceso de Minería de datos, los cuales puedan mostrar

resultados y ser programados desde donde se encuentre el usuario a través de una página web, un celular o un asistente personal.

V. ARQUITECTURA PROPUESTA

El proyecto contempla el diseño e implementación de multiagentes distribuidos programables vía web para minería de datos descriptiva para una aplicación de mercadotecnia. Se programarán los agentes en Lenguaje Java para hacerlos multiplataforma [3]. El programa de comunicación usuario-agente coordinador se programará en J2ME.

Para la realización de minería de datos descriptiva se utilizarán el enfoque de Asociación de Reglas, para una aplicación de "análisis de canasta" de una tienda comercial, es decir, se analizarán las compras de artículos de los clientes.

A. Aplicación

La aplicación de minería de datos es un "análisis de canasta". El análisis de canasta de compras toma su nombre del análisis que se hace de los artículos que un comprador lleva en el momento de la compra en el supermercado o tienda comercial en busca de afinidades para mejorar las ofertas o también para desarrollar las promociones adecuadas [1].

Sin embargo, este análisis se puede realizar en cualquier situación en donde se requiera identificar el tipo de productos o servicios que una persona consume o utiliza. A partir de la identificación de las afinidades se pueden identificar las oportunidades para venta cruzada (venta de artículos de manera conjunta), o para elevar el valor promedio de compra.

Entre los principales beneficios que una tienda comercial obtiene al realizar un análisis de canasta, están:

- Conocimiento del patrón de consumo de sus clientes
- Identificar oportunidades para la venta cruzada
- Desarrollo de ofertas más atractivas
- Ubicación de los productos dentro de la tienda

Para llevar a cabo el análisis, se deben realizar un proceso de minería de datos para encontrar las relaciones entre los artículos comprados en determinado espacio y tiempo.

B. Minería de datos descriptiva

La minería de datos descriptiva se aplica para describir patrones en datos existentes para guiar el proceso de toma de decisiones.

Para la realización de este tipo de minería, existen diversos algoritmos [4]:

- Asociación de reglas
- Regresión
- Clasificación

- **Agrupamiento o Clustering** La asociación de reglas utiliza reglas para definir patrones de comportamiento de los datos. La regresión utiliza técnicas estadísticas para encontrar la relación lineal entre los datos. La clasificación utilizan clases para describir los patrones de comportamiento y el agrupamiento forma grupos mutuamente excluyentes con los datos.

El algoritmo a implementar es la asociación de reglas, debido a que es el que se adapta mejor al análisis de canasta [4] que los otros algoritmos por la facilidad de encontrar en los datos reglas para identificar los patrones de comportamiento de las compras de los clientes. Por ejemplo, la regla de asociación $Compra_Cerveza \Rightarrow Compra_Pañales$ [30% confianza, 2% soporte] indica que “el 30% de las transacciones que contienen cerveza también contienen pañales; el 2% de todas las transacciones contienen ambos artículos”.

Para cada regla de asociación existen dos conceptos conocidos como el *soporte* y la *confianza*. El *soporte* se define como la probabilidad de que un registro satisfaga tanto a X como a Y . La *confianza* se define como la probabilidad de que un registro satisfaga a Y dado que satisface a X . Para el ejemplo anterior, 30% es el nivel de confianza, y el 2% es el *soporte* de la regla. Por lo tanto, la regla $X \Rightarrow Y$ tiene el *soporte* s en el conjunto de transacciones D si el $s\%$ de las transacciones en D contienen $X \cup Y$, y la regla $X \Rightarrow Y$ en el conjunto de transacciones D tiene *confianza* c si el $c\%$ de las transacciones en D que contienen X también contienen Y .

C. Multiagentes distribuidos

Los agentes deben ser capaces de realizar el proceso de minería de manera automática, es decir, arrancar el proceso a determinado tiempo de acuerdo con parámetros establecidos (que pueden ser modificados vía web). El ambiente de trabajo entre los agentes (Fig. 1) se especifica como una red de computadoras con n sitios sobre las cuales se encuentran bases de datos grandes con una estructura fija y son dinámicas en el sentido de que su contenido va variando en tiempo real (debido a la utilización propia de la base de datos como inserciones, eliminaciones y modificaciones), existirán n agentes de procesamiento que se encarguen de analizar la información minada de cada sitio y que es obtenida por un agente de monitoreo capaz de detectar cambios en la base de datos y realizar la extracción modificación de éstos.

Cada agente de procesamiento debe comunicar sus resultados parciales al agente coordinador el cual los pasa al agente recolector de resultados para que procese el resultado final de minería. El resultado final es enviado al agente de comunicación con el usuario, el cual se encarga de la visualización de los resultados y permite al usuario especificar los parámetros de minería que desee establecer para un nuevo análisis; éstos nuevos parámetros son enviados al agente coordinador para que redefina el proceso de minería de los agentes de procesamiento en los n sitios.

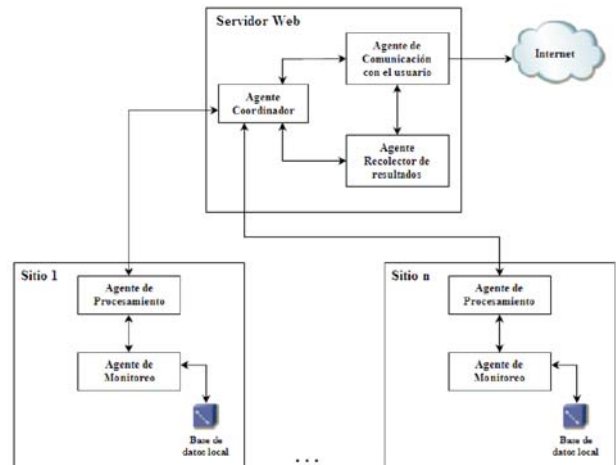


Fig. 1 – Arquitectura de comunicación.

La coordinación del sistema completo está a cargo del agente coordinador el cual coordina todo el proceso de minería de datos y la comunicación entre los diversos agentes. Al dividir el trabajo entre los agentes existentes en cada sitio el proceso de minería de datos se agiliza. En resumen, las funciones que cada agente debe realizar se ilustran en la Tabla 1.

Tabla 1
Funciones generales de los distintos agentes del sistema.

Agente	Funciones
Agente de monitoreo	Verifica constantemente si hubo cambios significativos en la base de datos local del sitio para reportarlos al agente de procesamiento.
Agente de procesamiento	Con los datos obtenidos de la base de datos, aplica un proceso de minería de datos descriptiva para generar un resultado, el cual es enviado al agente recolector de resultados.
Agente recolector de resultados	Recolecta los resultados de los n sitios para generar un resultado general, el cual es enviado al agente de comunicación según lo solicite éste. Para recibir todos los resultados, el agente recolector solicita al agente coordinador que establezca la comunicación con los n sitios y le envíe los resultados, de manera que el agente coordinador funciona como un agente de comunicación entre los demás agentes.
Agente de comunicación con el usuario	Establece la comunicación con el usuario vía web para enviarle a éste los resultados del proceso de minería de datos descriptiva, además de permitirle especificar los nuevos parámetros de minería de datos para realizar un nuevo análisis, los nuevos parámetros son enviados al agente coordinador para que este los comunique a los agentes de procesamiento de cada sitio.
Agente coordinador	Permite la comunicación entre los diferentes agentes del sistema, por lo que es el responsable de comunicar a los n sitios los cambios en los parámetros del proceso de minería de datos y de coordinar la recopilación de resultados de cada sitio.

Obviamente, se requiere una interfaz para la comunicación usuario-agente que permita programar al agente: especificar los parámetros de minería de datos, patrones que deseen encontrarse, visualización del conocimiento descubierto, entre otras cosas.

La importancia de esta interfaz de comunicación es que permita al usuario comunicarse con el agente principal desde cualquier parte donde éste se encuentre, por lo cual la interfaz se programará en html (para la página web) y en J2ME (para los dispositivos móviles) para que el agente principal pueda ser programado dinámicamente y mostrar resultados a través de paginas web y dispositivos móviles como celulares y asistentes personales (PDA). La Fig. 2 muestra esta arquitectura de comunicación con el agente principal.

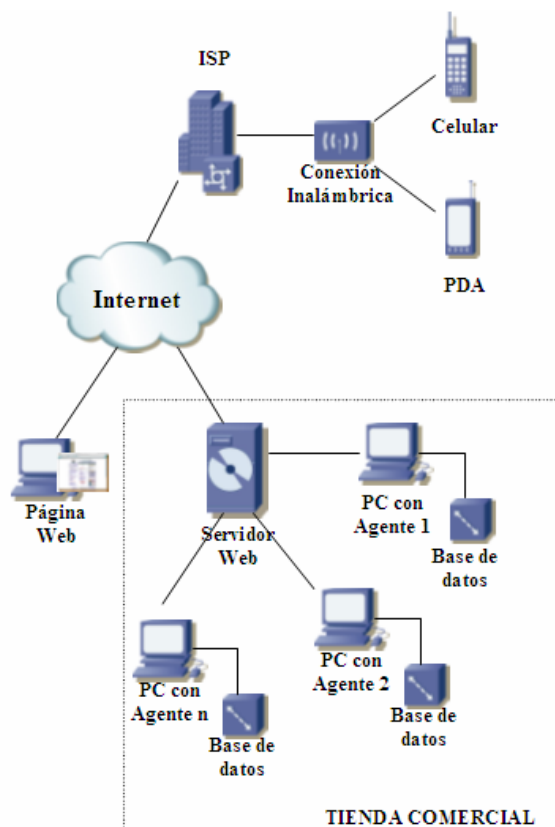


Fig. 2 - Arquitectura de red.

VI. CONCLUSIONES Y TRABAJO FUTURO

El presente proyecto aportará las bases de una nueva forma de realizar el proceso de minería de datos, de manera que los usuarios se vean beneficiados al contar con un proceso de minería de datos dinámico y flexible que obtenga mejores resultados y de manera más rápida.

VII. REFERENCIAS

- [1] Jennings, Nicholas R.. "Agent Technology: Foundations, Applications, and Markets." 1998, Hardcover.
- [2] Liu, Jiming. "Autonomous Agents and Multi-Agent Systems: Explorations in Learning, Self-Organization and Adaptive Computation." 2001, Hardcover.

- [3] Bigus, Joseph P.; Bigus, Jennifer; Bigus, Joe; "Constructing Intelligent Agents Using Java: Professional Developer's Guide." 2000, Paperback.
- [4] Jiawei, Han; Kamber, Micheline; Kaufmann, Morgan. "Data Mining: Concepts and Techniques." 2001, Publishers.
- [5] Thuraisingham B.. "Data Mining, Technologies, Techniques, Tools, and Trends." 1999, CRC press.
- [6] Written, Ian H., Eibe Frank., "Data Mining: Practical Machine Learning Tools with Java Implementations." 2000, Publishers.
- [7] Jean-Marc, Adamo. "Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms." 2001, Hardcover.
- [8] Linoff, Gordon, Berry, Michael J. A. "Data Mining Techniques : For Marketing, Sales, and Customer Support." 2000, Paperback.
- [8] Bradshaw, Jeffrey M.. "Software Agents." 2000, Paperback.
- [9] Attoui, Ammar. "Real-Time and Multi-Agent Systems (Practitioner Series)." 2000, Paperback

VII. BIOGRAFÍAS



Karina Mino Polanco, nació en el Distrito Federal en 1978. Obtuvo el grado de Licenciada en Informática en el Instituto Tecnológico de Tepic en el año 2001. Actualmente es candidata a obtener el grado de Maestro en Ciencias en Ciencias de la Computación con especialidad en sistemas distribuidos en el Instituto Tecnológico de Morelia.



Rogelio Ferreira Escutia, nació en Morelia, Michoacán en 1971. Obtuvo el grado de Ingeniero en Electrónica en el Instituto Tecnológico de Morelia en 1995 y el grado de Maestro en Ciencias de la Computación en Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Morelos en 1998. Actualmente es Profesor del Departamento de Sistemas y Computación y de la Maestría en Ciencias en Ciencias de la Computación en el Instituto Tecnológico de Morelia.