SELECCIÓN DE UNA TÉCNICA DE MINERÍA DE DATOS PARA LA CORRELACIÓN DE PRODUCTOS EN EL COMERCIO ELECTRÓNICO TIPO B2C

AUTOR
Robinson Montenegro Caicedo
Estudiante Programa Ingeniería de Sistemas
Universidad del Cauca
rmontenegro@unicauca.edu.co
COLOMBIA

AUTOR
Roberto Carlos Naranjo
Docente Programa Ingeniería
Universidad del Cauca
rnaranjo@unicauca.edu.co
COLOMBIA

Fecha de recepción del artículo: 12 de Abril de 2007 ***Fecha de aceptación del artículo: 14 de Mayo de 2007***
Articulo Tipo 1

RESUMEN.

La minería de datos es una poderosa tecnología que tiene como objetivo extraer conocimiento útil a partir de grandes cantidades de datos, esta se utiliza en diferentes sectores como el comercio electrónico, la banca, telecomunicaciones entre otros. Actualmente la minería de datos se está utilizando con mayor frecuencia en diversas organizaciones, y a nivel académico en los últimos años, se han venido realizando con mayor regularidad trabajos relacionados que involucran investigación, uso y desarrollo de herramientas que realizan análisis de información a través de las técnicas de minería de datos.

En este trabajo se pretende abordar la investigación de diferentes técnicas de minería de datos para solución de un problema en el entorno del comercio electrónico de tipo B2C. Para la selección de la técnica de minería de datos se debe tener en cuenta una serie de criterios de acuerdo al entorno del problema, además de ello, hay diferentes algoritmos que implementan la técnica, por lo cual es necesario investigarlos y seleccionar el que mejor se adecue a los requerimientos y necesidades desde la perspectiva del problema que se quiere solucionar.

PALABRAS CLAVE

Minería de datos
KDD
Técnicas de Minería de Datos
Comercio electrónico.

ABSTRACT

The Data Mining is a powerful technology that has as objective to extract useful knowledge from great amounts of data, this is used in different sectors as the e-commerce, the bank, the telecommunications among others, and for the extraction of this knowledge is needed a complex process which uses a series of techniques for the processing of the data. At the moment the mining of data is being used most frequently in diverse organizations and at academic level, in the last years, has come making with greater regularity related works that involve investigation, use and development of tools that make analysis of information through the techniques of data mining. In this work it is tried to embroider the investigation of different techniques from data mining for solution of a problem in the surroundings of the e-commerce of type B2C. For the selection of the technique of data mining are due to consider a series of criteria according to the surroundings of the problem, in addition to it there are different algorithms that implement a

technique of data mining, thus is necessary to investigate them and to select the one that adapts better to the requirements and necessities from the perspective of the problem that need to be solved.

KEYWORDS

Data Mining
KDD
Data Mining Techniques
Electronic commerce.

INTRODUCCIÓN

Para obtener conocimiento es necesario partir de la materia prima, que son los datos, los cuales se encuentran disponibles en gran cantidad gracias a las tecnologías de la información y las comunicaciones. Estos datos por lo general se encuentran en forma no refinada y para poder analizarlos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos[1]. Para realizar un análisis en profundidad de forma automática, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos, sin embargo, es la transformación de los datos en conocimiento y la aplicación de este lo que genera valor para una organización[1]. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

Para las organizaciones que realizan algún tipo de mercadeo, el conocimiento es algo imprescindible para tener éxito, por tanto encontrar asociaciones o correlaciones interesantes en los registros de las transacciones de negocios puede ayudar a la toma de decisiones en los procesos de marketing [1], por lo cual invierten recursos técnicos y económicos para la construcción o compra de herramientas para el análisis de información. Es en este entorno la minería de datos ofrece la posibilidad de llevar a cabo un proceso de descubrimiento de información automático [1], esto no es ajeno para las empresas del departamento del cauca que hacen negocios a través de Internet, sin embargo, la mayoría de estas cuentan con recursos limitados lo que dificulta su acceso a estas herramientas. El proyecto titulado "Módulo de Soporte a toma de decisiones para seleccionar estrategias publicitarias de comercio electrónico B2C" tiene como meta impulsar a las empresas del departamento al acercar una herramienta que brinde soporte a la toma de decisiones que esté enmarcada dentro del comercio electrónico B2C (ver Ítem 3), con lo cual se pretende brindar apoyo para las decisiones de marketing.

Aunque existen algunas de libre distribución como WEKA (Waikato Environment for Knowledge Analysis) [16], Orange (Data Mining Fruitful and Fun) [17] y TANAGRA - (A Free Software for Research and Academia Purposes) [18] entre otras, éstas cubren aspectos generales del proceso de minería de datos, y por lo general están orientadas a propósitos académicos presentando algunas dificultades en su adaptación tales como, modificaciones en el código fuente, tipo de lenguajes implementados, tipo de licencia de uso, formato de entrada de datos, entre otros, lo que dificulta su uso simple por parte de las empresa.

Para brindar una solución acorde a las necesidades de las empresas es necesario entender los objetivos y requerimientos desde la perspectiva de lo que se busca, convirtiendo entonces este conocimiento en la definición de un problema de minería de datos [19], ya que dependiendo del problema de información que se desea solucionar, existe una serie de técnicas que son aplicadas en la solución de diversos problemas.

Para la escogencia de una técnica de minería de datos se debe tener en cuenta una serie de consideraciones tales como el entorno del negocio, las fuentes de datos, tipo de clientes, entre otros, que afectan el desempeño de estas [2]. Además de ello existen diferentes algoritmos que implementan las técnicas de minería de datos y cuya utilización afecta el rendimiento de estas, por lo cual es necesario estudiarlos para determinar cual es el más indicado.

En lo que resta del documento se hablará del proceso de descubrimiento de conocimiento, el concepto de minería de datos, las operaciones de minería de datos, una descripción general de las técnicas de minería de datos mas usadas en el marketing, los criterios de selección de la técnica de minería de datos escogida, el uso de esta en el marketing, los algoritmos que la implementan y los criterios de selección del algoritmo escogido y se finalizará el trabajo con las conclusiones obtenidas.

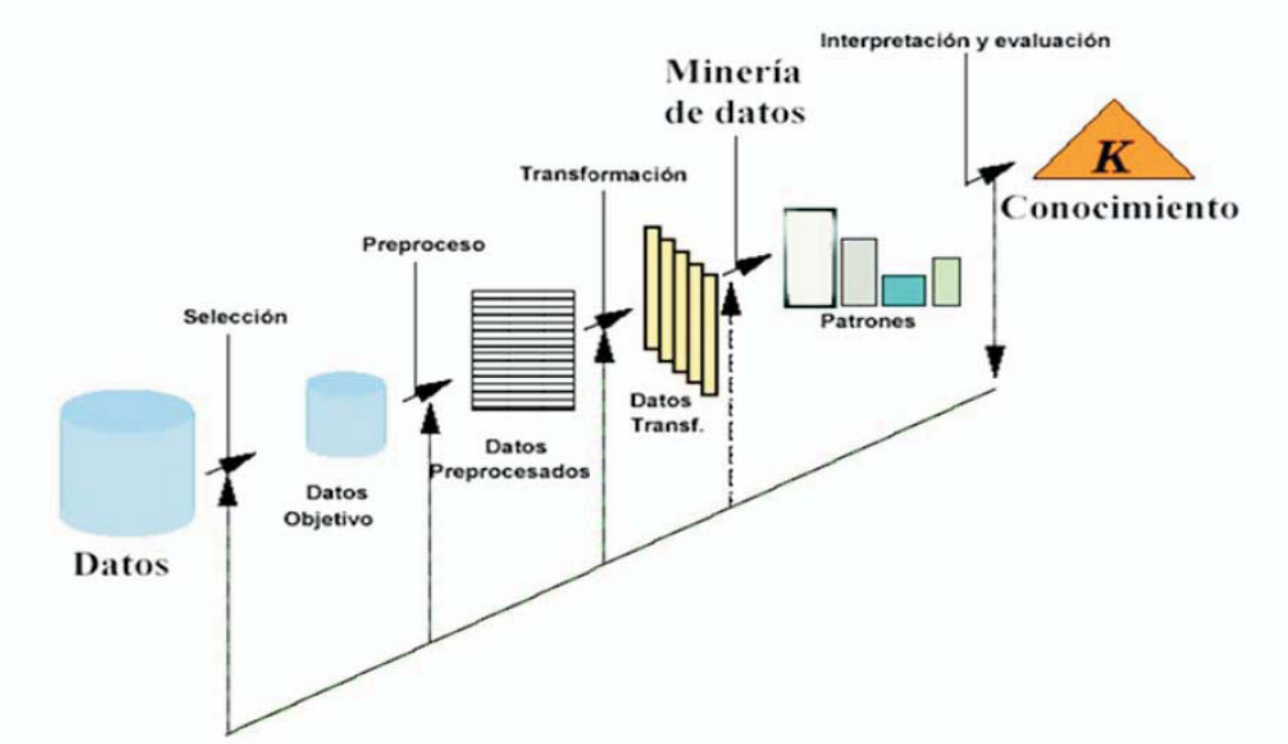
1. EL DESCUBRIMIENTO DE CONOCIMIENTO

La minería de datos es, en principio, una fase dentro de un proceso global denominado descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD)[4], aunque generalmente se asocia el concepto de minería de datos a todo el proceso, en lugar de la fase de extracción de conocimiento. El proceso de KDD es útil en el comercio electrónico para obtener conocimiento de las preferencias de compra de los clientes, y realizar campañas de marketing para aumentar los ingresos de una organización [15].

KDD se constituye de varias etapas que se ejecutan iterativa e interactivamente. El proceso es no trivial porque incluye acciones de cierta complejidad que involucran la búsqueda de estructuras, modelos y parámetros en la base de datos. Los patrones que se obtienen deben ser válidos con algún grado de certeza, novedosos preferiblemente para el usuario, al que deberán reportar algún tipo de utilidad.

El proceso KDD (ver figura 1) comienza con la definición y comprensión de un determinado problema y termina con el análisis de los resultados. Una de las propuestas más ampliamente extendida sobre las etapas o fases componentes del proceso KDD [5] incluye la comprensión del problema, la selección de los datos, su limpieza y preprocesamiento, la transformación y aplicación del método de descubrimiento (minera de datos) a utilizar y la interpretación de los patrones obtenidos o análisis de resultados.

[Figura] 1: Proceso KDD



1.1. MINERÍA DE DATOS

La minería de datos es una fase dentro del KDD y se define como "El proceso de extracción de información previamente desconocida, válida y útil de grandes bases de datos y el uso de la información para tomar decisiones cruciales de negocios" [6].

La minería de datos emplea una serie de técnicas las cuales son aplicadas para la solución de diversos problemas [7], sus herramientas predicen futuras tendencias y comportamientos, permitiendo tomar decisiones conducidas por un conocimiento obtenido de los datos. Para conseguir esto hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales [2].

1.2. OPERACIONES DE MINERÍA DE DATOS

Las cuatro mayores operaciones que se implementan en la minería de datos son análisis de caminos, segmentación de la base de datos, modelos predictivos y detección de desviación. Hay relación entre las operaciones y los algoritmos, y dependiendo del problema que se quiera resolver y las entradas de datos, se pueden mezclar las operaciones ya que no son mutuamente excluyentes [2]. A continuación se hará una descripción general de estas tareas.

1.2.1. Modelo Predictivo

Es aquel que esta relacionado con la experiencia de aprendizaje humana, en la cual se usan observaciones para formar un modelo de las características esenciales de algún fenómeno. La habilidad predictiva es crítica porque ayuda a hacer generalizaciones sobre el mundo que nos rodea y agregar nueva información dentro de un marco general [2]. En el comercio electrónico se utiliza este modelo en estrategias de marketing como cross selling, target marketing, aprobación de créditos y gerencia de retención de clientes.

1.2.2. Segmentación de la base de datos

La meta de la segmentación es la partición de la base de datos en segmentos de registros similares. Para particionar los registros se debe tener un numero de propiedades o características consideradas como homogéneas, los segmentos deben tener homogeneidad y heterogeneidad. La homogeneidad se refiere a que los registros en un segmento están próximos entre si, donde la proximidad es expresada como una medida dependiente de la distancia de los registros al centro del segmento. La heterogeneidad hace referencia a que los registros en diferentes segmentos no son similares unos de otros de acuerdo con una medida de similaridad [2], en el comercio, lo anterior se utiliza en la ejecución de estrategias de mercado futuras, tanto on-line como off-line, tales como envío de correo automático a aquellos clientes que se encuentren dentro de un cierto grupo, o presentación de contenidos específicos de publicidad e información según el tipo de cliente.

1.2.3. Análisis de Caminos

En contraste con el modelo predictivo y la segmentación de la base de datos, cuya meta es caracterizar el contenido de la

base de datos como un todo, las operaciones de análisis de caminos hacen una búsqueda en la base de datos para establecer relaciones entre registros individuales. Esas relaciones son llamadas asociaciones. Estas asociaciones se usan para cross selling, target marketing y movimiento de precio común. [2] Por ejemplo, se pueden utilizar para identificar asociaciones relacionadas con las compras a través del tiempo las cuales revelan información sobre la secuencia en la cual los clientes compran productos o servicios, todo esto, para entender perfiles de compra de los clientes a largo plazo y ofrecer promociones más oportunas.

1.2.4. Detección de desviación

Las operaciones de desviación son unas de las mejores herramientas para encontrar patrones ocultos en los datos. Dentro de la detección de desviación, la visualización es particularmente útil para percatarse de fenómenos que no se identifican en una muestra relativamente pequeña de los datos. Cuando se usa visualización es importante no tener una hipótesis preconcebida sobre el fenómeno que se está buscando [2] para evitar imprecisiones en los resultados. Esta operación tiene gran utilidad en las telecomunicaciones, en las cuales se usa un modelo para escanear las millones de transacciones que se realizan diariamente como un esfuerzo para detectar fraudes potenciales.

2. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de minería de datos implementan algoritmos específicos que son usados para realizar las operaciones, a continuación se hará descripción de las técnicas investigadas.

2.1. MÉTODOS ESTADÍSTICOS

El análisis de los datos estadísticos es un sistema de metodologías establecido para la explotación minera de datos. Históricamente, los primeros usos computarizados del análisis de datos fueron desarrollados con la ayuda de la estadística, extendiéndose del análisis de datos unidimensional al análisis de datos multidimensional. La estadística ofrece una variedad de métodos para la explotación de minería de datos, incluyendo diversos tipos de regresión y de análisis discriminantes [9].

2.2. ANÁLISIS DE CLUSTER

El análisis de cluster es un sistema de metodologías para clasificación automática de muestras en un número de grupos o clases usando una medida de asociación, de modo que las muestras en un grupo sean similares y las muestras que pertenecen a diversos grupos no sean similares. La entrada, para un sistema de análisis de cluster es una muestra y una medida de semejanza (o desemejanza) entre dos muestras. La salida del análisis de cluster es un número de grupos (clusters o racimos) en forma de una partición, o una estructura de particiones del grupo de muestra. Dentro de las aplicaciones del clustering más importantes se tiene el descubrimiento de distintos grupos entre los clientes, basados en sus patrones de compra [9].

2.3. ÁRBOLES DE DECISIÓN Y REGLAS DE DECISIÓN

Los árboles de decisión y las reglas de decisión son metodologías muy utilizadas en aplicaciones del mundo real de minería de datos como solución de gran alcance a los problemas de clasificación. En general, la clasificación es un proceso de aprendizaje de una función que ingresa un grupo de datos en una de varias clases predefinidas. Cada clasificación esta basada en algoritmos de aprendizaje inductivos y se dá como entrada a un sistema de muestras que consisten en vectores con valores cualitativos (también llamados vectores de característica) y de una clase correspondiente [9].

El atractivo de los árboles de decisión se debe a que representan reglas las cuales pueden ser expresadas fácilmente de modo que los seres humanos puedan entenderlas; la capacidad de explicar la razón de una decisión, es crucial. Por ejemplo, en marketing se debe describir los patrones del cliente a los profesionales de marketing, de modo que puedan utilizar este conocimiento para lanzar una campaña acertada de comercialización [10].

2.4. REGLAS DE ASOCIACIÓN

Las reglas de asociación son una de las técnicas principales de la explotación de minería de datos y es quizás la más común en el descubrimiento de patrones locales y sistemas de aprendizaje no supervisados[9]. Las técnicas de de reglas de asociación recuperan los patrones interesantes posibles en la base de datos, estas se aplican a los conjuntos de transacciones para descubrir todas las asociaciones y correlaciones entre las compras por parte de los clientes .

2.5. DETECCIÓN DE FRAUDES

Las técnicas de detección de fraude consisten en recopilar datos históricos para construir un modelo de conducta fraudulenta o potencialmente fraudulenta y encontrar instancias similares de esta conducta [2], estas técnicas se utilizan con mayor frecuencia las aplicaciones de minería de datos de organizaciones como las de cuidado medico, compañías de tarjetas de crédito, prestación de servicios y telecomunicaciones [2].

3. COMERCIO ELECTRÓNICO

En el Marketing es uno de los campos donde los éxitos de la minería de datos son más conocidos, cuanto más precisa sea la información que se tenga sobre los clientes, mayores posibilidades se tienen de aumentar los ingresos y rentabilizar al máximo las acciones. El objetivo fundamental puede resumirse en determinar quién comprará qué, cuándo y dónde [15]. Es por esto, que de acuerdo a las necesidades de este trabajo y acorde a ello, este trabajo esta enmarcado dentro del comercio electrónico, el cual se define como: ***"Intercambios mediados por la tecnología entre diversas partes (individuos, organizaciones o ambos), así como las actividades electrónicas y no electrónicas dentro y entre organizaciones que facilitan esos intercambios."*** [14].

Actualmente existen varios tipos de comercio electrónico, entre los cuales está el B2C que será objeto de estudio en el

presente artículo, el cual se lleva a cabo entre una empresa y el consumidor final, y va más allá de tener un catálogo en Internet con la posibilidad de que se pueda comprar en línea [11].

4. CRITERIOS DE SELECCIÓN

Para la selección de una técnica de minería de datos se ha de tener en cuenta una serie de consideraciones previas que afectan al desempeño de la técnica. El entender estas características y su impacto es útil para escoger la técnica que mejor se adecue a una determinada aplicación [2]. Teniendo en cuenta el entorno se seleccionaron los siguientes criterios [3]:

- **Habilidad para manejar datos con ruidos:** Las bases de datos a menudo contienen ruido en forma de imprecisiones o inconsistencias. Algunos procesos de validación de datos están mal diseñados y permiten introducir datos incorrectos a los usuarios.
- **Habilidad para manejar datos perdidos:** es importante darle un manejo apropiado a los datos ya que se pueden producir pérdidas si los datos se obtienen de diferentes fuentes.
- **Procesamiento de grandes volúmenes de datos:** es importante que la técnica posea la habilidad para manejar gran cantidad de información lo cual permite mayor precisión en el análisis.
- **Escalabilidad:** esta es una propiedad muy deseable en una técnica de minería de datos para futuras actualizaciones.
- **Procesamiento de diferentes tipos de datos:** es importante que la técnica seleccionada tenga la capacidad para poder manejar diferentes tipos de datos numéricos, cadenas, etc.
- **Capacidad predictiva:** Esta característica tiene gran influencia en la efectividad de la técnica de minería porque determina que tan buena es para la solución de un problema.
- **Facilidad de Operación:** La facilidad de integración y operación es otra característica importante para su utilización.
- **Capacidad explicativa:** Dependiendo de la técnica utilizada, el grado de procesamiento aplicable a los datos varía, por tanto, una técnica que sea fácil de entender y que requiera de poco preprocesamiento es más interesante para un usuario final.
- **Complejidad de implementación:** Es importante que la técnica seleccionada no presente un alto grado de complejidad para su implementación, lo que resulta conveniente para el desarrollo de una herramienta de minería.

4.1. RESULTADOS OBTENIDOS DEL ANÁLISIS DE LAS TÉCNICAS DE MINERÍA DE DATOS ESTUDIADAS

A cada uno de los criterios de selección escogidos se le asignó un valor numérico de uno a cinco, el cual indica el grado de impacto que tiene sobre la técnica de minería de datos seleccionada. Posteriormente se hizo una suma de los respectivos valores obtenidos de las diferentes técnicas estudiadas (Ver tabla 1) y de acuerdo a la investigación realizada, se obtuvo como resultado la información necesaria para la selección de la técnica de minería de datos.

Se observó que la técnica de reglas de asociación es la que mas se adecua a las necesidades específicas que deseamos cubrir de acuerdo a los criterios de selección.

Entre estos criterios se tiene, por ejemplo, que la fuente de

[Tabla] 1: Comparación de Criterios de las Diferentes Técnicas de Minería de datos

Tabla De comparación de Criterio de la Técnicas de Minería de Datos					
Técnicas / Criterios	Árboles de Decisión y Reglas de Decisión	Reglas de Asociación	Métodos Estadísticos	Análisis De Cluster	Detección de fraudes
Habilidad para manejar datos con ruidos	3	5	3	3	5
Habilidad para manejar datos Perdidos	3	5	3	3	5
Procesamiento de Grandes Volúmenes de Datos	4	5	3	4	5
Procesamiento de Diferentes Tipos de Datos	3	5	3	4	4
Capacidad Predictiva	4	4	3	4	4
Facilidad de Operación	4	5	5	4	4
Capacidad explicativa	4	4	4	5	4
Escalabilidad.	3	5	3	4	4
Complejidad de implementación	3	4	5	4	4
Total Puntuación	31	42	32	35	39

datos de la técnica proviene de sistemas operacionales que ya poseen un grado alto de depuración de los datos lo que evita impurezas que puedan afectar a la técnica [10], además esta técnica sugiere una búsqueda por toda la base de datos, realizando una clasificación en cada barrido, por lo tanto no hay límite establecido para la cantidad de datos que puede manejar. Tambien busca las características o cualidades presentes en las transacciones realizadas, las cuales pueden tener atributos de diferentes tipos, y por ende no es necesario hacer una conversión a un solo tipo de datos específico.

La capacidad predictiva de la técnica depende de las medidas de confianza y soporte establecidas, ya que esta técnica se basa en el conteo de posibles ocurrencias entre las combinaciones de ítems en una tabla de transacciones, no requiere consideraciones algorítmicas muy complejas y posee gran escalabilidad ya que realiza un barrido por la base de datos con lo cual puede operar sin mayores problemas con un número grande de datos .

5. REGLAS DE ASOCIACIÓN EN LAS TRANSACCIONES DE NEGOCIOS

Las reglas de Asociación encuentran asociaciones o correlaciones interesantes en los registros de las transacciones de negocios, donde una transacción T está formada por un conjunto de artículos o ítems. A un conjunto de ítems se le suele denominar itemset, en general, o itemset de grado k, cuando se especifica el número k de ítems que incluye. Una regla de asociación es una implicación X -> Y en la cual X e Y son itemsets de intersección vacía, es decir sin ítems en común. El significado de la regla de asociación es que las transacciones o tuplas que contienen a X también tienden a contener Y. Las reglas de asociación deben superar además un valor para dos medidas denominadas confianza y soporte de la regla. La confianza de la regla de asociación es la proporción de las transacciones que, conteniendo a X, también incluyen a Y, mientras que el soporte es la fracción de transacciones en la base de datos que contienen tanto a X como a Y [2].

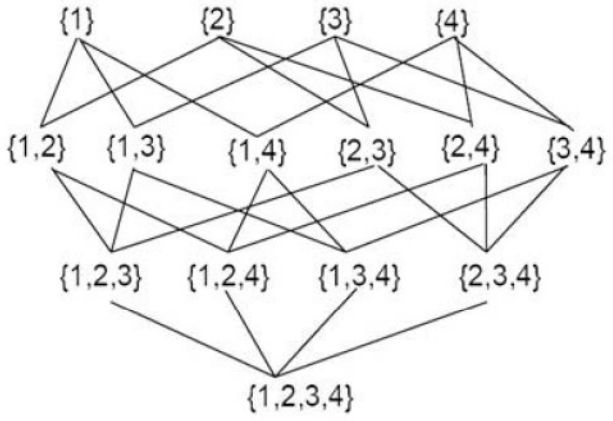
Las correlaciones encontradas en los registros de las transacciones de negocios para muchas organizaciones resultan interesantes. Actualmente con la masiva cantidad de datos que las organizaciones recolectan de sus procesos de negocio, el descubrimiento de asociaciones interesantes en los registros de transacciones puede ayudar en la toma de decisiones en los procesos de marketing [10]. Un típico ejemplo de regla de asociación es el ***"análisis de canasta de mercado"*** el cual busca descubrir asociaciones entre los atributos de una canasta, es decir, busca descubrir reglas para cuantificar la relación entre dos o más atributos [11].

5.1. ALGORITMOS PARA LA BÚSQUEDA ÍTEMSETS FRECUENTES

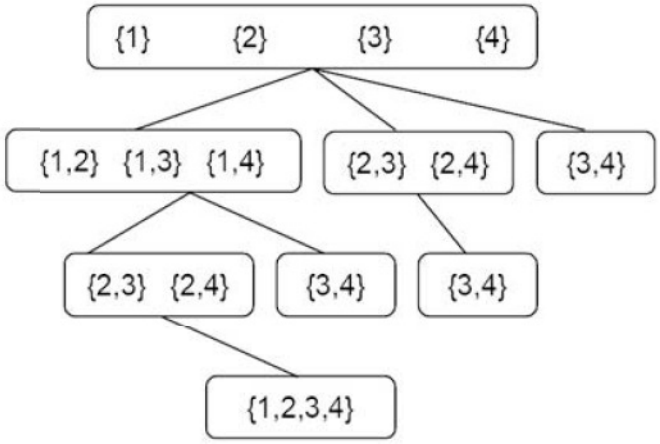
Un ***itemset*** es un conjunto de artículos contenidos en una transacción de compra de productos. Para la búsqueda de los ítemsets frecuentes se emplean dos formas comunes de búsqueda en árbol: primero a lo ancho (BFS, por sus siglas en inglés) y segundo en profundidad (DFS, por sus siglas en inglés), (figuras 2 y 3). Los algoritmos más comunes para el cálculo de los ítemsets frecuentes se muestran en la figura 4.

Con la estrategia BFS el valor del soporte de los (k-1) itemsets se determina antes de contar el soporte de todos los k-itemsets. Con la estrategia DFS, no se conocen todos los (k-1) itemsets, pero sí los necesarios (k-1) itemsets cuando se

[Figura] 2: Árbol usado en los algoritmos con estrategia BFS



[Figura] 3: Árbol usado en los algoritmos con estrategia DFS

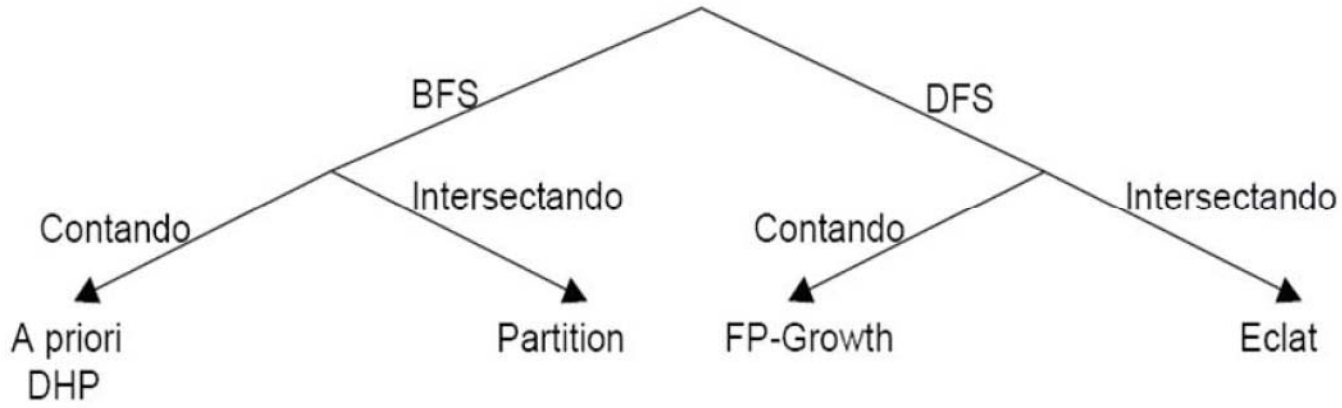


genera cada uno de ellos, pues trabaja de forma recursiva descendiendo por el árbol.

[12].

5.2. ALGORITMOS PARA REGLAS DE ASOCIACIÓN

[Figura] 4: Algoritmos para el cálculo de itemsets frecuentes



Para la técnica de reglas de asociación existe una serie de algoritmos. A continuación se hará una breve descripción de los investigados.

5.2.1. Algoritmo A priori

El algoritmo A priori trabaja de la siguiente manera: primero busca todos los conjuntos frecuentes unitarios (contando sus ocurrencias directamente en la base de datos), éstos se mezclan para formar los conjuntos de candidatos de 2-itemsets y seleccionan entre ellos los frecuentes. Considerando la propiedad de los conjuntos de ítems frecuentes, vuelve a mezclar éstos últimos y selecciona los frecuentes. Así sucesivamente se repite el proceso hasta que en una iteración no se obtengan conjuntos frecuentes [12].

Este algoritmo asume un orden entre los ítems y utiliza las siguientes notaciones: Ck, para el conjunto de candidatos de k-itemsets, Fk, para el conjunto frecuente de k-itemsets y asociado a cada itemset se encuentra el campo count para almacenar el soporte de dicho itemset.

5.2.2. Algoritmo DHP (Poda y hashing directa)

En el algoritmo de poda y hashing directa (DHP, Direct hashing and Pruning) se emplea una técnica de hash para eliminar todos los conjuntos de ítems innecesarios para la generación del próximo conjunto de ítems candidato. Cada (k+1)-itemset es añadido a una tabla hash en un valor hash dependiente de las ocurrencias en la base de datos de los conjuntos candidatos de k elementos que lo formaron, o sea, dependiente del soporte de los conjuntos candidatos de k elementos. Estas ocurrencias son contadas explorando en las transacciones de la base de datos. Si el soporte asociado a un valor hash es menor que el soporte mínimo, entonces, todos los conjuntos de ítems de k+1 elementos con este valor hash no serán incluidos entre los candidatos de k+1 elementos en la próxima iteración.

5.2.3. Algoritmo Partition

Este algoritmo en primer lugar propone particionar la base de datos en tantas partes como fueran necesarias para que todas las transacciones en cada partición sean alojadas en la memoria principal operativa. En contraste con otros algoritmos, este recorre la base de datos sólo dos veces. La primera vez, cada partición es minada independientemente para encontrar todos los conjuntos de ítems frecuentes en la partición y luego se mezclan éstos para generar el conjunto de los conjuntos de ítems candidatos. En la segunda iteración se cuenta la ocurrencia de cada candidato, aquellos cuyo soporte es mayor que el mínimo soporte especificado, se retienen como conjuntos frecuentes. Este algoritmo emplea el mecanismo de intersección entre conjuntos para determinar el soporte de dichos conjuntos, en este caso, cada ítem en una partición mantiene la lista de los identificadores de las transacciones que contienen a dicho ítem [12].

5.2.4. Algoritmo Eclat

EL algoritmo Eclat reduce la cantidad de operaciones de E/S, aunque atraviesa la base de datos sólo una vez. Se basa en realizar un agrupamiento (clustering) entre los ítems para aproximarse al conjunto de ítems frecuentes maximales y luego emplea algoritmos eficientes para generar los ítems frecuentes contenidos en cada grupo [12]. Para el agrupamiento hay dos métodos que son empleados después de descubrir los conjuntos frecuentes de dos elementos:

1.Por clases de equivalencia: esta técnica agrupa los itemsets que tienen el primer ítem igual.

2.Por la búsqueda de cliques maximales: se genera un

grafo de equivalencia cuyos nodos son los ítems y los arcos conectan los ítems de los 2-itemsets frecuentes. Se agrupan los ítems por aquellos que forman cliques maximales.

Con el primero de los métodos se generan todos los conjuntos de ítems frecuentes, con el segundo, se generan sólo los conjuntos frecuentes maximales, los restantes conjuntos frecuentes son subconjuntos de éstos [12].

5.2.5. Algoritmo FP-Growth

Este algoritmo esta basado en una representación de árbol de prefijos de una base de datos de transacciones dada (llamada FP-Tree), el cual puede almacenar considerables cantidades transacciones en la memoria. La idea básica del algoritmo FP-Growth puede ser descrita como un esquema de eliminación recursiva: en un primer paso de preprocesamiento se borran todos los ítems de las transacciones que no son frecuentes individualmente o no aparecen en el mínimo soporte de transacciones, luego se seleccionan todas las transacciones que contienen al menos un ítem frecuente (al menos una cantidad de esos que son frecuentes). Luego se hace el proceso recursivo para obtener una base de datos reducida. Al retorno, se remueven los ítems procesados de la base datos de transacciones en la memoria y se empieza otra vez, y así con el siguiente ítem frecuente [13].

5.3 CRITERIOS DE SELECCIÓN DEL ALGORITMO

Para la selección del algoritmo de Reglas de Asociación hay que tener en cuenta una serie de consideraciones que tienen impacto en la solución que se desea brindar [12]:

- ***Inserciones a la base de datos:*** Es importante que los algoritmos minimicen el recorrido por la base de datos, pues el número de reglas crece exponencialmente con el número de ítems considerados, lo cual afecta el rendimiento del algoritmo cuando se accesa constantemente a la base de datos.
- ***Costo computacional:*** es importante que el algoritmo no realice un gran numero de operaciones que agoten los recursos de maquina.
- ***Tiempo de Ejecución:*** se desea que el tiempo utilizado para la generación de reglas sea razonable.
- ***Rendimiento:*** es importante que el algoritmo realice las operaciones y procesos de forma eficiente.

Posteriormente se evaluaron los criterios para cada uno de los algoritmos y se observó que el algoritmo FP-Growth tiene ventajas operacionales sobre los otros algoritmos, entre las cuales están: que no requiere de la generación de ítems candidatos, requiere de pocos accesos a la base de datos, no requiere de un costo computacional elevado, y tiene un tiempo de ejecución superior a los otros algoritmos como por ejemplo el A priori [10].

5.4 DISEÑO DEL ALGORITMO FP-GROWTH

Con el objetivo de implementar este algoritmo para el proyecto, a continuación se muestran las etapas de funcionamiento del mismo. El corazón del algoritmo FP-Growth es el cómputo de un Árbol FP de una base de datos proyectada, esto es, una base de datos de transacciones que contiene un ítem específico, esta proyección de base de datos es procesada recursivamente [13].

El algoritmo hace un preproceso de la base de datos de transacciones de la siguiente forma: en un escaneo inicial se determina la frecuencia de los ítems, aquellos ítems que aparecen en las transacciones con menor cantidad de un mínimo especificado por el usuario, se descartan de las transacciones, dado que no pueden ser parte de un grupo de ítems frecuentes. Los ítems en cada transacción son almacenados y luego se ordena su frecuencia de forma descendente en la base de datos [13].

Luego se borran todos los ítems infrecuentes de la base de datos de transacciones, y se pasa al Árbol FP. Un Árbol FP es básicamente un árbol de prefijos para las transacciones, esto es, que cada camino representa el grupo de transacciones que comparten el mismo prefijo, cada nodo corresponde a un ítem. Todos los nodos que referencian al mismo ítem se enlazan juntos en una lista, esto es que todas las transacciones que contienen un ítem específico pueden encontrarse fácilmente y contarse al atravesar la lista. Esta lista puede ser accesada a través de la cabeza, el cual también expone el número total de ocurrencias del ítem en la base de datos. Posteriormente el algoritmo mina la estructura del árbol FP encontrando las correlaciones entre los ítems frecuentes (proceso de generación de reglas) [13].

6. CONCLUSIONES

- La investigación de las diferentes técnicas de minería de datos y su empleo en la solución de diversos tipos de problemas de análisis de información nos ayudan a tener un conocimiento general del tema para desarrollar trabajos futuros en otras áreas de conocimiento.
- Para la escogencia de una técnica de minería de datos es necesario entender las necesidades propias del trabajo a desarrollar y tener en cuenta consideraciones y criterios para seleccionar una técnica adecuada que satisfaga los requerimientos propios del mismo.
- La generación de reglas de asociación, fue la técnica de minería de datos que se selecciono, porque es un método utilizado en análisis de canasta de mercado que permite encontrar regularidades en la conducta de compras de los clientes. Se generan reglas de la forma "Si condición entonces consecuencia", que son muy fáciles de comprender para la persona encargada de analizar los resultados.
- La investigación de técnicas de Minería de datos aplicadas a problemas específicos permitió delimitar los requerimientos del proyecto y el uso del Algoritmo FP-Growth uno de los algoritmos mas rápidos y eficientes para minar de reglas de asociación, permitiendo la realización del proceso de minería de datos de forma rápida.

7. REFERENCIAS

[1].José Alberto Castañeda García Miguel Ángel Rodríguez Molina "La Minería de Datos como herramienta de Marketing: Delimitación y Evaluación del resultado" Facultad de CC. EE.

Departamento de Comercialización e Investigación de mercados Universidad de Granada Granada España - 2005.
[2].CABENA, Hadjinjan., STADLER, Verhees Zanasi., Discovering Data mining From Concept To Implementation. Prentice Hall PTR. Upper Saddle River. 1998 - 224 paginas.
[3].Alejandro Amat Bedmar - "Ingeniería De Conocimiento Minería De Datos Empresariales" - M.S. / E.T.S. Ingeniería Informática de la Universidad de Granada - 2005.
[4].Garlan D., «Software Architecture: a Roadmap,» The Future of Software Engineering, ACM Press, pp.91-101, 2000. J. Zhao, «Applying Slicing Technique to Software Architectures,» Proc. Fourth IEEE International Conference on Engineering of Complex Computer Systems (ICECCS98), pp.87-98, Agosto - 1998.
[5].Fayyad U., Piatetsky-Shapiro G., Smyth P.,Discovery and Data Mining. AAAI Press/The MIT Press - 1996.
[6].KENNETH C. LAUDON, Jane P. Administración de la información y toma de decisiones, Resúmenes de los principales capítulos del libro, Management Information Systems Organization and Technology. Documento. Universidad de Taparaca. Chile - 2005.
[7].KIMBALL, Ralph., ROSS, Margy. The Data Warehouse Toolkit The Complete Guide to Dimensional Modeling. Libro. McGraw Hill. Second Edition 2002 - 464 paginas.
[8].Piatestky-Shapiro G. Report on the AAAI-91 Workshop on Knowledge Discovery in Databases, IEEE Expert 6(5), pp.74-76 Octubre - 1991.
[9].Mehmed Kantardzic - Data Mining: Concepts, Models, Methods, and Algorithms - John Wiley & Sons -2003 -343 paginas.
[10].Jiawei Han, Micheline Kamber - Data Mining: Concepts and Techniques Simon Fraser University Morgan Kaufmann Publishers - 2002 - 550 paginas.
[11].Daniel T. Larose "Discovering Knowledge in Data: An Introduction to Data Mining" - John Wiley & Sons - 2004 -240 paginas.
[12].Roxana Danger Mercaderes - Rafael Berlanga Llavori- Informe técnico -"Búsqueda de Reglas de Asociación en bases de datos y colecciones de textos" - Departamento de Computación, Universidad de Oriente, Santiago de Cuba 2001.
[13].Christian Borgelt-Department of Knowledge Processing and Language Engineering -School of Computer Science, OttovonGuericke - University of Magdeburg 2005.
[14].RAYPORT JEFFREY, F., JAWORSKI BERNARD J. Comercio Electrónico. McGraw Hill. - 2000 - 456 paginas.
[15].NUÑEZ, Fernando Alberto., LUGONES, Fernando Alberto., Modelos de Negocios en Internet visión poscrisis. McGrawHill. 2001 - 384 paginas.
[16]. Waikato ML Group. The Waikato environment for knowledge analisys. http://www.cs.waikato.ac.nz/ml/weka - 2007
[17].Faculty of Computer and Slovenia Information Science, University of Liubliana. Orange, fruitful and fun. http://www.ailab.si/orange - 2007.
[18].R Rakotomalala. Tanagra project. http://chirouble.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html - 2007.
[19].The CRISP-DM consortium. CRISP-DM Step by step data mining guide. -2000 -. Documento. http://www.crisp-dm.org/CRISPWP-0800.pdf ,2007.