Facultad 3

Universidad de las Ciencias Informáticas

Trabajo de Diploma para optar por el título de Ingeniero Informático

Título: Generalización de Itemsets en la etapa de pre-procesamiento para la extracción de reglas de asociación.

Autor: Carlos Alberto Hernández Miranda

Tutor: Msc. Julio Cesar Diaz Vera

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Carlos Alberto Hernández Miranda	Julio Cesar Diaz Vera
_____	_____
Autor	**Tutor**

DATOS DE CONTACTO

Julio Cesar Diaz Vera, graduado de Ingeniería en Telecomunicaciones en la Universidad Martha Abreu ubicada en Villa Clara, Cuba. Máster en Gestión de Proyectos, grado científico alcanzado en la Universidad de la Ciencias Informáticas, La Habana, Cuba. Posee varias publicaciones nacionales e internacionales referentes al tema abordado en esta investigación, alcanzando una experiencia de más de 10 años.

Correo electrónico: jcdiaz@uci.cu

AGRADECIMIENTOS

Quisiera agradecer primeramente a mis padres por darme la vida, la seguridad y el sustento. Por sus consejos oportunos, por su amor, su paciencia y su dedicación. A mis abuelos Carmen, Miguel y Gladis que siempre están presentes en cada momento de mi vida y que me han guiado a través de los años, gracias por estar siempre para mí. A mis hermanos que tal vez por su edad no entiendan el significado del sacrificio y la dedicación de cinco años de estudios, pero que espero que vean en mí un ejemplo para sus días por venir. Al resto de mi familia, a mi tío Migue, el primer master en ciencias de la familia pero espero que no el ultimo. A Yohanka, Inés, Efraín, Boris y la Nena, que forman parte de mi familia también, gracias por su apoyo incondicional, por su cariño y por aceptarme en su casa como un hijo o un nieto más de la familia, los quiero.

También quisiera expresar mi más profundo agradecimiento a mis hermanos “vandálicos” durante estos cinco años de carrera, a Bernardo, Andy, Daniel, Roberto y Boris; compañeros de mil batallas, la mayor parte de ellas victoriosas. A Yelenis mi novia y mi amiga que celebra mis logros y llora mis angustias como si fueran las de ellas, gracias mi amor por estos años, y sobre todo por soportarme. A las chicas VIP Analay y Janet, que también han sido hermanas para mí. Al resto de mis amistades Arturo y Dianela que estamos en esto desde primer año, a mis hermanos del barrio Pocholo y Jansel este logro también es de ustedes.

Gracias a todas las personas que han contribuido a mi formación como profesional. A mis profesores, a Julio tutor de este trabajo, amigo incondicional y paño de lágrimas insustituible, gracias por tu apoyo. A mis compañeros de clases durante todos estos años y a los que no han podido terminar la carrera.

DEDICATORIA

A la memoria de mi abuela Lidia Ferrán, veladora de mis primeros pasos y sueños, educadora y formadora de la persona que soy y seré…

RESUMEN

La complejidad computacional de los algoritmos de extracción de reglas de asociación es uno de los principales problemas de esta área del conocimiento y los trabajos encaminados a mitigar este inconveniente están enmarcados en la etapa de pre-procesamiento de los datos o sobre los propios algoritmos. Este trabajo propone un marco integrado para la generalización de ítems en la etapa de pre-procesamiento del minado de reglas de asociación, basado en relaciones taxonómicas contenidas en el conocimiento previo de los usuarios, expresado en una ontología de dominio específico. La clasificación de mercancías de acuerdo al riesgo de fraude es una de las tareas principales de la aduana de cualquier país, esta actividad entra en contradicción con el facilitamiento del comercio internacional que es otra de las misiones de la aduana, por lo que la aplicación de técnicas de minería de reglas de asociación podría arrojar resultados positivos. El caso de estudio propuesto para la validación de la solución está enmarcado en el contexto anterior.

PALABRAS CLAVE

Minería de datos, reglas de asociación, ontología, generalización de ítems.

ÍNDICE DE CONTENIDOS

AGRADECIMIENTOS I
DEDICATORIA II
RESUMEN III
INTRODUCCIÓN. 7
1. CAPÍTULO 1: MARCO TEÓRICO 11
1.1 Minería de datos. 11
1.2 Reglas de asociación. 12
1.3 Medidas de evaluación de las reglas de asociación. 13
1.4 Reglas de asociación generalizadas. 15
1.5 Extracción de reglas de asociación. 16
1.6 Utilización de conocimiento previo para la extracción de reglas de asociación generalizadas. 18
1.7 Ontologías. 18
1.8 Utilización de ontologías en la minería de datos. 21
1.9 Utilización de ontologías en las reglas de asociación. 21
1.10 Trabajos relacionados. 22
1.11 Conclusiones parciales 23
2. CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL ALGORITMO 24
2.1 Introducción. 24
2.2 Propuesta de solución. 27
2.3 Implementación. 29
2.4 Conclusiones parciales 31
3. CAPÍTULO 3: VALIDACIÓN 32
3.1 Evaluación y validación. 32
3.2 Descripción de la demostración 34
3.2.1 Recursos utilizados en la demostración. 35
3.2.2 Conjuntos de datos 35
3.2.3 Casos de prueba 36
3.3 Discusión de los resultados 38
3.4 Conclusiones parciales 40
CONCLUSIONES 42
RECOMENDACIONES 43
BIBLIOGRAFÍA 44
ANEXOS 47
GLOSARIO 48

ÍNDICE DE TABLAS

Tabla 1: Muestra de datos asociados a las declaraciones de mercancías aduanales. 25
Tabla 2: Muestra de datos generalizados asociados a las declaraciones de mercancías aduanales. 27
Tabla 3: Características de los conjuntos de datos utilizados en los casos de pruebas. 35
Tabla 4: Casos de prueba para el parámetro k=1. 37
Tabla 5: Casos de prueba para el parámetro k=2. 37
Tabla 6: Casos de prueba para el parámetro k=3. 38
Tabla 7: Reducción porcentual de la instancia de entrada. 39
Tabla 8: Tiempos de ejecución para los casos de prueba. 40

ÍNDICE DE FIGURAS

Figura 1: Fragmento de taxonomía de la ontología de mercancías. 26
Figura 2: Fragmento de un fichero con el formato OWL/XML. 29
Figura 3: Funcionamiento de la función Generalizar. 30

INTRODUCCIÓN

La dicotomía de la misión de la aduana de un país entre el control del tráfico de mercancías y el facilitamiento del comercio internacional hace que sea deseable para las mismas poder contar con mecanismos que a partir de datos previos permitan predecir aquellas cargas de mayor riesgo para que sean estas las que pasen a examen físico. El cobro de los impuestos de aduana constituye una fuente importante de ingresos al país, su monto está estrechamente relacionado al tipo de mercancía y la cantidad de esta involucrada en la operación de importación-exportación. Existen tres comportamientos fundamentales que son utilizados para evitar el cumplimiento de estas obligaciones cuando las mercancías pasan por la aduana, estos pueden ser categorizados como ocultamiento, declaración de menores cantidades y la incorrecta clasificación de las mercancías. Todos estos comportamientos encajan dentro de la denominación de fraude aduanal (Laleh, et al., 2009).

La mayoría de los países consumen importantes recursos materiales y humanos con el fin de garantizar, mediante el chequeo físico, la veracidad de lo declarado en las operaciones aduanales. Este proceso tiene dos elementos negativos, el primero de ellos asociado a que el volumen de tráfico mercantil actual hace muy difícil examinar más del 10% de las mercancías y comúnmente solo el 1% de las inspeccionadas son detectadas como fraudulentas (Laleh, et al., 2009), el segundo elemento está asociado al compromiso de las aduanas en la facilitación del comercio que se ve obstaculizado en la medida que aumenta el chequeo físico.

La mayoría de los países, sobre todo los desarrollados, utilizan sistemas de información que reducen la cantidad de papel y agilizan el despacho de las mercancías. Al mismo tiempo han introducido técnicas no intrusivas, como los rayos Gamma y los rayos X, para agilizar la revisión física de las mercancías. A pesar de ello los volúmenes actuales de tráfico mercantil hacen imposible que las aduanas puedan mantener los estándares de tiempo para el despacho de mercancía a la par que realizan revisión física del 100 % de la carga.

En este escenario la utilización de técnicas de minería de datos que analicen la información histórica y permitan crear modelos predictivos pudiese ser relevante para alcanzar resultados positivos. En particular la minería de reglas de asociación que posibilita la obtención de reglas relevantes para la clasificación del riesgo de fraude en las declaraciones de mercancías.

Es un hecho, que en los últimos años el volumen de los datos almacenados ha crecido de forma considerable, ya sea por la informatización de las tareas o por el almacenamiento de la información en formato digital. Contar con registros históricos ha incrementado el interés por poder utilizar estos datos para obtener información desconocida y de valor para producir un impacto positivo sobre la organización. Este fenómeno, según (Ruiz, 2010) ha favorecido el surgimiento de nuevas herramientas que permiten manejar grandes volúmenes de datos y a su vez adquirir información oculta en ellos que pueda ser de utilidad en algún sentido. Estas herramientas forman parte del campo de la *Extracción de Conocimiento* (KD, Knowledge Discovery).

El término *extracción de conocimientos de bases de datos* (KDD, Knowledge Discovery Databases) es usado para nombrar el proceso de descubrir conocimiento útil en los datos, y la minería de datos (DM, Data Mining) es la aplicación de algoritmos específicos para extraer *patrones* de los datos, entendiendo patrones en un sentido amplio (relaciones, correlaciones, tendencias, agrupamientos y clasificaciones) (Fayyad, et al., 1996).

La extracción de reglas de asociación (Agrawal, et al., 1993) es una técnica de minería de datos que ha recibido mucha atención en los últimos años. Originada primeramente en el contexto del análisis de bolsas de compras para organizar estrategias de *marketing*, planificación de almacenes y promoción de artículos. En la actualidad el uso de las reglas de asociación se extiende más allá del *marketing* probando su aplicación en campos como la genética y la medicina.

Uno de los principales problemas que ha sido objeto de investigación en el área radica en la gran cantidad de reglas que son generadas y en la dificultad de utilizar una gran parte de estas dentro del proceso de toma de decisiones ya sea porque son obvias, demasiado generales, demasiado específicas o porque no tienen interés para el usuario final (Cao, et al., 2010).

La mayoría de las investigaciones en esta área están orientadas a disminuir la complejidad computacional del minado de reglas de asociación y a aumentar la calidad de las reglas extraídas. Comúnmente los esfuerzos realizados en la temática siguen tres direcciones diferentes:

- Definir mecanismos más eficientes para cubrir el espacio de búsqueda.
- Explotar estructuras de datos más eficientes.

- Utilizar conocimiento previo del dominio particular.

De manera general los trabajos que pretenden obtener reglas de mayor calidad se centran en la etapa de pos-procesamiento mientras que los que pretenden disminuir la complejidad computacional trabajan directamente sobre los algoritmos o en la etapa de pre-procesamiento.

En este trabajo se toma como **problema a resolver:** ¿Cómo disminuir la complejidad computacional de la extracción de reglas de asociación en las declaraciones de mercancías mediante la reducción del tamaño de la instancia sin perder generalidad en las reglas generadas?

Para llevar a cabo esta investigación se propone como **objetivo general:** Implementar un algoritmo que sea capaz de generalizar un grupo de ítems, de acuerdo a una jerarquía semántica, de manera que se reduzcan los itemsets candidatos y por tanto la cantidad de reglas generadas. Para alcanzar este objetivo se define como **objeto de estudio:** Minado de reglas de asociación y como **campo de acción:** Etapa de pre-procesamiento para el minado de reglas de asociación. Los objetivos específicos y tareas a desarrollar para cumplimentar la investigación fueron desarrollados siguiendo la propuesta de (Berndtsson, et al., 2008) y se listan a continuación.

Objetivos específicos:

- Establecer el marco conceptual de referencia.

Utilizando el método de Análisis Bibliográfico se trata de recopilar la mayor cantidad de información referente a tecnologías utilizadas en el desarrollo de algoritmos de generalización de ítems, y decidir si se puede utilizar alguna implementación en particular o es necesario desarrollar una propia, por lo que se plantean las tareas siguientes:

  - ✓ Recopilación de la bibliografía referente al tema.
  - ✓ Selección de la bibliografía relevante.
  - ✓ Análisis de la información de relevancia.

- Definir los requisitos especiales de los datos.

- Proponer el algoritmo de generalización.
- Definir el modelo de componentes.

Para el desarrollo de este objetivo se utilizará el método de Implementación y se llevarán a cabo las siguientes tareas:

- ✓ Selección de la estrategia de generalización.
- ✓ Definición de un marco arquitectónico.
- ✓ Implementación de las funciones asociadas al algoritmo.

- Probar la validez del resultado.

Para dar cumplimiento a este objetivo se utilizará el método de Casos de Estudio y se deben cumplir las siguientes tareas:

- ✓ Validación del resultado obtenido.
- ✓ Presentación de los resultados.

Cumplidos estos elementos se espera como **posibles resultados:** Algoritmo de generalización de ítems que posibilite la reducción de la instancia de entrada a los algoritmos de generación de reglas de asociación.

La estructura capitular de la investigación se explica a continuación. En el capítulo uno se desarrolla el marco teórico en el que está basada la investigación; en dicho capítulo se ofrecen elementos teóricos y definiciones formales fundamentales en el objeto de estudio del presente trabajo. En el segundo capítulo se describe el análisis y diseño de la solución propuesta; para desarrollar este apartado se explica el funcionamiento de la solución a través de ejemplos sencillos y de la descripción del algoritmo propuesto. Para finalizar en el último capítulo se desarrolla la evaluación y validación de la solución mediante el método de “Demostración” y las posteriores conclusiones de la investigación.

1. CAPÍTULO 1: MARCO TEÓRICO

1.1 Minería de datos.

El término minería de datos es comúnmente asociado al de KDD, aunque ambos están muy relacionados no son lo mismo (Zhou, et al.). Según (Fayyad, et al., 1996) KDD es el proceso no trivial de identificar patrones en los datos que sea válido, novedoso, potencialmente útil y comprensible; no trivial se refiere a que es un proceso que no puede ser superficial, los patrones que se identifiquen no deben estar a simple vista y deben producir algún efecto positivo sobre el conocimiento humano. Por otra parte (Mannila, et al., 2001) define como minería de datos el análisis de conjuntos de datos (a menudo de gran tamaño) de observación para encontrar relaciones insospechadas y para resumir los datos en nuevas formas que son comprensibles y útiles para el propietario de los datos, la cual se tomará como referencia en el presente trabajo.

Se ha tomado la definición de patrones de la siguiente forma: dado una serie de hechos D, un lenguaje L, y una medida de certeza C, un patrón P es enunciado en L que describe las relaciones entre varios subconjuntos de D con una certeza dada mediante C, tal que P es más sencillo (en algún sentido) que la enumeración de todas las relaciones entre dichos subconjuntos (Frawley, et al., 1992).

En la aplicación de la minería de datos es necesario reunir las experiencias de varias disciplinas de estudios como las técnicas de *"machine learning"*, reconocimiento de patrones, estadísticas, bases de datos y visualización para hacer frente al tema de la extracción de información. Otras áreas que contribuyen son las redes neuronales, análisis de datos espaciales y procesamiento de señales (Han, et al., 2001).

Los usos más comunes dados a la minería de datos son listados a continuación (Larose, 2004):

- Descripción: encontrar la manera de describir los patrones y las tendencias de los datos.
- Clasificación: para examinar los registros que contienen información sobre una variable objetivo categórica, que puede dividirse en varias clases o categorías y crear conjuntos de entrenamiento. Con base en las clasificaciones de los conjuntos de entrenamientos estas se le asignan a los nuevos registros.

- Estimación: similar a la clasificación, salvo que la variable de destino es más numérica que categórica.
- Predicción: es similar a la clasificación y la estimación a excepción de que para la predicción, los resultados están en el futuro.
- Agrupamiento (clustering): para agrupar los registros, observaciones o casos en grupos, que son colecciones de registros similares entre sí y diferentes a los registros de otros grupos.
- Asociación: para encontrar los atributos que "van de la mano". Las reglas de la asociación son de la forma: "Si antecedente, entonces consecuente", junto con una medida del soporte y la confianza asociados a la regla.

1.2 Reglas de asociación.

Formalmente (Agrawal, et al., 1993) una regla de asociación (RA) puede definirse como:

Sean I un conjunto finito de ítems, D una base de datos donde cada transacción T tenga un único identificador y contenga un conjunto de ítems. La regla de asociación es una implicación de la forma:

$$X \rightarrow Y \qquad (1)$$

Donde $X, Y \subset I$, son conjuntos de ítems llamados itemsets cumpliendo que $X \cap Y = \phi$. Se tomará a X antecedente y a Y consecuente de la regla de asociación anterior. Siguiendo este formalismo la implicación {sobres, papel} → {sellos} denota la aparición conjunta de sellos, sobres y papel en las compras de un establecimiento de correo. Las reglas de asociación también traen relacionadas medidas que indican su calidad y/o valor para el usuario.

Existen varios tipos de RA, que permiten definir subconjuntos de reglas, optimizando la ejecución de los algoritmos, reduciendo los volúmenes de reglas a almacenar y posibilitando al mismo tiempo la recuperación del conjunto original de reglas relevantes o confiables (Medina, et al., 2007). Estos tipos de RA son los siguientes:

- Reglas de asociación representativas (RAR) (Kryszkiewicz, 1998a).
- Reglas de asociación de mínima condición y máxima consecuencia (RAMM) (Kryszkiewicz, 1998b).

- RA generalizadas (RAG) (Srikant, et al., 1995).

El concepto de RAR fue introducido en 1998 por Marzena Kryszkiewicz para hacer frente al problema de la gran cantidad de RA que pueden ser generadas. Las RAR tienen la característica de ser el conjunto mínimo de RA a partir del cual se puede generar el total de reglas que satisfagan los umbrales de soporte y confianza mínimos a través de un operador de cubrimiento (C); este operador es generado moviendo para el antecedente parte de los ítems del consecuente manteniendo como consecuente parte o todos los ítems no movidos. Formalmente se puede definir entonces una RAR si verifica la expresión (Kryszkiewicz, 1998a):

$$RAR = \{r \in RA \mid \neg\exists r' \in RA, r' \neq r, r \in C(r')\} \quad (2)$$

$$C(X \rightarrow Y) = \{X \cup Y \rightarrow V \mid Z, V \subseteq Y, Z \cap V = \phi, V \neq \phi\} \quad (3)$$

Se ha indicado el conjunto de reglas que cumplen con el umbral de soporte y confianza como RA.

En (Kryszkiewicz, 1998b) se demuestra que RAMM ⊆ RAR facilitando el análisis de contextos con un mínimo de ítems, con el inconveniente de que este tipo de regla no permite obtener el total de reglas que satisfacen el umbral de soporte y confianza mínimo. Una RAMM confirma la expresión:

$$RAMM = \{r: X \rightarrow Y \in RAR \mid \neg\exists r': X' \rightarrow Y' \in RAR, r' \neq r, X \subseteq X', Y \subseteq Y'\} \quad (4)$$

Las RAG se describen en el apartado 2.4.

1.3 Medidas de evaluación de las reglas de asociación.

Según (Agrawal, et al., 1994), el problema en el minado de reglas de asociación es generar todas las reglas con un soporte **s** al menos mayor que el mínimo soporte especificado por el usuario ***s_min*** y con una confianza ***c*** al menos mayor que la mínima confianza ***c_min*** que especifica el usuario. El significado intuitivo de una regla de asociación está en que las transacciones en la base de datos que contienen los elementos de X también tienden a contener los elementos de Y. Por ejemplo el 98% de los clientes que compran gomas y accesorios de autos también solicitan algún servicio automotriz. Ese 98% es a lo que se

le llama confianza de la regla de asociación. Formalmente (Agrawal, et al., 1994) la confianza de la regla X→Y de un conjunto de transacciones D se define como:

$$conf(X) = \frac{|\{T \in D | X \cup Y \subseteq T\}|}{|\{T \in D | X \subseteq T\}|} \quad (5)$$

Entonces la regla X → Y está en las transacciones T con un nivel de confianza ***c*** si ***c%*** de transacciones de T que contienen a X también contienen a Y. El soporte (Agrawal R, 1994) se define como las veces que aparece un ítem X en el total de transacciones:

$$s(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|} \quad (6)$$

Para una regla de asociación X → Y el soporte se define entonces como:

$$S(X \to Y) = \frac{|\{T \in D | X \cup Y \subseteq T\}|}{|D|} \quad (7)$$

Si se supone que en las compras de una tienda el 90% de las transacciones en las que un cliente compra pan y mantequilla también compra leche, y estos tres elementos aparecen en el 5% de las transacciones, en ese caso la regla {pan, mantequilla} → {leche} tiene un nivel de confianza de 0.90 y un soporte de 0.05.

La confianza es una medida de la fortaleza de la regla y el soporte tiene un significado estadístico. Por ejemplo si la regla $X, Y \to Z$ tiene mucha más confianza que $X, A \to Z$ significa que cuando se encuentre *X* en una transacción es más probable encontrar *Z* si *X* está acompañada de *Y que de A*. En ese caso $X, Y \to Z$ es más fuerte que $X, A \to Z$ y se puede tomar más seriamente en el análisis de los datos. El soporte es una medida del significado estadístico y se puede tomar de la forma: si $S(X \to Y) \approx s(X) \times s(Y)$ entonces es probable que X e Y sean independientes y pueden ocurrir simultáneamente en transacciones, sin embargo si $S(X \to Y) \gg s(X) \times s(Y)$ ocurre el caso contrario, X e Y son dependientes entre sí.

A pesar de que las métricas anteriores son las más conocidas existen otras formas de evaluar la calidad de las reglas de asociación. La Medida de Importancia de una Regla (RIM, Rule Importance Measure) (Li, et al., 2006) es una medida objetiva del grado de interés para evaluar la importancia de una regla de

asociación. La RIM se complementa con ERIM (Enhanced Rule Importance Measure) (Li, et al., 2009) y es una medida que toma en cuenta el factor subjetivo integrando el conocimiento de dominio a la evaluación RIM de la regla.

Según (Medina, et al., 2007) la confianza tiene como limitante que no es capaz de detectar la independencia estadística ni la dependencia negativa entre consecuente y antecedente, ya que no considera el soporte del consecuente. La aplicación del Factor de Certeza (FC) en las reglas de asociación fue propuesta por (Berzal, et al., 2001) para expresar (en forma de índice) el por ciento del incremento (o decremento) de la probabilidad condicional de Y respecto a X relativo a la magnitud del intervalo definido por la probabilidad de Y de la siguiente forma:

$$FC(X \to Y) = \begin{cases} \frac{conf(X \to Y) - s(Y)}{1 - s(Y)}, & si\ conf(X \to Y) > s(Y) \\ \frac{conf(X \to Y) - s(X)}{s(Y)}, & si\ conf(X \to Y) < s(Y) \\ 0 & ,\ en\ otro\ caso \end{cases} \qquad (8)$$

Con respecto a la limitación antes mencionada de la confianza el FC es superior para determinar las dependencias positivas aunque mantiene la imposibilidad de discriminar las dependencias negativas (Medina, et al., 2007).

Autores como (Brin, et al., 1997) (Dong, et al., 1998) (Ahn, et al., 2004) proponen diversas medidas de evaluación de reglas de asociación basadas en métodos estadísticos o lógicos que pudiesen ser aplicables en contextos específicos.

1.4 Reglas de asociación generalizadas.

Los usuarios generalmente están interesados en generar reglas que estén esparcidas por varios niveles de una taxonomía (jerarquía es un); esto se debe a que en un conjunto de datos significativamente grande, las reglas de asociación que involucran conceptos a niveles muy bajos de la taxonomía no alcanzan valores de soporte altos y pueden ser descartadas reglas relevantes. Además este tipo de reglas de asociación pueden facilitar la poda de reglas redundantes o poco interesantes.

Una regla de asociación generalizada es un tipo de regla que se obtiene cuando los ítems se expresan a través de una taxonomía de conceptos. Los soportes en estos tipos de reglas se determinan considerando

que una transacción T “soporta” un ítem de una RAG si ese ítem está en T o es un ancestro de algún ítem de T. Se dice que una transacción T “soporta” el itemset X si T “soporta” cada ítem de X (Srikant, et al., 1995); entonces una RAG se puede definir como cualquier regla $X \rightarrow Y$ cuyos ítems representan conceptos de una jerarquía o taxonomía, cumpliéndose que ningún ítem de Y es concepto ancestro de X, ya que la regla de la forma $X \rightarrow ancestro(X)$ es trivial, con una confianza de 100% por lo que además sería redundante.

Por ejemplo se puede inferir que las personas que compran ropas de exteriores tienden a comprar también botas para excursión, entonces las personas que compran chaquetas compran botas de excursión y las que compran pantalones compran botas de excursión. Sin embargo, el soporte de la regla *ropa de exteriores → botas de excursión* pude no ser la suma de los soportes de *chaquetas → botas de excursión* y *pantalones → botas de excursión*; ya que algunas personas pueden haber comprado chaquetas, pantalones y botas en la misma transacción. Por lo que *ropas de exteriores →botas de excursión* puede ser una regla válida, mientras que *chaquetas → botas de excursión* y *pantalones → botas de excursión* no lo son. En el primer caso, la generalización de la regla puede no cumplir con el soporte mínimo mientras que el segundo es posible que no cumpla con la confianza mínima requerida por el usuario ya que pocas personas compran chaquetas y botas de excursión, pero muchas de las que compran algún tipo de ropas de exteriores compran botas de excursión, así muchas asociaciones significativas se pasan por alto si se restringe el minado solo a los últimos niveles de abstracción de la taxonomía (Srikant, et al., 1995).

1.5 Extracción de reglas de asociación.

El algoritmo propuesto por (Agrawal, et al., 1993) fue el AIS (por las iniciales de cada uno de los autores), este consiste en hacer pases múltiples sobre el conjunto de datos donde en cada pase se computan los nuevos itemsets que son generados. En cada lectura de los datos se calcula el soporte de los itemsets generados y si cumplen con el mínimo requerido pasan a ser itemsets candidatos, de lo contrario son desechados. El algoritmo SETM (Set Oriented Mining) (Houtsma, et al., 1993) fue propuesto para el minado de reglas de asociación usando operaciones relacionales en ambientes de bases de datos relacionales, motivado por el deseo de utilizar el Lenguaje Estructurado de Consulta (SQL por sus siglas en inglés).

En (Agrawal, et al., 1994) se proponen dos algoritmos, el Apriori y el AprioriTid así como un tercero, que es una combinación de los anteriores, llamado AprioriHybrid. Los dos primeros reducen la cantidad de elementos que se generan en cada pasada. El AprioriTid utiliza una codificación de la base de datos para cada pasada en lugar de utilizar la base de datos completa. Como se mencionó el AprioriHybrid combina características de los anteriores algoritmos, se emplea Apriori para la completa explotación del conjunto de datos en pases iniciales, pero cambia a AprioriTid cuando se requiere un tamaño de codificación para encajar en la memoria disponible.

El algoritmo Partition presentado por (Savasere, et al., 1995) y su modificación por (Toivonen, 1996) proponen la idea de seleccionar una muestra aleatoria en dos pases y uno respectivamente al conjunto de datos, y utilizarla para determinar reglas de asociación representativas que es muy probable que ocurran también en la base de datos completa.

Cuando son usadas técnicas estándares para el minado de reglas de asociación (como el algoritmo Apriori y sus variantes) pueden causar consumo de recursos exponencial en el peor de los casos. Por lo tanto, puede tomar mucho tiempo de la Unidad Central de Procesamiento (CPU por sus siglas en inglés) el proceso de extracción de las reglas de asociación. El minado de forma exhaustiva de todas las reglas que satisfacen la restricción de soporte mínimo definido, puede dar lugar a la generación de un número excesivo de reglas. Entonces, el usuario final tendrá que determinar cuáles son las reglas que valen la pena utilizar, por lo tanto, es mayor el número de las reglas de asociación derivadas y más difícil su revisión (Triantaphyllou, 2010). Si la base de datos es muy densa, entonces la situación anterior puede ser aún peor. El tamaño de la base de datos (léase también cualquier formato de entrada de los datos) también juega un papel vital en los algoritmos de minería de datos (Toivonen, 1996).

El núcleo de la mayoría de los algoritmos usados en la actualidad es el algoritmo Apriori. Por lo tanto, es muy conveniente para desarrollar un algoritmo que tiene complejidad polinómica y aún así ser capaz de encontrar reglas de buena calidad; ventaja que aprovecha el algoritmo RA1 dando lugar a variaciones como la ARA1 con excelentes resultados en comparación con sus predecesores (Triantaphyllou, 2010).

1.6 Utilización de conocimiento previo para la extracción de reglas de asociación generalizadas.

En el campo de la minería de datos una cuestión importante es cómo incorporar el conocimiento previo del usuario en el proceso de obtener nuevos patrones y comportamiento de los datos. Una dificultad de esta práctica (Zhou, et al.) es la representación cualitativa y cuantitativa del análisis de las reglas obtenidas. Ya que la mayoría de los investigadores y profesionales se dedican al desarrollo de herramientas y métodos de descubrimiento que sean capaces de extraer patrones estadísticamente fuertes que satisfagan un criterio, generalmente soporte y nivel de confianza. Normalmente después de un proceso exhaustivo de minería de datos que puede tener un rango de duración de horas a días, el usuario puede quedar "confundido" por la abundancia de reglas obtenidas, cuando solo un pequeño grupo de ellas son las que realmente le interesan.

(Zhou, et al.) proponen un método para aprovechar el conocimiento previo de los usuarios en la generación de reglas de asociación. Mediante el uso de una red Bayesiana consistente en un grafo directo y acíclico (DAG por sus siglas en inglés) para representar las preferencias y conocimientos previos de los usuarios y utilizarlos en el proceso de descubrimiento para obtener el tipo de reglas que le interesan al usuario.

Otros enfoques proponen representar el conocimiento previo mediante el uso de ontologías (ver epígrafe 1.8).

1.7 Ontologías.

El término ontología lo introdujo Thomasius y Wolf, en el siglo XVII, para designar, precisamente, en la filosofía primera de Aristóteles, la parte de la metafísica que *estudia el ser en cuanto tal*. Sin embargo, el origen epistemológico del término se encuentra mucho antes, en la disciplina filosófica surgida con Parménides (h. 540-h. 450 a.C.), cuyo objetivo era también, el estudio del ser (Céspedes, 2005).

Las ontologías se ocupan de las categorías generales del ser, entendidas de forma abstracta, de las que participa el ser concreto. En el entorno de la hipertextualidad, la ontología ha sido definida como: una representación explícita y formal de una conceptualización compartida. El término conceptualización corresponde a una parte del mundo o universo que es objeto de tratamiento, como una forma de entender

y describir un dominio, por lo que constituye un modelo abstracto de algún fenómeno en el mundo; se construye a partir de identificar los conceptos que componen un dominio del conocimiento, y las relaciones relevantes establecidas entre dichos conceptos, por lo que la base de toda ontología es una taxonomía (o clasificación) de conceptos (Céspedes, 2005).

En la ciencia de la computación y la ciencia de la información, una ontología es una representación formal de un conjunto de conceptos dentro de un dominio y de relaciones entre dichos conceptos. Es usada para razonar acerca de propiedades de un dominio, y quizás usada para definir el dominio. Una ontología brinda un vocabulario compartido, el cual puede ser usado para modelar un dominio, el tipo de objetos y/o conceptos que existe, y sus propiedades y relaciones (Concepción, et al.).

Una ontología define un vocabulario común para investigadores que necesitan compartir la información en un dominio y contiene definiciones de conceptos básicos y sus relaciones que no pueden ser interpretadas por una máquina. Según los criterios analizados, una ontología no es más que una forma de modelar términos manejables, lo que hace posible una comprensión común de la estructura de la información, es decir que los términos con los que se trabajan tengan el mismo significado y se hable en un lenguaje común. Una ontología está formada por clases o conceptos, propiedades o atributos de las clases y relaciones entre clases. Teniendo en cuenta lo anterior las ontologías pueden ser aplicadas en varios contextos (Céspedes, 2005):

- Representar explícitamente y a través del lenguaje natural, el conocimiento implícito y el operacional de los grupos que integran la estructura, permitiendo la comprensión de las mismas.
- Posibilitar el aprendizaje organizacional.
- Rehusar el conocimiento representado para el desarrollo de servicios y productos de información, especialmente en procesos de búsqueda y recuperación.
- Posibilitar una cultura de colaboración a escala grupal y/o organizacional a través del proceso de consenso en la construcción de la ontología.

1.7.1 Componentes de una Ontología

Los componentes de una ontología varían de acuerdo al dominio de interés y a las necesidades de los desarrolladores. Por lo general los componentes son los siguientes (Nuñez, 2007):

- **Clases:** Las clases son la base de la descripción del dominio de conocimiento de la ontología. Generalmente se organizan en taxonomías a las que usualmente se les aplican mecanismos de herencia.
- **Relaciones:** Representan las interacciones entre los conceptos del dominio. Las ontologías por lo general contienen relaciones binarias; el primer argumento llamado dominio y el segundo rango.
- **Funciones:** Son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de una ontología.
- **Instancias:** Representan objetos determinados de un concepto.
- **Taxonomía**: Conjunto de conceptos organizados jerárquicamente. Definen las relaciones entre los conceptos pero no los atributos de éstos.
- **Axiomas:** Se usan para modelar sentencias que son siempre ciertas. Los axiomas permiten, junto con la herencia de conceptos, inferir conocimiento que no esté indicado explícitamente en la taxonomía de conceptos.
- **Propiedades (Slots)**: Son las características o atributos que describen a los conceptos.Para un concepto dado, las propiedades y las restricciones sobre éstos son heredadas por las subclases y las instancias de la clase.

1.7.2 Tipos de Ontologías

De acuerdo al nivel de generalización se distinguen cuatro tipos de ontologías (Garea, et al., 2009):

- Las ontologías de alto nivel describen los conceptos generales como el espacio, tiempo, materia, objeto, evento, acción, los cuales son independientes de un problema o dominio en particular.

- Las ontologías de dominio y de tareas describen, respectivamente, el vocabulario relacionado a un dominio genérico, por ejemplo, medicina, o una tarea o actividad genérica, como diagnosis, especializando los términos introducidos en la ontología de alto nivel.
- Las ontologías de aplicación describen conceptos dependiendo de un dominio y de una tarea en particular, la cual es una especialización de ambas ontologías relacionadas (ontología de dominio y ontología de tarea).

1.8 Utilización de ontologías en la minería de datos.

Las ontologías fueron introducidas en las técnicas de minería de datos alrededor del año 2000. Dadas sus características que les permiten ser una eficaz herramienta de comunicación automática entre computadoras y/o humanos para el razonamiento, representación y reutilización del conocimiento, se han convertido en una precondición necesaria para una eficiente aplicación de estas técnicas. Las Ontologías de Dominio de Conocimiento o de Conocimiento Previo organizan los dominios de conocimiento. Además de ser de gran importancia en varios niveles del proceso de descubrimiento de conocimiento, las Ontologías de Metadatos describen el proceso de construcción de items y las de Procesos de Minería de Datos codifican el proceso de minado, contribuyendo a la elección de la tarea más apropiada según el problema a resolver (Marinica, et al.).

La relación entre minería de datos y ontologías según (Nigro, et al., 2008) es bidireccional. De ontologías a minería de datos se puede incorporar conocimiento al proceso a través de ontologías, además de la interpretación y validación del conocimiento minado. De minería de datos a ontologías, se pueden incluir como una instancia de entrada de información o como el resultado del proceso, entonces el análisis se hace sobre la ontología.

1.9 Utilización de ontologías en las reglas de asociación.

Según (Xiong, et al., 2010) el mínimo soporte y confianza de un número de reglas de asociación fuertes extraídas hace que el análisis y consecuente utilización de estas sea difícil. (Xiong, et al., 2010) proponen un índice de calidad de las reglas de asociación basado en la integración de aspectos objetivos y subjetivos con el fin de cuantificar la calidad de las reglas de asociación basados en una Ontología de Dominio. En la construcción de la ontología se tienen en cuenta el dominio de conocimiento, y el conocimiento previo, combinado con el conocimiento experto y la familiarización de los usuarios con el

conjunto de datos. El minado de las reglas de asociación se hace bajo la guía de la ontología y las reglas de asociación derivadas cumplen con los parámetros deseados además de expresar el propósito de los usuarios.

La metodología de emparejamiento AROMA (Association Rule Ontology Matching Approach) (Jerome, et al., 2007) tiene como objetivo encontrar las relaciones de inclusión (y también de equivalencia) entre entidades (clases o propiedades) de dos estructuras jerárquicas diferentes proveídas por datos textuales. Las ontologías se definen como vocabularios estructurados que describen las relaciones entre conceptos. El lenguaje OWL permite describir conceptos (llamados clases) y organizarlos en taxonomías (de clases y de propiedades) mediante el uso de la relación inclusión.

1.10 Trabajos relacionados.

(Agrawal, et al., 1993) Introducen el problema del minado de reglas de asociación en grandes conjuntos de datos, así como la necesidad de generarlas con un mínimo soporte y confianza que aseguren la calidad de las mismas. Los principales problemas del minado de reglas de asociación son por una parte la calidad de las reglas generadas, así como la complejidad computacional de los algoritmos de minado. Los trabajos en la etapa de pos-procesamiento tratan de dar solución al primer caso y los que trabajan directamente sobre los algoritmos o en la etapa de pre-procesamiento dan solución a la segunda problemática.

Los principales trabajos relacionados con el minado de reglas de asociación están enfocados en aumentar la calidad de las reglas generadas, así como disminuir la complejidad computacional de los algoritmos de minado. La utilización de conocimiento previo del dominio particular puede ser relevante para disminuir la complejidad computacional de los algoritmos de minado de reglas de asociación. Según (Chen, et al., 2003) la utilización de una ontología de conocimiento previo para subir o elevar el nivel de generalización de los ítems posibilita la obtención de reglas de asociación generalizadas con un mayor soporte, este procedimiento es conocido como *“Raising”* (subir o elevar). De acuerdo con el caso de estudio propuesto por (Zhou, et al., 2007) los valores de soporte y confianza de las reglas de asociación siempre tienden a aumentar si se aplica el método *“Raising”.*

1.11 Conclusiones parciales

En este capítulo se analizaron los principales elementos referentes a la minería de reglas de asociación, como parte de las técnicas de minería de datos. Se relacionaron los conceptos asociados al tema que tienen influencia para alcanzar los objetivos de la investigación. Se hace un estudio en el área del conocimiento referida a las ontologías, sus principales aplicaciones en el campo de la minería de datos y particularmente en la extracción de reglas de asociación. Concluyendo que estas son de gran interés para representar el conocimiento previo del negocio en la etapa de pre-procesamiento de los algoritmos de extracción de reglas de asociación. Se mencionan además las principales técnicas de extracción de reglas de asociación y su complejidad computacional, centrándose en el método Apriori como base de la mayoría de los algoritmos de minado de reglas.

2. CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL ALGORITMO

2.1 Introducción.

En las transacciones del despacho de mercancía de la aduana se procede a cobrar el impuesto asociado a la operación que se desea hacer de acuerdo al tipo de carga. La declaración de mercancías como documento oficial contiene o debe contener la información verídica del contenido de las cargas y además las operaciones que se desean realizar sobre ellas, sirviendo de base para determinar el monto a pagar. En ocasiones la declaración de mercancías no es 100% real de acuerdo a la existencia física. Esto ocurre con el objetivo de burlar las directivas de pago de los regímenes aduaneros.

La clasificación del riesgo de fraude mercantil se ocupa de determinar qué o cuáles declaraciones de mercancías son las más propensas a tener problemas en las transacciones aduaneras. Un ejemplo sería el conjunto de datos siguiente (Tabla 1) donde se almacenan los datos referentes al origen o país donde se efectúa más del 50% de la manufactura de la mercancía, el campo “mercancía” muestra, para mejor comprensión, la descripción correspondiente a la codificación establecida en el Sistema Armonizado de Clasificación de Productos (SACLAP[1]), aunque en la práctica se almacena el código determinado; el tipo de moneda con que se procede a efectuar el pago, el importe total a pagar del impuesto o tasa que ampara esa declaración, la cantidad de bultos que contiene la carga, el monto de los gastos por flete en USD, el peso bruto total del producto, el tipo de declaración de mercancía que puede ser completa, provisional, anticipada o incompleta; almacenándose solo el identificador correspondiente a cada clasificación, la nacionalidad del medio de transporte en el que vienen los productos, los datos del Agente de Aduana o Apoderado debidamente registrado como declarante y por último si fue o no mal clasificada.

[1] Constituye una nomenclatura internacional orientada a los aranceles de aduana y a las estadísticas del Comercio Exterior.

origen	mercancia	t_mon	monto	c_bultos	flete	peso	tipo_dm	nac	decl	m_clas
colombia	carne porcina congelada deshuesada	USD	20000	1	612	2000	2	brasil	1	si
canada	carne pollo congelado en trozos	USD	1000	2	520	500	1	canada	5	no
colombia	carne porcina fresca deshuesada	USD	55000	5	460	8000	2	EEUU	2	si
brasil	carne porcina fresca en trozos	USD	4500	3	702	1000	3	brasil	2	si
argentina	carne pollo fresca sin trocear	USD	8000	6	821	6000	1	uruguay	4	no
venezuela	carne porcina congelada en trozos	USD	2000	8	963	25000	2	venezuela	1	si
brasil	carne equina fresca en trozos	USD	4500	3	426	1000	3	chile	2	no
mexico	carne pollo congelado sin trocear	USD	150000	10	148	55500	1	EEUU	2	si
brasil	carne equina fresca en trozos	USD	4500	3	426	1000	3	chile	2	no

Tabla 1: Muestra de datos asociados a las declaraciones de mercancías aduanales.

De la tabla anterior se podrían extraer varias reglas de asociación, entre ellas por ejemplo:

- ***dm-mercancía**: carne porcina congelada deshuesada, **dm-origen**: colombia →**dm-m_clas**: si*
- ***dm-mercancía**: carne porcina fresca deshuesada, **dm-origen**: colombia →**dm-m_clas**: si*

Sin embargo estas reglas serían descartadas ya que el elemento *dm-mercancía* en ambos casos el soporte es de 0,1 por lo que no cumpliría con un umbral mínimo aceptable para estos datos que sería a partir de 0,3. Una solución a este problema sería generalizar las reglas de asociación, en particular el atributo mercancía que está afectando el soporte de las mismas, si se lleva a cabo este proceso, teniendo en cuenta que los elementos *dm-origen* y *dm-m_clas* tienen los mismos valores se obtendría la regla de asociación siguiente:

- ***dm-mercancía**: carne porcina, **dm-origen**: colombia →**dm-m_clas**: si*

En este caso particular esta regla no sería descartada por no cumplir con el soporte mínimo ya que al generalizar el atributo mercancía el soporte de este aumentaría a 0,4.

De acuerdo a la codificación de la instancia de entrada utilizada en los algoritmos de extracción de reglas de asociación, cada atributo se convierte en sus respectivos valores. Según el criterio anterior el atributo mercancía se convierte en todos los valores distintos asociados a él, dejando de existir como mercancía y transformándose en 9 elementos y así sucesivamente con todos los atributos de la tabla, aumentando considerablemente el tamaño de la instancia, en este caso particular la codificación sería la siguiente:

$$dm - origen: colombia, \dots, dm - origen: n, dm - mercancia: carne\ porcina\ congelada\ deshuesada, \dots, dm - mercancia: n, \dots, dm - atributo\ n: valor\ n$$

En el proceso de extracción de reglas de asociación las particularidades de los atributos generalmente no aportan información relevante ni afectan la calidad de las reglas generadas, sin embargo por ser tan específicos, en la mayoría de los casos no verifican las ocurrencias mínimas sobre el total de transacciones para que cumplan con los umbrales de soporte y confianza mínimos requeridos. A partir de los valores que toma el atributo mercancía, se pueden generalizar mediante el uso de la taxonomía de conceptos de una ontología de dominio específico de la siguiente forma:

- ***carne porcina*** *congelada deshuesada,* ***carne porcina*** *fresca en trozos,* ***carne porcina*** *congelada en trozos* y ***carne porcina*** *fresa deshuesada* en ***carne porcina****.*
- ***carne pollo*** *congelada en trozos,* ***carne pollo*** *fresca en trozos* y ***carne pollo*** *congelada sin trocear* en ***carne pollo***.

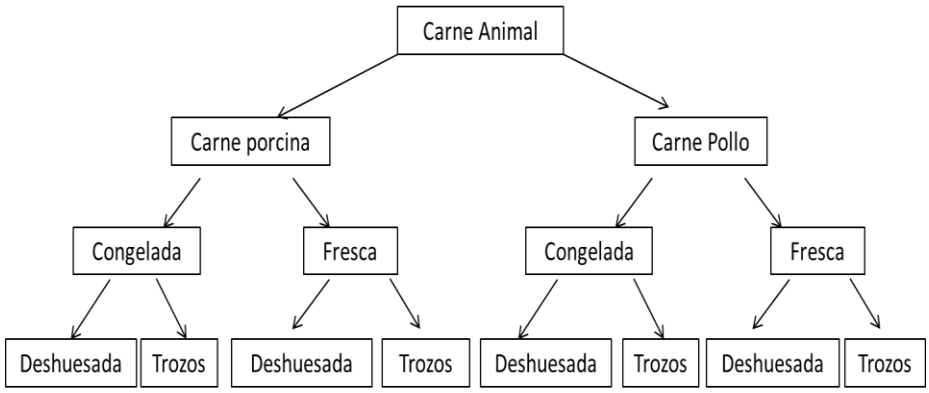

Figura 1: Fragmento de taxonomía de la ontología de mercancías.

Reduciendo la transformación del valor mercancía de 9 a 3 elementos lo que disminuye el tamaño de la instancia de entrada y por tanto la complejidad computacional de la extracción de las reglas de asociación.

origen	mercancia	t_mon	monto	c_bultos	flete	peso	tipo_dm	nac	decl	m_clas
colombia	carne porcina	USD	20000	1	612	2000	2	brasil	1	si
canada	carne pollo	USD	1000	2	520	500	1	canada	5	no
colombia	carne porcina	USD	55000	5	460	8000	2	EEUU	2	si
brasil	carne porcina	USD	4500	3	702	1000	3	brasil	2	no
argentina	carne pollo	USD	8000	6	821	6000	1	uruguay	4	si
venezuela	carne porcina	USD	2000	8	963	25000	2	venezuela	1	no
brasil	carne equina	USD	4500	3	426	1000	3	chile	2	no
mexico	carne pollo	USD	150000	10	148	55500	1	EEUU	2	si

Tabla 2: Muestra de datos generalizados asociados a las declaraciones de mercancías aduanales.

Con una muestra pequeña como la anterior, determinar la generalidad que existe entre los valores de los atributos puede obtenerse de forma manual; en la práctica las bases de datos de la aduana contienen millones de transacciones y cientos de clasificaciones de mercancías haciendo imposible realizar el procedimiento anterior manualmente. Por lo tanto se requiere de un método que informatice el proceso de generalización.

2.2 Propuesta de solución.

Se ha diseñado un algoritmo para la generalización de elementos a partir del método *Raising* (Zhou, et al., 2007) y con la utilización del conocimiento previo expresado en una ontología. Este método propone utilizar una representación del dominio específico para obtener los niveles de abstracción de los elementos de la base de datos y poder generalizarlos. Estos autores definen el método *Raising* como se muestra a continuación:

Una operación R^k puede ser llamada *Raising* (elevar) una tupla a un nivel k si dada una tupla $T =< N_s, D >$ donde N_s es un conjunto $\{N_1, N_2, \dots N_n\}$ de valores de un atributo y D el resto de atributos o elementos de información. N_i está contenido en una taxonomía de conceptos en una ontología O. En O cada N_i es único, correspondiéndole un número de $m_i (m_i \geq 1)$ ancestros $< A^k_{i\_1}, A^k_{i\_2}, \dots A^k_{i\_m_i} >$, todos a nivel k. Si N_i está a un nivel mayor que k, entonces no tiene ancestros a nivel k, en ese caso $A^k_{i\_1} = N_i$. De otra forma R^k es la operación que toma T como entrada y devuelve la tupla:

$$T^k = R^k(T) =< A^k_{1\_1}, A^k_{1\_2}, \dots A^k_{1_{m_1}}, A^k_{2_1}, A^k_{2_2}, \dots A^k_{2_{m_2}}, A^k_{n_1}, A^k_{n_2}, \dots A^k_{n_{mn}}, D >$$

Para cada T en O excepto para la tupla $< raíz(O), D >$, entendiéndose que para la raíz no existen ancestros posibles por lo que la operación $R^k(< raíz(O), D >)$ queda indefinida. Durante el proceso se tiene en cuenta la duplicación de datos, así cada $A^k_{i\_j}$ que aparezca en T^k es único y todos los duplicados son removidos.

2.2.1 Algoritmo.

El procedimiento desarrollado depende de cuatro parámetros suministrados por el usuario en un formato definido para su correcto funcionamiento. El primer parámetro es una ontología *O* de dominio específico que debe contener la taxonomía de conceptos correspondiente. El tipo de formato de la ontología debe ser *OWL/XML* y la nomenclatura de las clases debe coincidir con la nomenclatura de los datos utilizada en la base de transacciones que se desea generalizar. El dominio de la ontología puede estar enmarcado en un solo atributo de la base de transacciones, por lo que se podrá generalizar ese atributo en particular o puede abarcar todos los atributos contenidos en la base de datos.

La base de transacciones *D* mencionada anteriormente es un archivo de texto con todas las transacciones de la base de datos original correspondiente al dominio descrito en *O*. Cada elemento de *D* debe estar separado por coma y tener el formato ***tabla-atributo:valor***, la primera línea debe contener todos los elementos de la base de transacciones y posteriormente las transacciones a razón de una transacción por línea. Los dos parámetros restantes son un número entero que representa el nivel *k* al que el usuario desea generalizar el atributo *A* en la base de transacciones. Los niveles son tomados desde el nodo raíz de la taxonomía de *O* comenzando por cero hasta los nodos hojas.

El algoritmo de generalización utiliza los parámetros descritos anteriormente para generalizar los valores de los elementos contenidos en *D*. La taxonomía descrita en *O* permite verificar el nivel al que se encuentra cualquier elemento. Dado el nivel al que se desea generalizar un valor se obtiene de *O* el valor del elemento correspondiente al nivel especificado. El valor generalizado pasa a reemplazar todos los valores del atributo que se encuentren a niveles inferiores que el definido por el usuario. Luego de la ejecución de este proceso se obtendrá como salida la base de transacciones modificada *D'* donde todos los valores de *A* se encontrarán a un nivel de abstracción igual o superior al nivel *k* del parámetro de entrada especificado por el usuario.

A continuación se presenta la descripción del algoritmo propuesto:

Algoritmo de generalización

Entradas: O, D, k, A

1. Obtener de O los valores generalizados al nivel k de los valores de A.
2. Reemplazar en D los valores de A con nivel inferior a k por los nuevos valores generalizados al nivel k.

Salida: D'

2.3 Implementación.

La implementación del algoritmo de generalización se forma de dos componentes fundamentales: la interacción con el fichero *.owl en el formato *OWL/XML* que contiene la ontología y la interacción con el fichero que contiene el conjunto de datos al que se le va a aplicar la generalización. Cada uno de estos elementos son combinados para lograr el correcto funcionamiento del algoritmo.

Para el trabajo con el fichero *.owl solo interesan las relaciones taxonómicas que en él se definen. Teniendo en cuenta que una taxonomía en su conjunto grafo acíclico dirigido, la mayoría de las operaciones que se realizan son recursivas, siendo este tipo de operación la más recomendada para estas estructuras. Cada declaración de una clase es tomada como un nodo y las declaraciones de subclase conectan cada clase con su antecesor lo que brinda el mecanismo para recorrer toda la estructura en forma de un Tipo de Dato Abstracto (TDA) grafo.

```
1   <Declaration>
2       <Class IRI="#TELECOMUNICACIONES"/>
3   </Declaration>
4   <Declaration>
5       <Class IRI="#REDES_INALAMBRICAS"/>
6   </Declaration>
7   <SubClassOf>
8       <Class IRI="#TELECOMUNICACIONES"/>
9       <Class IRI="#REDES_INALAMBRICAS"/>
10  </SubClassOf>
```

Figura 2: Fragmento de un fichero con el formato OWL/XML donde se declaran 2 clases y la relación taxonómica entre ellas.

Las especificaciones descritas anteriormente facilitan las operaciones que son necesarias realizar sobre este fichero: obtener los antecesores a un nivel específico de un elemento y por consiguiente obtener el nivel de un elemento.

El segundo componente es el encargado de realizar los procedimientos sobre el conjunto de datos contenidos en un fichero con el formato y las restricciones descritas en el apartado 2.2.1. Las operaciones definidas para este componente son en su mayoría cíclicas debido a que se hace necesario recorrer todas las líneas del fichero y dentro de cada línea, cada elemento que la compone. Simultáneamente a este recorrido se construye el fichero de salida del proceso de generalización de forma tal que se lee una línea del fichero de entrada, se modifica dicha línea de acuerdo a los parámetros de generalización y se escribe la línea modificada en el fichero de salida. Los elementos generalizados que reemplazarán los elementos del fichero de entrada en el fichero de salida se obtienen mediante la función de obtener los antecesores a un nivel específico en el componente de interacción con el fichero *.owl.

El procedimiento descrito anteriormente posibilita un ahorro considerable de memoria ya que no necesita cargar todos los elementos del fichero de entrada y la salida en memoria, además de trabajar con datos pequeños en cada iteración lo que posibilita que las operaciones que se realicen sobre estos sean mucho más rápidas.

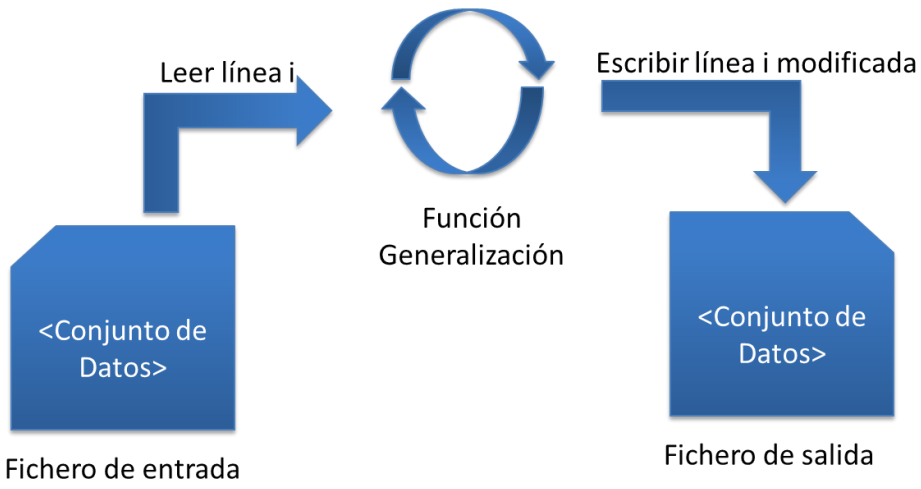

Figura 3: Funcionamiento de la función Generalizar.

2.4 Conclusiones parciales

En la investigación se ha desarrollado un marco integrado para la generalización de ítems en la etapa de pre-procesamiento del minado de reglas de asociación basado en una estructura taxonómica contenida en una ontología de dominio específico. El marco permite definir el nivel de abstracción al que se desea generalizar, permitiendo pre-procesar el conjunto de datos utilizado como entrada a los algoritmos de extracción de reglas de asociación. El mismo, cuenta con una arquitectura en la que el motor de minería de datos con el dominio de especificación son independientes. La arquitectura propuesta permite cambiar segmentos de código en el algoritmo o extender el dominio de la ontología, sin afectar el nivel de aplicación ni el conjunto de datos.

3. CAPÍTULO 3: VALIDACIÓN

3.1 Evaluación y validación.

Evaluar y validar que la solución desarrollada y las afirmaciones referentes a la solución sean aceptables para la comunidad investigadora es un paso fundamental de cualquier investigación. Los siguientes métodos proporcionan una vía para realizar la evaluación y validación de una solución desarrollada (Vaishnavi, et al., 2008):

- Demostración

  Se basa en demostrar que una solución es realizable y válida en una situación predefinida. Se aplica desarrollando una o varias situaciones particulares de un problema y trabaja para ese conjunto de situaciones predefinidas. El método es especialmente relevante cuando la demostración de una solución en sí misma se considera una contribución, consta de dos etapas fundamentales: la construcción de la solución que permite afirmar que la solución es realizable y la demostración de que la solución es razonable para un conjunto de situaciones predefinidas. Como resultado, la demostración puede exponer las deficiencias de la solución o por el contrario que es viable y aceptable. Las pruebas exhaustivas aumentan la confianza en la solución, si las situaciones de prueba están diseñadas apropiadamente, entonces la construcción de la solución y sus pruebas para estas situaciones puede demostrar la validez de la misma.

- Experimentación

  La experimentación es utilizada para validar o rechazar un conjunto de hipótesis relacionadas con las afirmaciones acerca de la solución. Estas hipótesis no pueden ser probadas matemática o lógicamente, por lo que es necesario generar un conjunto de datos del sistema y luego utilizar esta información para validar o rechazar las hipótesis. La experimentación ayudará a establecer resultados asociados con la solución del problema de investigación en situaciones donde la recogida y análisis de datos es el único método factible de validación.

- Simulación

  La simulación es usada para evaluar y validar una solución a un problema complejo tal que dicha solución no pueda ser demostrada matemáticamente como válida. La evaluación y validación de la solución en el ámbito de la vida real es poco viable y costoso, entonces el problema y su solución deben ser modelados con precisión en una computadora. Para la realización de la simulación es

necesario contar con el modelo conceptual del problema y su solución para que sean simulados en una computadora y un conjunto de datos de prueba iniciales. Este método ofrece una forma razonable y rentable de evaluación y validación de una solución y brinda la alternativa de poner a prueba la solución en la vida real lo que puede ser a la vez costoso y consume mucho tiempo, o tal vez ni siquiera sea factible.

- Uso de métricas

El uso de métricas se propone para evaluar el desempeño de la solución y para probar o argumentar las hipótesis que se han hecho en relación con el rendimiento de la solución. Es necesario para su correcta aplicación determinar si existen o no las métricas que son apropiadas para medir el rendimiento de la solución y comparar los resultados. Si tales parámetros no existen, entonces es necesario determinar si existen o no las métricas para medir el desempeño de problemas similares al problema a evaluar. En tal caso, se necesita argumentar que el uso de las métricas elegidas es una forma razonable de evaluación y validación de una solución.

- Marcadores

Los marcadores son utilizados para demostrar que una solución tiene un rendimiento razonable o es mejor que alguna otra solución disponible. Es usado generalmente cuando no hay métricas disponibles para medir el rendimiento de la solución y se hace necesario probar que la solución desarrollada es superior a otras soluciones. Si no existen marcadores disponibles para validar la solución es efectivo crear un escenario o varias clases de escenarios para evaluar la solución y demostrar su superioridad a otras soluciones disponibles. Para la aplicación de este método es necesario identificar el marcador a usar para la evaluación y validación y de no existir crear uno propio. Los marcadores proporcionan una vía objetiva de evaluación o de comparación de la solución.

- Razonamiento lógico

El razonamiento lógico es generalmente usado para argumentar la validez de la solución desarrollada y no es posible utilizar una prueba matemática formal para establecer la validez de la solución. El razonamiento lógico es válido también en contextos donde el problema puede ser demasiado complejo, o puede que no sea posible formular el problema y los criterios de solución en un marco formal. Las construcciones y los supuestos del problema son, sin embargo, lo suficientemente precisos para que pueda construirse un argumento lógico basado en las hipótesis de la solución. Este método

podría servir como un suplemento o como alternativa a la evaluación experimental. El razonamiento lógico está compuesto por 3 fases fundamentales: la identificación de los supuestos (axiomas), identificación de las reglas (deducciones) y la construcción del modelo lógico.

- Pruebas matemáticas

  Las pruebas matemáticas consisten en demostrar matemáticamente las afirmaciones que se hacen acerca de la solución que se ha desarrollado. Las afirmaciones hipotéticas para la solución deben expresarse cuantitativamente, los aspectos esenciales del problema y la solución se pueden expresar formalmente en un sistema lógico cerrado. Este modelo ofrece la forma más fuerte de validación de las afirmaciones hechas sobre la solución.

La aplicación de estos métodos varía de acuerdo a su idoneidad y viabilidad con la que se puede establecer la validez de una solución. La demostración provee la forma más sencilla de la validación. Puede, sin embargo, ser adecuado si la solución es novedosa y resuelve un problema para el cual no existe ninguna solución previa. Por otro lado, las pruebas matemáticas constituyen la forma más fuerte de validación. La certeza del razonamiento lógico depende en gran medida de la precisión de sus argumentos y suposiciones. La experimentación y simulación son útiles cuando el problema es complejo y es inviable efectuar una demostración matemática. El uso de métricas y la evaluación comparativa, mecanismos para cuantificar afirmaciones respecto a una solución, son generalmente útiles cuando se utiliza en la experimentación y la simulación (Vaishnavi , et al., 2008).

El uso de unos u otros métodos de validación está relacionado en gran medida con las características del problema estudiado y de los parámetros requeridos por la comunidad de investigadores como alternativas válidas de aprobación de la materia en cuestión. La presente investigación utiliza el método de "Demostración" como mecanismo de validación. Las características del trabajo se adaptan correctamente a este método siendo ideal para corroborar la validez de la solución. La "Demostración" constituye el patrón de validación más utilizado en las publicaciones científicas para el área de las ciencias de la computación de acuerdo a los trabajos de (Shaw, 2002), (Cañete, 2002), (Shaw, 2003).

3.2 Descripción de la demostración

Para la demostración de la solución se proponen tres casos de pruebas sobre conjuntos de datos de varias dimensiones y generalizando a diferentes niveles de abstracción de la taxonomía de tipos de

mercancías que se encuentra descrita en la ontología pasada por parámetro. En cada uno de los casos se evaluará el tiempo de ejecución además de la reducción de la cantidad de ítems y su repercusión en el tamaño de la instancia de entrada al algoritmo de extracción de reglas de asociación.

3.2.1 Recursos utilizados en la demostración.

El hardware disponible para realizar la evaluación y validación de la solución es una computadora ACPI Multiprocessor PC con una motherboard Intel Rogers City DG965RY, procesador DualCore Intel Core 2 Duo E4500 a 2.20GHz y una memoria DDR2 SDRAM Kingston 9905320-007.A00LF con 1GB de capacidad. El sistema operativo sobre el que se realizaron los casos de prueba es OpenSuse v11.4 con una arquitectura de 32 bits. La implementación de la solución se desarrolló sobre la plataforma Java (TM) Platform Standard Edition Runtime Environment Version 6 y el entorno integrado de desarrollo NetBeans en su versión 6.9.

3.2.2 Conjuntos de datos

Los conjuntos de datos seleccionados para la validación de la solución fueron extraídos de las bases de datos reales de la aduana relacionados con las declaraciones de mercancías. Cada conjunto de datos contiene 33 atributos lo que equivale al mismo número de columnas. Para hacer un análisis del rendimiento del algoritmo de generalización se incrementó el número de tuplas extraídas para conformar cada conjunto, de un número inicial de 1000 hasta 10000 tuplas pasando por un conjunto intermedio de 3000 tuplas. Para el correcto funcionamiento del procedimiento cada conjunto de datos fue convertido a transacciones con el formato de entrada requerido por el algoritmo de generalización (ver apartado 2.1) lo que trae como consecuencia el aumento de la cantidad de columnas que se muestra en la Tabla 3. A continuación (Tabla 3) se presentan las características (cantidad de ítems, cantidad de transacciones, tamaño de la instancia y el tamaño del fichero en kb) de cada uno de estos conjuntos de datos.

	Ítems	Transacciones	Tamaño instancia	Tamaño (kb)
Conjunto de datos 1	442	1000	195364×10^3	411
Conjunto de datos 2	615	3000	1134675×10^3	1205
Conjunto de datos 3	1003	10000	1006009×10^4	3896

Tabla 3: Características de los conjuntos de datos utilizados en los casos de pruebas.

Acorde con los tipos de mercancías relacionados en los conjuntos de datos se utilizó una ontología basada en la clasificación de mercancías vigente en la aduana. Este parámetro se mantendrá constante para todos los conjuntos de datos. Para obtener un resultado comparativo de acuerdo a la reducción de ítems con respecto al nivel de generalización de los elementos se propone realizar la generalización de cada uno de los conjuntos de datos a diferentes niveles.

La clasificación de mercancías vigente en la aduana permite crear una ontología que contenga una taxonomía de acuerdo al tipo de mercancía basada en la enmienda SACLAP. Teniendo en cuenta que la taxonomía disponible para realizar la evaluación y validación cuenta con 4 niveles; de ellos el cuarto nivel corresponde a los elementos que están en los conjuntos de datos. Se propone generalizar los elementos al mayor nivel de abstracción posible pasando como parámetro k=1, al nivel medio k=2 y a un nivel bajo k=3.

3.2.3 Casos de prueba

Como se explicó en la sección 3.2.2 se cuentan con tres conjuntos de datos de dimensiones diferentes para realizar la evaluación y validación del algoritmo de generalización desarrollado en esta investigación. Acorde a cada conjunto de datos se desarrolló un caso de prueba como se muestra a continuación:

- Caso 1

  Para este caso particular se seleccionó el conjunto de datos 1 (ver tabla 3) que cuenta con 1000 transacciones que tienen un total de 442 ítems lo que implica el mismo número de columnas. Para este conjunto de datos con los recursos descritos en el apartado 3.2.1 el tiempo de ejecución del procedimiento de generalización para el nivel k=1 fue de 6 segundos, disminuyendo el tamaño de la instancia de entrada 86464×10^3 con respecto a su tamaño inicial. Modificando el nivel de generalización a k=2, se obtuvo una reducción de la instancia de 7426×10^4 en un tiempo de 4 segundos y en ese mismo tiempo de ejecución para el nivel de generalización k=3 la instancia de entrada se redujo en 24755×10^3.

- Caso 2

  El caso de prueba 2 se evaluó con el conjunto de datos 2 (ver tabla 3) que cuenta con 3000 transacciones para un total 615 ítems. El tiempo de ejecución del algoritmo fue de 16 segundos con los recursos descritos en el apartado 3.2.1, lo que implicó una reducción para el nivel de

generalización k=1 del tamaño inicial del conjunto de datos de 375648×10^3. Para el nivel de generalización k=2 la reducción de la instancia de entrada obtenida fue de 320352×10^3 en un tiempo de ejecución de 13 segundos. Finalmente para el parámetro de nivel de generalización k=3 la reducción de la instancia es de 104487×10^3 en un tiempo de 9 segundos.

- Caso 3

  El caso de prueba 3 se ejecutó sobre el conjunto de datos 3 (ver tabla 3) con un total de 10000 transacciones y 1003 ítems. Para este conjunto de datos el tiempo de ejecución para el nivel de generalización k=1 fue de 48 segundos, lo que implicó una reducción del tamaño de la instancia de entrada de 212128×10^4 en comparación con el tamaño inicial. Para los niveles de generalización k=2 y k=3 se obtuvieron reducciones de 17972×10^5 y 57333×10^4 respectivamente en tiempos de ejecución de 38 y 25 segundos en ese orden.

En las siguientes tablas se muestran los resultados de los casos de prueba descritos anteriormente. La columna *Reducción* contempla la diferencia del tamaño de la instancia de entrada con respecto a su tamaño inicial.

Caso	Entrada		Salida		Reducción
	Conjunto de datos	Tamaño instancia	Ítems	Tamaño instancia	
1	1	195364×10^3	330	108900×10^3	86464×10^3
2	2	1134675×10^3	503	759027×10^3	375648×10^3
3	3	1006009×10^4	891	793881×10^4	212128×10^4

Tabla 4: Casos de prueba para el parámetro k=1.

Caso	Entrada		Salida		Reducción
	Conjunto de datos	Tamaño instancia	Ítems	Tamaño instancia	
1	1	195364×10^3	348	121104×10^3	7426×10^4
2	2	1134675×10^3	521	814323×10^3	320352×10^3
3	3	1006009×10^4	909	826281×10^4	17972×10^5

Tabla 5: Casos de prueba para el parámetro k=2.

Caso	Entrada		Salida		Reducción
	Conjunto de datos	Tamaño instancia	Ítems	Tamaño instancia	
1	1	195364×10^3	413	170589×10^3	24775×10^3
2	2	1134675×10^3	586	1030188×10^3	104487×10^3
3	3	1006009×10^4	974	948676×10^4	57333×10^4

Tabla 6: Casos de prueba para el parámetro k=3.

3.3 Discusión de los resultados

En la obtención de resultados positivos en la generalización de ítems tiene mucha importancia el cubrimiento de la ontología sobre el atributo que se está generalizando. Si en la taxonomía contenida en la ontología que se utiliza en el proceso de generalización no cubre totalmente el dominio específico del atributo (no contiene todos los valores del atributo que están en el conjunto de datos) que se desea generalizar, quedarán transacciones cuyos valores no podrán modificarse y por tanto darían al traste con la reducción del tamaño de la instancia, por tanto la utilización de una ontología que represente de manera correcta y completa (en la medida de lo posible) el dominio de interés es definitorio para alcanzar buenos resultados en la aplicación del método.

El nivel de la taxonomía al que se desea realizar la generalización tiene un papel fundamental en la reducción del tamaño de la instancia de entrada. A medida que se aumenta el nivel de los conceptos de la taxonomía disminuye considerablemente el tamaño de la instancia. En los casos de prueba descritos anteriormente se comprobó que para el conjunto de datos 1 la reducción de la instancia al nivel k=1 de generalización fue de un 44%, 38% para el nivel 2 y de un 12% para el nivel 3. Este comportamiento es similar para los conjuntos de datos 2 y 3 (ver Tabla 7).

Conjunto de datos	Nivel de generalización k	Reducción	Reducción %
1	1	86464×10^3	44
	2	7426×10^4	38
	3	24775×10^3	12
2	1	375648×10^3	33
	2	320352×10^3	28
	3	104487×10^3	9
3	1	212128×10^4	21
	2	17972×10^5	17
	3	57333×10^4	5

Tabla 7: Reducción porcentual de la instancia de entrada.

En los resultados descritos anteriormente se puede apreciar que el tamaño de la instancia de entrada disminuye a medida que aumenta el nivel de abstracción al que se generalizan los elementos. En el proceso de generalización de ítems se debe tener en cuenta que aunque la mayor reducción del tamaño de la instancia se obtiene abstrayendo al mayor nivel posible, esto no siempre es conveniente debido a que se puede perder generalidad e información en las reglas generadas. Es aconsejable en la mayoría de los casos generalizar a niveles medios de la taxonomía obteniendo reducciones considerables y sin perder generalidad ni información relevante en las reglas de asociación generadas posteriormente. El nivel de generalización es un parámetro al que se le debe prestar especial atención con vistas a la automatización del proceso, una idea interesante en línea con lo propuesto en este trabajo, sería definir ontologías de preferencias para los usuarios que puedan especificar el nivel de generalización útil para ellos en cada caso.

El tiempo de ejecución está directamente relacionado con el tamaño de la instancia de entrada. Otro elemento importante en el tiempo de ejecución del procedimiento de generalización es la cantidad de ítems eliminados durante el proceso, dado que en caso de eliminación deben realizarse operaciones adicionales que aumentan el uso de los recursos. El tiempo de ejecución aumenta de forma exponencial con respecto a la cantidad de tuplas por lo que sería aconsejable particionar las transacciones en partes de 1000 tuplas (este valor fue efectivo en el hardware en el que se ejecutaron las pruebas del sistema) ya

que a esta cantidad alcanza su mayor rendimiento el procedimiento de generalización propuesto (ver Tabla 8).

Conjunto de datos	Nivel de generalización k	Reducción	Tiempo de ejecución (seg.)
1	1	86464×10^3	6
	2	7426×10^4	4
	3	24775×10^3	4
2	1	375648×10^3	16
	2	320352×10^3	13
	3	104487×10^3	9
3	1	212128×10^4	48
	2	17972×10^5	38
	3	57333×10^4	25

Tabla 8: Tiempos de ejecución para los casos de prueba.

La cantidad de transacciones de los conjuntos de datos es otro elemento a tener en cuenta en la generalización de ítems para disminuir el tamaño de la instancia de entrada. A medida que aumentan las transacciones generalmente aparecen nuevos valores para cada uno de los atributos del conjunto de datos. Si el cubrimiento de la ontología que se está utilizando es total sobre el atributo al que se le está aplicando la generalización este aumento no tiene consecuencias. Por otra parte si en las nuevas transacciones aparecen valores que no están recogidos en la taxonomía, la reducción del tamaño de la instancia de entrada se ve afectada con respecto a la cantidad de transacciones.

3.4 Conclusiones parciales

Se determinó que el método “Demostración” es el adecuado para realizar las pruebas de validación para la solución propuesta en el presente trabajo; definiéndose los elementos fundamentales de la demostración propuesta así como los recursos con que se cuenta para efectuar los casos de prueba. Para demostrar el correcto funcionamiento del procedimiento de generalización sobre datos reales se extrajeron conjuntos de datos de las bases de datos de la Aduana General de la República de Cuba.

A partir de los casos de prueba aplicados se pudo comprobar que el procedimiento desarrollado disminuye el tamaño de la instancia de entrada. Se ofrecieron elementos concretos que demuestran la efectividad de la generalización como método de disminución de la cantidad de ítems y por tanto de la instancia de entrada a los algoritmos de minado de reglas de asociación y la disminución de la complejidad computacional de estos últimos. Se ofrecieron elementos comparativos tanto para expresar la validez de la solución como para seleccionar el conjunto de parámetros que permitan obtener un rendimiento superior de la solución propuesta.

CONCLUSIONES

En este trabajo se implementó un algoritmo para reducir el tamaño de la instancia en el procesamiento de reglas de asociación, usando conocimiento previo del negocio reflejado en una ontología. Los resultados alcanzados en la demostración de la solución permiten concluir lo siguiente:

- El algoritmo desarrollado permite reducir el tamaño de la instancia en algoritmos de extracción de reglas de asociación.
- Existe una relación directa entre el tamaño de la entrada y la tasa de reducción que alcanza el algoritmo.
- A medida que se generaliza a niveles más cercanos a la raíz la tasa de reducción crece, pero es posible que se pierdan reglas relevantes.

RECOMENDACIONES

Se recomienda integrar el presente trabajo con la solución “Eliminación de itemsets no significativos para los usuarios” desarrollada por el estudiante Bernardo Corrales Armbruster y con el algoritmo Apriori-Like para el minado de reglas de asociación desarrollado por el estudiante Andy Fernández Garabote. La integración de estas soluciones daría como resultado un marco integrado para la minería de reglas de asociación. En consecuencia de ganar en optimización se recomienda la implementación del algoritmo de generalización en un lenguaje de bajo nivel cercano a *C*, esto permitiría incluir la solución como una librería de PHP lo que facilitaría su integración con el sistema GINA (Gestión Integral de Aduanas). También sería conveniente poder incorporar otros mecanismos de reducción de la instancia de entrada a los algoritmos de minado de reglas de asociación en la etapa de pre-procesamiento.

En este estudio se comprobó la existencia de una relación estrecha entre el nivel de generalización y la reducción alcanzada, por cuestiones de tiempo no fue posible desarrollar métodos automáticos que permitieran escoger el nivel de generalización a aplicar en cada caso. Por lo que se recomienda con vistas a mejorar los resultados alcanzados, integrar a la solución mecanismos que posibiliten determinar cuál es el nivel de generalización recomendado. A juicio del autor, en este sentido sería relevante el uso de ontologías de preferencias de los usuarios.

BIBLIOGRAFÍA

Agrawal R, Srikant R. 1994. *Fast algorithms for mining association rules.* s.l. : Proceedings of the 20th, 1994.

Agrawal, R. and Imielinski, T.and Swami, A. 1993. *Mining Associations between sets of items in massive databases.* ACM-SIGMOD International Conference on Data : s.n., 1993.

Agrawal, R. and Srikant, R. 1994. *Fast Algorithms for Mining Association Rules.* Proceedings of the 20th VLDB Conference, Santiago, Chile : s.n., 1994.

Agrawal, R., Imielinski, T. and Swami, A. 1993. *Minning associations between sets of items in massive databases.* 1993.

Ahn, KI and Kim., Jae-Yearn. 2004. *Efficient Mining of Frequent Itemsets and a Measure of Interest for Association Rule Mining.* Journal of Information & Knowledge Management : s.n., 2004.

Berndtsson, Mikael , et al. 2008. *Thesis Projects: A Guide for Students in Computer Science.* 2008.

Berzal, F., et al. 2001. *A new framework to assess association rules.* Proceedings of the 4th International Conference on Intelligent Data Analysis : s.n., 2001.

Brin, S., et al. 1997. *Dynamic itemset counting and implication rules for market basket data.* 1997.

Cabena, P., et al. 1998. *Discovering Data Mining: From Concept to Implementation.* Upper Saddle River, NJ : s.n., 1998.

Cañete. 2002. *¿Qué se entiende, en España, por Investigación en Ingeniería del Software?* s.l. : MIFISIS, 2002.

Cao, Longbing , et al. 2010. *Domain Driven Data Mining.* 2010.

Céspedes, Zulia Ramírez. *Las ontologías como herramienta en la Gestión del Conocimiento.* Ciudad de La Habana,Cuba : s.n.

Chen, X., Zhou, X. and Scherl, R.: Geller, J. 2003. Using an interest ontology for improved support in rule mining. *Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery.* 2003.

Concepción, Raimil Cruz, Cruz, Liz Mary and Reyes, Liannet Lucia. *Utilización de la Representación Formal de Conocimiento en la Toma de decisiones.*

Dong, G. and Li, J. 1998. *Interestingness of discovered association rules in terms of neighbourhood-based unexpectedness.* 1998.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. *Advances in Knowledge Discovery and Data Mining.* s.l. : AAAI/MTI Press, 1996.

Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. 1992. *Knowledge discovery in databases - an overview.* 1992.

Garea, Llano Eduardo and Vera, Voronisky Francisco. 2009. Alineamiento de Ontologías en el dominio geoespacial. La Habana : CENETAV, SerieAzul RT_010, 2009.

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques.* 2001.

Houtsma, H. and Swami, A. 1993. *Set Oriented Mining of Association Rules.* 1993.

Jerome, David, Guillet, Fabrice and Briand, Henri. 2007. *Association Rule Ontology Matching Approach.* Polytechnic School of Nantes University, France : s.n., 2007.
Kryszkiewicz, M. 1998a. *Representative Association Rules.* s.l. : Lectures Notes in Artificial Intelligence 1394. Research and Development in Knowledge Discovery and Data Mining. Springer-Verlag., 1998a.
—. 1998b. *Representative Association Rules and Minimum Condition Maximum Consequence Association Rules.* s.l. : In proceedings of Principles of Data Mining and knowledge Discovery, 1998b.
Laleh, Naeimeh and Abdollahi Azgomi, Mohammad. 2009. *A Taxonomy of Frauds and Fraud Detection Techniques.* Tehran, Iran : Iran University of Science and Technology, 2009.
Larose, D. T. 2004. *Discovering Knowledge in Data: An Introduction to Data Mini.* Wiley : s.n., 2004.
Li, Jiye and Cercone, Nick. 2006. Introducing a Rule Importance Measure. *Lecture Notes in Computer Science.* s.l. : Springer Berlin / Heidelberg, 2006.
Li, Jiye, et al. 2009. *Enhancing Rule Importance Measure Using Concept Hierarchy.* 2009.
Mannila, H., Hand, D. J. and Smyth, P. 2001. *Principles of Data Mining.* Cambridge : The MIT Press, 2001.
Marinica, Claudia, Guillet, Fabrice and Briand, Henri. *Post-Processing of Discovered Association Rules Using Ontologies.* s.l. : LINA – Ecole polytechnique de l'Université de Nantes, France.
McGuinness, L., F, Natalya and Noy, Deborah. 2005. *Desarrollo de Ontologías -101:Guía Para Crear Tu Primera Ontología.* Stanford : s.n., 2005.
Medina, Pagola J. E., et al. 2007. *Generación de conjuntos de ítems y reglas de asociación.* La Habana, Cuba : Serie Gris, CENETAV, 2007.
Michail, Amir. 2000. *Data Mining Library Reuse Patterns.* Seattle, WA : Dept. of Computer Science and Engineering, 2000.
Morales, Yanssel Urquijo. 2009. *Análisis y diseño de un agente semántico basado en ontologías para el dominio de la salud.* La Habana : s.n., 2009.
Nigro, Hector Oscar, González Cisaro, Sandra Elizabeth and Xodo, Daniel Hugo. 2008. *Data Mining with Ontologies: Implementations, Findings, and Frameworks.* 2008.
Nuñez, Esmeralda Ramos y Haydemar. 2007. *Ontologías:Componentes, metodologías, lenguajes,herramientas y aplicaciones.* Caracas : s.n., 2007.
OMG, Organización Mundial de Aduanas. 1997. *Convenio de Kyoto - Directivas de control aduanero.* Kyoto : http://www.wcoomd.org/Kyoto/20Sp/cap6, 1997.
Ruiz, María Dolores Jiménez. 2010. *MODELADO FORMAL PARA REPRESENTACION Y EVALUACION DE REGLAS DE ASOCIACION.* Granada : s.n., 2010.
Savasere, A., Omiecinski, E. and Navathe, S. 1995. *An Efficient Algorithm for Mining Association Rules in Large Databases.* Austin, TX, U.S.A. : Data Mining Group, Tandem Computers, Inc., 1995.
Shaw, M. 2002. *What makes good research in software engineering.* s.l. : FOR TECHNOLOGY TRANSFER (STTT). SPRINGER BERLIN / HEIDELBERG, 2002.

—. 2003. *Writing good software engineering research papers: minitutorial.* Washington : Proceedings of the 25th International Conference on Software Engineering, ICSE, 2003.
Srikant, Ramakrishnan and Agrawal, Rakesh. 1995. *Mining Generalized Assosiation Rules.* San Jose, California : IBM Almeden Research Center, 1995.
—. *Mining Generalized Assosiation Rules.* San Jose, California : IBM Almeden Research Center.
Toivonen, H. 1996. *Sampling Large Databases for Association Rules.* Proceedings of the 22nd VLDB Conference, Bombay, India : s.n., 1996.
Triantaphyllou, Evangelos. 2010. *DATA MINING AND KNOWLEDGE DISCOVERY VIA LOGIC-BASEDMETHODS.* Baton Rouge, Louisiana, USA : Springer, 2010.
Vaishnavi , Vijay K. and Kuechler Jr., William. 2008. *Design Science Research Methods and Patterns Innovating Information and Communication Technology.* NewYork : Auerbach Publications is an imprint of the , an informa business , 2008.
Vaishnavi, Vijai K and Kuechler, William Jr. 2008. *Design Science Research Methods and Patterns.* New York : Auerbach Publications, 2008.
Xiong, Xia Shi, Fan, Li and Lei, Zhang. 2010. *Ontology-based Association Rule Quality Evaluation Using Information Theory.* 2010.
Yaqin, W. and Yuming, S. 2010. *Classification Model Based on Association Rules in Customs Risk Management Application.* International Conference on Intelligent System Design and Engineering Application : s.n., 2010.
Zhou, X. and Geller, J. 2007. *Raising, to Enhance Rule Mining in Web Marketing with the Use of an Ontology.* New Jersey Institute of Technology : s.n., 2007.
Zhou, Zonglin, et al. *Rule Mining with Prior Knowledge.*

ANEXOS

"[Insertar anexos]"

GLOSARIO

"[Insertar glosario]"