UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

Departamento de Ingeniería y Gestión de Software

Almacén de datos para el análisis de reglas de asociación

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autor:

José Carlos García Hernández

Tutor:

MSc. Julio C. Díaz Vera

Ing. Guillermo M. Negrín Ortiz

La Habana, mayo de 2019

"Año 61 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaro ser el autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

_____ _____

Nombre del autor
José Carlos García Hernández

Nombre tutor
MSc. Julio César Díaz Vera

_____

Nombre Tutor
Ing. Guillermo Manuel Negrín Ortiz

DATOS DE CONTACTO

Síntesis del Tutor

A mis padres por su paciencia.

AGRADECIMIENTOS

A mis padres por todo su amor y apoyo incondicional durante toda mi vida.

A mis tutores Julio y Guille por dedicarme más tiempo del que podían y debían.

A mi oponente por todos sus comentarios y ayuda finalizando este proceso.

A mi profesora Dariela por su apoyo en estos años.

A todos los profesores que de una forma u otra contribuyeron con mi formación profesional.

RESUMEN

La extracción de reglas de asociación es uno de los problemas más estudiados y utilizados en minería de datos. El 80% del trabajo involucrado con la extracción de reglas está asociado a los procedimientos necesarios para la limpieza de los datasets. Tradicionalmente este tipo de trabajo se realiza sobre ficheros planos debido a que la mayoría de los algoritmos de extracción tienen este tipo de entrada. En este trabajo se propone un mecanismo para la extracción de reglas de asociación que tiene como entrada datos multidimensionales. La utilización de este tipo de entrada reduce la carga necesaria para el preprocesamiento del fichero y facilita que los usuarios puedan transformar los datos usando operaciones OLAP (Procesamiento Analítico en Línea, por sus siglas en inglés). Los resultados alcanzados en la experimentación muestran que el mecanismo de extracción mejora los tiempos de extracción de las propuestas en el estado del arte por varias unidades de magnitud.

Palabras claves: reglas de asociación, minería OLAP, modelos multidimensionales

ABSTRACT

Association rules extraction is one of the most treated and studied topics in Data Mining. 80% of the papers involving association rules extraction is related to the cleaning task. Traditionally this type of work is carried out over plain text due to most of the extraction algorithms use this kind of input. This research proposes a mechanism for rule extraction that uses as input multidimensional data. The use of this kind of input reduces the overall load involves in pre-processing the file and facilitates the users be able to transform data using OLAP operations. The results obtained in the experiments show that the extraction mechanism improves extraction time improves the variants over those proposed in the state of the art in several magnitude units.

Keywords: *association rules, OLAP mining, mutidimensional models*

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

En la actualidad el uso de bases de datos por parte de las empresas se ha vuelto fundamental ya que la utilización de estas les permite crear nuevas estrategias de marketing y mantener un mayor control tanto de ventas como de inversiones. Sin embargo, la masividad de datos supone un problema, demasiada información hace complicada la organización y la extracción de conocimientos, ante esta dificultad entra en juego la minería de datos.

La minería de datos constituye el centro del proceso de KDD (Knowledge Discovery in Data bases) (Dhanabal & Shantharajah, 2015), ya que se centra en la búsqueda de patrones dentro de grandes colecciones de datos para encontrar un hecho o hechos en común de estas. Esta técnica de extracción de conocimiento consta de siete fases entre las cuales se encuentran: selección, extracción, limpieza, transformación, minería de datos, evaluación y difusión.

La fase de selección se centra en la integración y recopilación de datos, determinando la utilidad de las fuentes. Dado que los datos provienen de diferentes fuentes es necesario la exploración mediante técnicas de análisis exploratorio buscando entre otras, distribución de los datos, su simetría, normalidad y la correlación existente entre ellos. Seguidamente es necesario la fase de limpieza de los datos ya que pueden contener valores atípicos, valores faltantes y valores erróneos. En esta fase se analiza la influencia de los datos atípicos, se imputan los valores faltantes y se eliminan o corrigen los datos incorrectos.

De ser necesario, se lleva a cabo la transformación de los datos, generalmente mediante técnicas de reducción o aumento de la dimensión y escalado multidimensional. La fase de minería es el momento de decidir cuál es la tarea a realizar (clasificar, agrupar, etc.) y se escoge la técnica descriptiva o predicativa a utilizar para llevar a cabo las tareas de minería. Las fases que más tiempo consumen son las fases de extracción y limpieza pues la mayoría de los algoritmos de minería de datos necesitan transformar un conjunto de transacciones a texto plano.

Hoy en día la información es un bien fundamental para todas las empresas. La información que es capturada en los ambientes operacionales es clave para la toma de decisiones. Con el objetivo de obtener, almacenar y procesar la información se utilizan diferentes sistemas de información. Los sistemas de información empresariales usualmente manejan fuentes de datos heterogéneas y complejas como pueden ser datos, operaciones o base de datos relacionales, para realizar las tareas de análisis comúnmente se utilizan almacenes de datos.

Los almacenes de datos reciben los datos a partir de los sistemas operacionales, debido a la cantidad y complejidad de estos sistemas los datos deben ser homogenizados y limpios antes de ser almacenados. Cuando los datos han sido cargados en el almacén la empresa tiene una visión centralizada de su negocio desde diferentes perspectivas. Estas perspectivas pueden ser adaptadas de acuerdo con la necesidad de la

empresa, la adaptación se hace mediante el diseño de diferentes caminos de análisis comúnmente denominados dimensiones. Las dimensiones permiten analizar los hechos con diferentes granularidades y representan las características topológicas del análisis (Bawane & Deshkar, 2015).

Los almacenes de datos permiten reducir en gran medida el tiempo necesario para la aplicación de técnicas de minería de datos, debido a que disponen de juegos de datos limpios y estandarizados (Marco-Ruiz, Moner, Maldonado, Kolstrup, & Bellika, 2015). Una de las técnicas de minería de datos más utilizadas es la de extracción de reglas de asociación. Sería provechoso realizar extracción de reglas de asociación sobre cubos de datos y de esta forma reducir en gran medida las tareas de selección, extracción y limpieza. Sin embargo, la mayoría de los algoritmos de extracción solo trabajan sobre ficheros planos.

Ante la problemática anteriormente planteada y para guiar el desarrollo de la investigación se plantea como **problema a resolver:** ¿Cómo implementar un modelo de extracción de reglas de asociación que aproveche las bondades de los cubos de datos?

Teniendo como **objeto de estudio:** minería de datos en cubos OLAP.

Para dar solución al problema a resolver se define como **objetivo general:** desarrollar un modelo para la extracción de reglas de asociación en cubos OLAP.

Delimitando como **campo de acción:** algoritmos de extracción de reglas de asociación sobre cubos OLAP.

Para cumplimiento al objetivo general se definen los **objetivos específicos:**

1. Establecer el marco conceptual para el desarrollo de la investigación.
2. Implementar el modelo para la extracción de reglas de asociación en cubos OLAP.
3. Validar la propuesta.

Dando solución al problema y cumplimiento al objetivo general como **posible resultado** de la investigación se tiene: modelo para la extracción de reglas de asociación en cubos OLAP.

La tesis se estructura en los siguientes capítulos:

Capítulo 1.

Se presenta la descripción de cada una de las etapas del método científico revisión sistemática de la bibliografía, las cuales son: planificación, ejecución y resultado. Además, se definen las etapas para el minado de reglas de asociación en almacenes de datos.

Capítulo 2.

Se presentarán los procedimientos utilizados para la obtención de reglas de asociación sobre cubos de datos. Para cada uno de los componentes del procedimiento se detallan los supuestos teóricos, las herramientas y

técnicas necesarias para su implementación. A lo largo del capítulo se utiliza un caso de estudio como prueba de concepto que garantiza la aplicabilidad del procedimiento.

Capítulo 3.

Se presenta la validación de la propuesta de solución, la cual se realiza haciendo uso de un cuasi experimento en el que se medirá el tiempo de respuesta necesario para obtener un modelo de reglas de asociación en dos escenarios: utilizando la variante de extracción de reglas de asociación propuesta en la presente tesis y utilizando la variante de extracción de reglas de asociación que utilizan los algoritmos Apriori y Fp-growth sobre ficheros de texto plano. Seguidamente se aplican trasformaciones a los datos para simular la interacción del usuario y se ejecutan las tres variantes de extracción de reglas de asociación evaluando el tiempo de respuesta.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

Introducción

En este capítulo se presenta la descripción de cada una de las etapas de método científico revisión sistemática de la bibliografía las cuales son: planificación, ejecución y resultado. Además, se definen las etapas para el minado de reglas de asociación en almacenes de datos.

1.1. Revisión sistemática de la bibliografía

En la realización del capítulo se utilizó el método de revisión sistemática de la bibliografía (Bonidia, Rodrigues, Avila-Santos, Sanches, & Brancher, 2018), este método permite hacer una rigurosa y exhaustiva evaluación de la bibliografía de un tema determinado, en este caso sobre la extracción de reglas de asociación en cubos de datos. Durante la ejecución de este método se recopilan evidencias extraída a través de la búsqueda sistemática de artículos relacionados con el tema y de la síntesis de dichos artículos. La realización de este método consta de 3 fases: la planificación, la conducción y el análisis. A continuación, se evidencia la aplicación de cada una de estas fases ya desarrolladas hacia el tema de esta tesis y las funciones que cumplen en la investigación.

1.2.1 Planificación y ejecución de la investigación

La planificación es la fase en la cual se definen las preguntas científicas a las cuales se les va a dar respuesta durante el desarrollo de la investigación, se logra también en esta fase se realiza la integración de las especificaciones de interés y la identificación de las bases de datos utilizadas. Se definen las palabras claves, las estrategias de búsqueda y los criterios de inclusión y exclusión por los cuales se va a regir la selección de los artículos.

A continuación, se listan las preguntas de investigación (PI):

PI1- ¿Existen actualmente estudios que propongan un modelo genérico para la extracción de reglas de asociación en almacenes de datos?
PI3- ¿Cuáles son los algoritmos más utilizados para extraer reglas de asociación?
PI4- ¿Cómo se pude calcular el soporte y confianza de los itemset en un almacén de datos?
PI5- ¿Cuáles son los dataset utilizados para realizar los estudios seleccionados?
PI7- ¿Cuáles son los resultados alcanzados por los estudios a revisar?

Definición de los criterios de inclusión y exclusión de los artículos:

Inclusión:

- Estudios en inglés.
- Estudios en español.
- El tipo de estudio puede ser artículo, conferencia.

- Todos los estudios tienen que ser relevantes para el tema, o sea todos los estudios tiene que estar en el área de la minería de datos y los modelos multidimensionales.

Exclusión:

- Estudios que no cumplen con la calidad requerida para esta investigación.
- Estudios fuera del contexto del trabajo.
- Estudios que estén escritos en cualquier otro idioma que no sean los definidos en los criterios de inclusión.
- Estudios publicados antes del año 2014.

Calidad:

- Todos los estudios seleccionados tienen q arribar a una o varias soluciones.

Además de los criterios antes planteados en aras de encontrar estudios con la mayor calidad posible se analizará también de cada uno de los estudios encontrados el título, el resumen, las palabras claves, los algoritmos utilizados y los resultados finales.

Selección de las bases de datos y método de búsqueda:

La selección de las bases de datos con las que se va a trabajar es de vital importancia durante la fase de la planificación de la búsqueda, ya que el estudio seleccionado debe cumplir con la calidad requerida, tiene que provenir de bases de datos relacionadas con la ingeniería informática o ser bases de datos de revistas científicas cien por ciento fiables para la investigación, por lo que teniendo en cuenta los elementos mencionados anteriormente se seleccionaron las siguientes bases de datos:

- ACM Digital Library: https://wwww.dl.acm.org
- IEEE Xplorer Digital Library: https://www.ieeexplorer.org
- Springer: https://www.springer.com/gp

La selección de la expresión de búsqueda es el último paso durante la etapa de planificación. El método seleccionado para esta investigación fue el de recuperación booleana, este método consiste en dividir la búsqueda en espacios identificando un subconjunto de documentos en una colección de acuerdo con los criterios de la consulta (Bonidia,2018). En el caso de la investigación se selecciona la siguiente relación de palabas claves: (Datawarehouse OR Datamart OR OLAP OR multidimensional OR Datacube) AND (Datamining OR Association Rules).

Plan de ejecución:

Esta etapa incluye cinco pasos:

- Realizar la búsqueda en las bases de datos seleccionadas.
- Comparar los resultados de búsquedas para excluir trabajos repetidos.
- Aplicar los criterios de inclusión, exclusión y calidad.
- Evaluación de todos los estudios que pasaron la revisión inicial.
- Síntesis de datos.

La siguiente figura muestra un diagrama de flujo de cómo funcionan los criterios de inclusión y exclusión sobre el total de estudios obtenidos de las bases de datos anteriormente presentadas. Para la realización de este diagrama fue necesario el uso de la herramienta StArt (State of the Art through Systemantic Reviews) que como su nombre lo indica es una herramienta para apoyar la revisión sistemática de la bibliografía obtenida. En una primera búsqueda en las bases de datos el total de artículos obtenidos fue 1115. Se eliminaron los artículos anteriores al año 2014 quedando un total de 220 estudios, luego basado en los criterios de exclusión la cantidad de artículos se vio reducida a 85 y por último basados en el criterio de calidad anteriormente definidos la cifra se vio reducida a 15 artículos con los que se procede a trabajar durante la investigación.

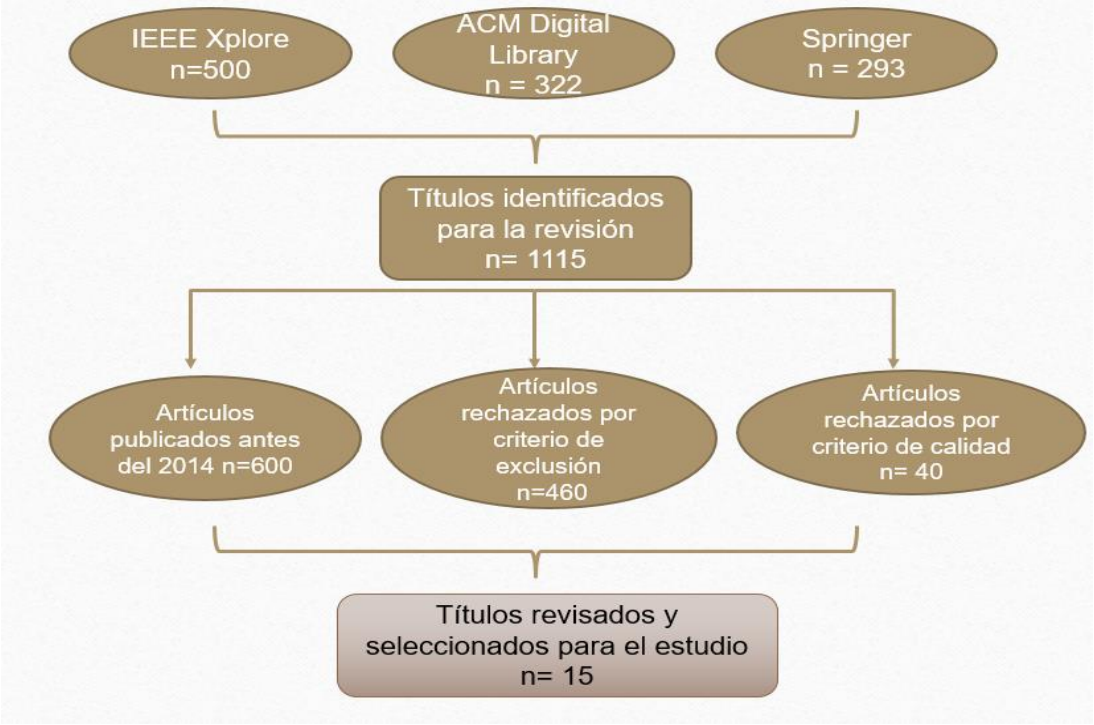

Figura 1 Selección de los artículos.

1.2.2 Resultados

En esta sección se muestran los resultados de la revisión sistemática a través de la respuesta a las preguntas definidas en la fase de planificación.

PC1- ¿Existen actualmente estudios que propongan un modelo genérico para la extracción de reglas de asociación en almacenes de datos?

Para dar respuesta a la pregunta anterior el autor desarrolla la siguiente tabla.

Tabla 1 Identificador de los artículos y sus títulos correspondientes.

ID de los artículos	Título del articulo
3260(Bawane & Deshkar, 2015)	Integration of OLAP and Association rule mining
3262(Zou, Liu, Qin, & Ma, 2014)	Research and Application of Association Rule Mining Algorithm Based on Multidimensional Sets
3263(D. Liu, Wu, Gu, Ma, & Wang, 2017)	A Multidimensional Time-series Association Rules Algorithm based on Spark
3264(Fisun, Kulakovska, & Horban, 2015)	Generation of Frequent Item Sets in Multidimensional Data by Means of Templates for Mining Inter-Dimensional Association Rules
3265(M. Usman, Usman, & Ahmad, 2014)	A Conceptual Model for Multi-level Mining and Visualization of Association Rules
3266(Arincy & Sitanggang, 2015)	Association Rules Mining on Forest Fires Data using FP-Growth and ECLAT Algorithm
3267(Fisur, Horban, & Dvoretskyi, 2018)	Methods of searching for association dependencies in multidimensional databases

3269(Bhavsar & Arolkar, 2014)	Multidimensional Association Rule Based Data Mining Technique for Cattle Health Monitoring Using Wireless Sensor Network
3270(Yokobayashi & Miura, 2018)	Multidimensional Data Mining Based on Tensor Model
3272(F. Liu, Zhou, Wang, Wang, & Zhang, 2018)	Identification of Hypertension by Mining Class Association Rules from Multi-dimensional Features
3273(Wang, Zeng, Xiong, & Yang, 2017)	Finding Main Causes of Elevator Accidents via Multi-Dimensional Association Rule in Edge Computing Environment
3275(Noh, Son, Park, & Chang, 2017)	In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining
3280(Abuelyaman & Elgimari, 2014)	A Prototype for a Data Mining Based Pathfinder to Sudanese Universities
3282(Poli, 2015)	Fuzzy Data Mining and Web Intelligence

La tabla anterior demuestra la existencia de estudios que abordan el tema de extracción de conceptos en el modelo multidimensional. Para dar una visión general de los artículos se propone separarlos en las siguientes clases:

C1-estudio del mercado: los estudios dentro de esta clase se tratan del uso de la minería de datos en apoyo a la toma de decisiones empresarial.

C2- **ciencia aplicada**: los artículos agrupados en esta clase están asociados a la relación de coautoría entre los estudios seleccionados.

C3-análisis energéticos: los estudios bajo este dominio se basan en la recopilación de información de las diferentes fuentes de energía.

C4-incendios forestales: los trabajos orientados a este tema se basan en el estudio de la información recopilada de incendios forestales.

C5-control del personal: eta clase de trabajos se basan en la utilización de almacenes de datos y las técnicas de minería en buscar de un mayor control del personal dentro de la empresa.

C6-enfoque médico: los artículos agrupados en esta clase enfocan sus estudios en el campo de la medicina.

C7-Estudios de accidentalidad: los estudios enfocados a este tema investigan principalmente los accidentes de distintos tipos.

En la siguiente tabla se muestra la organización de los estudios según su clase.

Tabla 2 Agrupación de los artículos según sus clases.

Clases	ID artículo
Estudio del Mercado	3260,3267
Ciencia aplicada	3262,3280,3291,3264, 3265,3282, 3270
Análisis energético	3263
Incendios forestales	3266
Estudios de accidentalidad	3273
Enfoque médico	3269,3272
Control del personal	3275

PC3- ¿Cuáles son los algoritmos más utilizados para extraer reglas de asociación en cubos de datos?

La siguiente tabla muestra los algoritmos que fueron utilizados durante la investigación de cada artículo seleccionado:

Tabla 3 Algoritmos utilizados por cada artículo .

ID del artículo	Algoritmos
3260	Apriori(Yuan, 2017)
3262	MSMINE(Fournier-Viger et al., 2017)
3263	MTAR(D. Liu et al., 2017)
3264	Rule templates(Marinica & Guillet, 2010)
3265	Multi-D-Slicing, n-D cube Search(Muhammad Usman, 2017)

3266	FP-Growth(Zeng, Yin, Liu, & Zhang, 2015), ECLAT(Heaton, 2016)
3267	Apriori
3269	Apriori-TID(Sarma & Mishra, 2016), DHP(Soni, Sharma, & Jain, 2016)
3270	Apriori
3272	Apriori
3273	Apriori
3275	Apriori
3280	Apriori
3282	-
3291	-

Durante el análisis de los artículos mencionados se pudo observar que el algoritmo más utilizado fue Apriori.

PC4- ¿Cómo se pude calcular el soporte y confianza de los itemset en un almacén de datos?

Los estudios seleccionados se refieren al uso de las medidas de soporte y confianza de distintas formas:

Los umbrales utilizados para el análisis son los de soporte y la confianza para generar las reglas. Los valores obtenidos después del cálculo se comparan con el umbral, o sea el valor definido por el usuario. Si el valor de conteo obtenido es mayor o igual que el valor del umbral, se toma el conjunto de elementos en consideración. soporte y confianza se calculan de la siguiente forma:

Soporte(X): cantidad de veces que aparece el dato X en todas las transacciones dividido por el total de transacciones.
Confianza(X⇒Y): soporte(X U Y) / soporte(X).

PC5- ¿Cuáles son los dataset utilizados para realizar los estudios?

La siguiente tabla muestra la proveniencia de los dataset con los que trabaja cada estudio de los seleccionados:

Tabla 4 Dataset de los artículos seleccionados.

ID del artículo	Dataset
3260	Datos reales de canasta de compras en supermercados en la India.
3262	Resumen de evaluaciones de la universidad de Chiefeng
3263	Datos de la red eléctrica de Beijín
3264	EMS
3265	Base de datos de incendios forestales de Bogor
3266	Datos reales de diferentes canastas de compras
3267	Datos reales de la salud del ganado usando sensores inalámbricos del ganado en la India.
3269	Datos reales del satélite meteorológico Himawari
3270	Datos reales tomados de la universidad de Northwestern
3272	Base de datos de accidentes en elevadores en Beijín
3273	Datos reales de estudiantes de la universidad Sundance
3275	Datos reales tomados de www.eia.gov/consumption/commercial/data/ 2003
3280	-
3282	-
3291	Dataset reales tomados de la UCI(Universidad de California Irving)

Luego de analizar la disponibilidad de los dataset utilizados se decidió seleccionar un dataset proveniente de la UCI (http://www.archive.uci.edu) debido a que la mayoría de los estudios utilizan dataset de los cuales no está disponible su uso.

PC7- ¿Cuáles son los resultados alcanzados por los estudios a revisar?

Tabla 5 Resultado de cada investigación estudiada.

ID de los artículos	Resultado
3260	Mejora del tiempo de ejecución del algoritmo Apriori.
3262	Se encuentran modelos de reglas de asociación más compactos.
3263	Mejora el rendimiento de los algoritmos.
3264	Formalización de un mecanismo de representación utilizando plantillas
3265	Formalización de un mecanismo para la extracción de regla.
3266	Encuentra relaciones entre los puntos calientes y las áreas de baja precipitación.
3267	Se encuentran reglas de asociación interdimensionales
3269	Se identifica asociación entre las enfermedades del ganado.
3270	Disminuye la complejidad espacial de los algoritmos de extracción.
3272	Se mejora el rendimiento de los métodos de extracción de reglas de asociación.
3273	Se mejora el rendimiento del algoritmo Apriori.
3275	Se encuentran relaciones entre los edificios y consumo de energía

3280	Se encuentran asociaciones entre el fracaso escolar, el tipo de estudiante y los cursos seleccionados.
3282	Se extiende el modelo de reglas de asociación en cubos de datos para tratar la lógica difusa.
3291	Se propone una metodología para construir modelos de reglas de asociación sobre cubos de datos.

1.2. Etapas para el minado de reglas de asociación en almacenes de datos

A partir de los elementos presentados en las secciones anteriores se puede establecer un grupo de principios que guiarán la obtención de reglas de asociación sobre cubos de datos, teniendo en cuenta la definición de extracción de conocimientos en bases de datos (KDD Knowledge Discovery in Data bases) referenciada en (Pupezescu y Rădescu, 2016) y establecida como el proceso de descubrimiento de patrones novedosos potencialmente útiles y no triviales en bases de datos. En primer lugar, la etapa que más tiempo y esfuerzo consume dentro del KDD es la asociada a la limpieza de los datos y al considerar a un almacén de datos como fuente de datos primarios esta carga se reduce drásticamente.

Un almacén de datos es una colección de datos orientados a tema, integrados, temporales, y no-volátiles para la toma de decisiones (Inmon and Hackathorn 1994). De las propiedades mencionadas anteriormente es de particular interés para la investigación la integración, que plantea que la base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes y limpios. Esta característica del almacén favorece el proceso de extracción de conocimientos ya que no es necesario ejecutar costosas rutinas de limpieza de datos.

Otro aspecto importante para el KDD está asociado al entendimiento del dominio que se va a estudiar en este sentido los almacenes de datos cuentan con un modelo lógico claro y bien definido denominado modelo multidimensional.

Los componentes del modelo multidimensional son definidos por Carlos Molina de la siguiente forma (Martínez-Rojas, Marín, Molina, & Vila, 2015):

- **Dimensión**: es una dupla d = (l, ≤d, l⊥, l⊤) donde l = {l1, ..., en} tal que cada li es un conjunto li = {ci1, ..., sin} tal que, li ∩ lb = ∅ si i = j, ≤d es una relación de orden parcial tal que li ≤d si ∀sij ∈ li ⇒

∃cup ∈ lek/sij ⊆ cup. l⊥ y l⊤ son dos elementos tal que ∀li ∈ l l⊥ ≤d li y li ≤d l⊤. A cada elemento li se denomina nivel, para identificar el nivel li de l.

Las dimensiones establecen el contexto del análisis sobre los hechos. Para acceder a los hechos a diferentes niveles de detalle, se puede definir jerarquías sobre las dimensiones, estableciendo los niveles de granularidad posibles. Cada uno de estos niveles será un conjunto de nombres o etiquetas que definen subconjuntos de elementos de los niveles inferiores que agrupan.

- **Hecho**: Considerando un conjunto de atributos A1, ..., An con dominios D1, ..., Dn, se denomina hecho a cualquier h = (x1, ..., xn) tal que xi ∈ Di ∀i = 1, ..., n, es decir, cualquier n-tupla definida sobre los dominios de los atributos que interesa estudiar.

Las variables del dominio que se quiere analizar definirán los hechos del DataCubo. Estos serán los que en mayor medida acotarán el dominio de análisis del DataCubo que se defina, limitando qué medidas se pueden obtener. Las dimensiones aportan la contextualización de estas variables, localizando cada hecho dentro del espacio que definen. Cada uno de estos hechos será una agrupación de variables (una tupla).

Finalmente, un data cubo es una tupla C = (D, lb, F, A) tal que D = (d1, ..., dn) es un conjunto de dimensiones, lb = (l1b, ..., lnb) un conjunto de niveles tal que li pertenece a di, F = R ∪ ∅ donde R es el conjunto de hechos, ∅ un símbolo especial y A es una aplicación definida como A: l1b × ... × lnb → F, que para cada conjunto de valores de las dimensiones devuelve el hecho relacionado con estas coordenadas.

Desde el punto de vista de la extracción de conocimientos el modelo multidimensional aporta ventajas al cliente que puede realizar diferentes operaciones sobre el mismo dentro de ellas son importantes las asociadas a las tecnologías OLAP (On-line Analytical Processing).

OLAP se caracteriza por ser una categoría de software que permite a los analistas, directivos y ejecutivos acceder a los datos de forma rápida, consistente e interactiva a través de una amplia variedad de vistas de la información que han sido obtenidas de datos sin procesar (OLAPCouncil). Para obtener las vistas se aplican las denominadas funciones OLAP a los datos.

Las operaciones son:

- Roll-up para subir el nivel de granularidad en una jerarquía
- Drill-Down para hacer lo contrario.
- Slice para restringir las dimensiones a analizar
- Dice para restringir valores de la tabla de hechos
- Pivot para hacer una tabulación cruzada

La técnica de minera de datos a utilizar en la investigación es la de extracción de reglas de asociación.

Una regla de asociación se define como:

Sea I = {i1, i2, ..., im} un conjunto de literales llamados elementos. Sea D un con junto de transacciones, donde cada transacción T es un conjunto de elementos tal que T ⊆ I. Una transacción T contiene un conjunto de elementos X si y solo si X ⊆ T.

El algoritmo más utilizado en los artículos analizados para la extracción de reglas de asociación en almacenes de datos con tecnología OLAP es Apriori. El mismo está basado en el principio de “clausura descendente del soporte de los ítemsets”, el cual plantea que el soporte de cualquier subconjunto de un conjunto de ítems es mayor o igual que el soporte de ese conjunto de ítems. Lo que implica que todo subconjunto de un conjunto frecuente de ítems es también frecuente y que todo súper conjunto de un conjunto de ítems infrecuentes es también infrecuente.

Existen un gran número de métricas para evaluar la calidad de las reglas de asociación, las seleccionadas para la investigación fueron, soporte y confianza, ya que son las más utilizadas en los artículos seleccionados. El comportamiento de dichas métricas es de la siguiente forma, una regla X → Y tendrá en D una confianza con valor c si el c % de las transacciones en D que contienen X también contienen Y. La regla tendrá un soporte con valor s en el conjunto de transacciones D si el s % de las transacciones en D contienen a los elementos X ∪ Y. La confianza da una medida de la calidad de la regla (la fuerza de la implicación). Normalmente se desea encontrar asociaciones entre elementos con un soporte razonablemente alto. Las reglas con confianza alta y soporte alto son denominadas reglas fuertes.

A continuación, se muestran los pasos que se siguen para cumplir el objetivo propuesto en esta investigación:

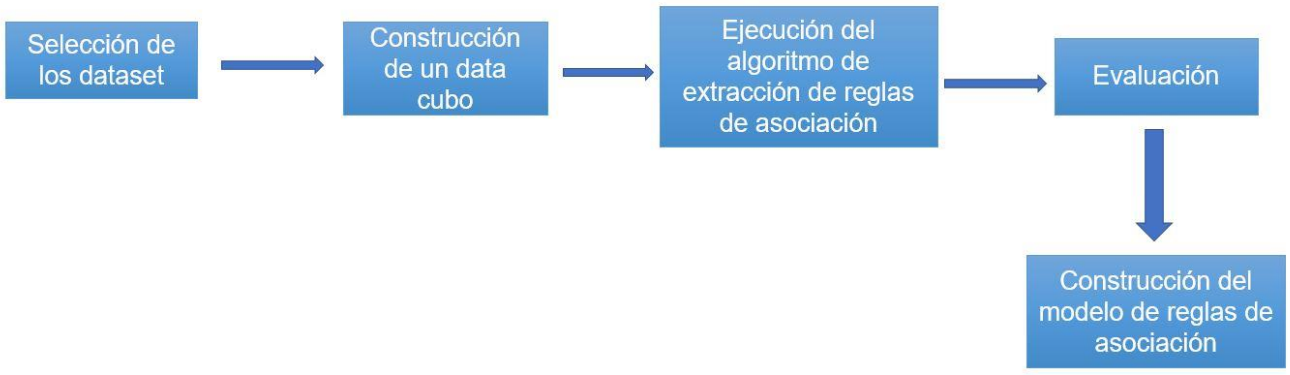

Figura 2 Procedimiento de construcción del modelo de regla de asociación.

Los dataset seleccionados se utilizan para comprobar la efectividad de la propuesta. A partir de los mismos se crea un modelo para almacenar los cubos de datos usando tecnología ROLAP y el gestor de base de datos Postgree SQL. El modelo es poblado a partir de transformaciones realizadas con la ayuda de la herramienta PDI. Una vista minable contiene el conjunto de datos sobre las que se va a ejecutar el algoritmo, se obtiene aplicando alguna de las operaciones OLAP sobre el cubo de datos. El algoritmo de extracción es implementado en PLPGSQL y las métricas utilizadas en la evaluación también son tomadas en cuenta en esta etapa. El modelo resultante es entregado al usuario final.

La evaluación de la propuesta se realiza utilizando la métrica tiempo de ejecución. El objetivo es comparar el tiempo necesario para obtener un modelo de reglas de asociación sobre un fichero plano y luego con el modelo multidimensional correspondiente a dicho fichero.

1.4 Conclusiones parciales

La aplicación del método de revisión sistemática de la bibliografía permitió definir cinco pasos fundamentales (selección del dataset, construcción del datacube, ejecución del algoritmo de extracción, evaluación, construcción del modelo de reglas de asociación) para la extracción de reglas de asociación en modelos multidimensionales. En el capítulo se propuso el procedimiento PRMA que contiene cinco etapas y que permite la extracción de reglas de asociación.

CAPÍTULO 2. DESCRIPCIÓN DE LA PROPUESTA

2.1. Introducción

Durante el desarrollo de este capítulo se va a definir un procedimiento para la obtención de reglas de asociación sobre cubos de datos. Para cada uno de los componentes del procedimiento se detallan los supuestos teóricos y, las herramientas y técnicas necesarias para su implementación. A lo largo del capítulo se utiliza un caso de estudio como prueba de concepto que garantiza la aplicabilidad del procedimiento.

2.2 Procedimiento para el Minado de Reglas de Asociación en almacenes de datos (PMRA)

En el capítulo uno se estableció, a partir del análisis del estado del arte, las fases para la extracción de reglas de asociación del cubo de datos (ver figura 2). A continuación, se describen cada una de las etapas del procedimiento.

2.2.1 Selección del dataset

En esta etapa se selecciona el dataset sin procesar a partir de una fuente de datos. Las fuentes de datos para la construcción de almacenes de datos pueden ser ficheros planos, bases de datos relacionales o datos extraídos directamente de fuentes en línea. De forma general cualquier repositorio que contenga o combine información no estructurada, semiestructurada o estructurada.

No es necesario que las fuentes cumplan ningún formato de entrada determinado, pero en la medida en que menor estructuración tengan los datos mayor tiempo será necesario en la fase de extracción y limpieza. Es recomendable presentar meta información sobre el dataset a utilizar que incluya la cantidad de atributos, cantidad de casos, número de datos nulos que contiene el data set y número de atributos continuos.

En el caso del estudio que será utilizado a modo de prueba de concepto se escogió como fuente de datos el dataset Adult, proveniente del repositorio de la UCI (http://www.archive.uci.edu). En la tabla 6 se presenta la meta información asociada al mismo.

Tabla 6 Relación de datos de la fuente.

Cantidad columnas	Cantidad filas	Cantidad datos nulos	Cantidad atributos continuos
14	48842	2428	5

2.2.2 Construcción de un datacube

En la investigación se hace uso de la definición formal de cubo de datos brindada por (Carlos Molina, 2013).La cual plantea que un cubo de datos es una tupla $C = (D, lb, F, A)$ tal que $D = (d1, \ldots, dn)$ es un conjunto de dimensiones, l$b = (l1b, \ldots, lnb)$ un conjunto de niveles tal que lib pertenece a di, F = R ∪ ∅ donde R es el conjunto de hechos y ∅ un símbolo especial y A es una aplicación definida como A : $l1b \times \ldots \times lnb \rightarrow$ F, donde para cada conjunto de valores de las dimensiones devuelve el hecho relacionado con estas coordenadas.

Si para un $a = (a1, \ldots, an)$ se tiene ∅, indica que para esta combinación de valores no existe un hecho definido. Para comenzar los análisis, se define un datacube al mayor nivel de detalle y sobre él se irá operando. A este datacube de partida se le llamará datacube básico.

Para la implementación del datacube se usa el enfoque de almacenamiento relacional donde a cada dimensión le corresponderá una tabla dentro de la base de datos cuyo nombre debe empezar con el prefijo dim_nombre de la dimensión. Cada dimensión tendrá una llave primaria subrogada con nombre id_dim_nombre de la dimensión y además contará con una llave única de negocio que está compuesta por los elementos significativos de la dimensión.

Los niveles dentro de la dimensión se representarán con un nuevo atributo que opcionalmente puede tener el prefijo Lv.

Los hechos serán recogidos en una tabla que comenzará con el prefijo fact_nombre del hecho cuyo identificador será la combinación de los identificadores de todas las dimensiones de la base de datos. Como un elemento distintivo dentro del modelo se añadirá al hecho la métrica de soporte que facilitará el cálculo de los itemset frecuentes en la tabla de hechos, esta métrica estará representada por el atributo cantidad. La estructura correspondiente con un almacén de datos se muestra en la figura 3.

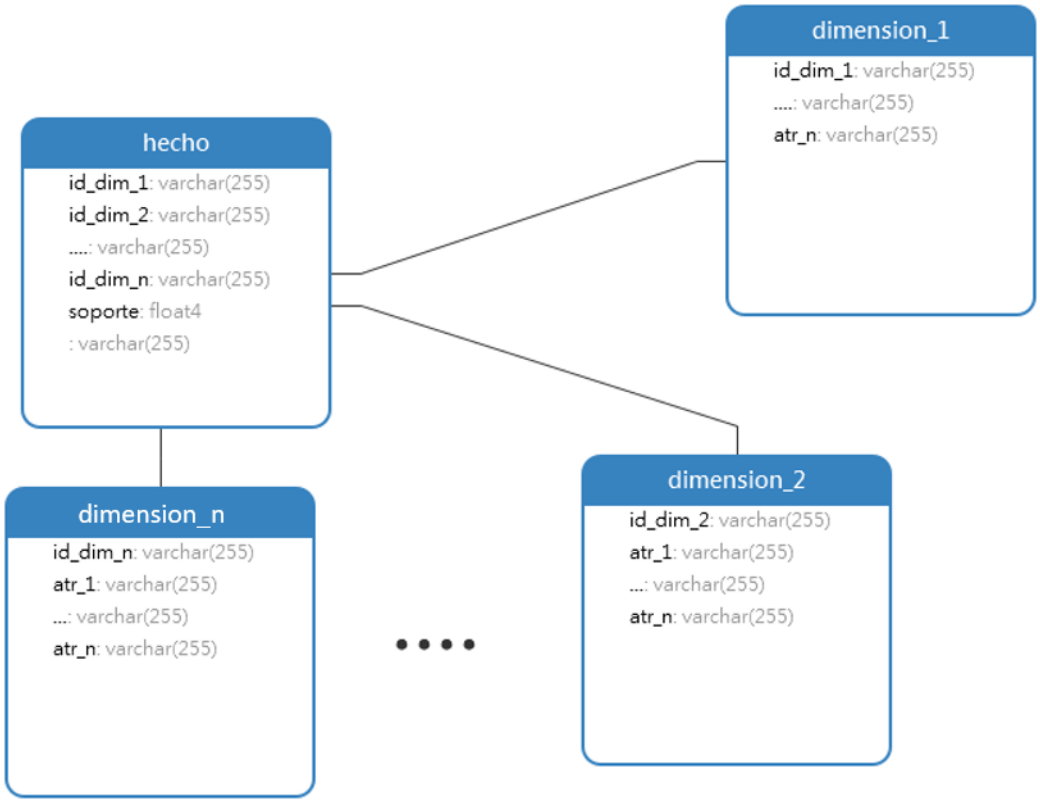

Figura 3 Ejemplo de diagrama de un Datawarehouse.

La figura 4 muestra el diagrama correspondiente con el almacén de datos del censo:

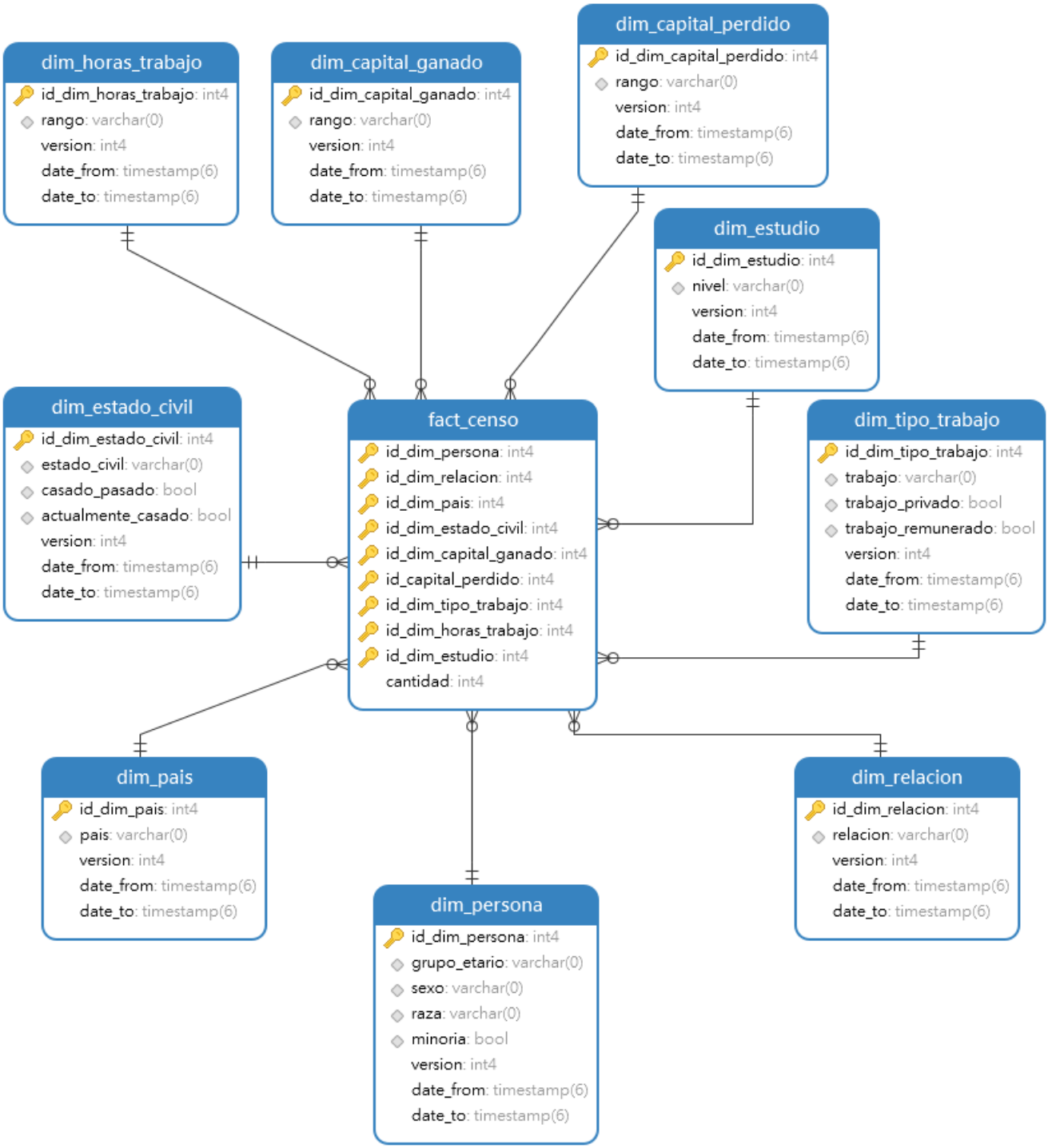

Figura 4 Diagrama del datawarehouse del censo.

Existen tres variantes importantes para almacenar un cubo de datos las cuales son ROLAP (OLAP Relacional), MOLAP (OLAP Multidimensional) o HOLAP (OLAP Híbrido)

ROLAP:

La tecnología más utilizada en almacenes de datos se llama ROLAP, donde los datos analíticos están representados en un formato relacional. Por lo tanto, los diferentes tipos de datos fuente de OLTP se convierten a ROLAP para realizar el procesamiento analítico en datos heterogéneos en un enfoque unificado (Sen, 2017).

La arquitectura ROLAP, accede a los datos almacenados en un datawarehouse para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.

HOLAP:

Un sistema HOLAP soporta e integra el almacenamiento de los datos multidimensionales y relacionales de una forma equivalente con el objetivo de beneficiarse de las correspondientes características y técnicas de optimización. Además, lo importante de esta técnica es que los datos se encuentran almacenados en bases de datos relacionales y las métricas están previamente computadas.

MOLAP:

Los modelos MOLAP pueden almacenar eficientemente dimensiones múltiples utilizando una tecnología de almacenamiento de matrices poco densas. La idea básica es eliminar las posibles celdas vacías (Sen,2017).

MOLAP (Multidimentional On Line Analytic Processing) precalcula los Cubos Multidimensionales y los almacena físicamente.

Este proceso se puede resumir a través de los siguientes pasos:

1. Se seleccionan los Indicadores, Atributos, Jerarquías, etc., que compondrán el Cubo Multidimensional.
2. Se precalculan los datos del Cubo, es decir, todas las combinaciones posibles entre los niveles de las Jerarquías de cada Dimensión.

3. Se ejecutan las consultas sobre los datos precalculados del Cubo.
4. Cada vez que se actualiza el Datawarehouse se debe precalcular y guardar el Cubo, para que contenga los nuevos datos.

Se decide desarrollar un enfoque mixto que no llega a ser HOLAP ya que se van a almacenar los datos usando tecnología relacional y se almacenarán en vistas materializadas el cómputo de todas las funciones de agregación. Esta variante es menos eficiente que almacenar todas las agregación con tecnología multidimensional, pero permite optimizar el desempeño de motores relacionales de bases de datos a la hora de dar respuesta a las consultas:

```
create MATERIALIZED VIEW cube_nombre as select 1.0*(sum(cantidad))/(select sum(cantidad) from fact_censo) as soporte,
id_dim_1,....,id_dim_n from fact_nombre
group by
CUBE(id_dim_1,....,id_dim_n,cantidad)
```

Figura 5 Código SQL de una vista materializada.

La vista materializada conforma el cubo de datos utilizando la información que se encuentra en la tabla de hecho, computando el soporte para cada una de las posibles combinaciones del ítemset. Dentro del select de la vista materializada se encuentra el cálculo del soporte que no es más que la suma de la cantidad de apariciones de una combinación de itemset determinada entre la cantidad de apariciones de todas las combinaciones de itemset, se multiplica dicho resultado por 1.0 con el fin de obtener resultados decimales. Para agrupar se utiliza la cláusula group by cube con el fin de obtener todas las combinaciones posibles de unir ítemsets.

En el caso de estudio que se está trabajando, la vista materializada que conforma el cubo de datos se define en la figura 6.

```
create MATERIALIZED VIEW cube_censo as select 1.0*(sum(cantidad))/(select sum(cantidad) from fact_censo) as soporte,
id_dim_persona,id_dim_relacion,id_dim_estado_civil,id_dim_capital_ganado,id_capital_perdido,
id_dim_pais,id_dim_tipo_trabajo,id_dim_horas_trabajo,id_dim_estudio from fact_censo
group by
CUBE(id_dim_persona,id_dim_relacion,id_dim_estado_civil,id_dim_capital_ganado,id_capital_perdido,
id_dim_pais,id_dim_tipo_trabajo,id_dim_horas_trabajo,id_dim_estudio,cantidad)
```

Figura 6 Código SQL de la vista materializada cube_censo.

Después de definido el almacén y el cubo de datos se requiere definir el procedimiento de extracción, transformación y carga asociado al poblado de las tablas en la estructura relacional. Se utiliza la

herramienta PDI (Pentaho Data Integration). En el caso de estudio fueron necesarias diez transformaciones, las cuales son listadas a continuación:

En la tabla 7 se muestra la transformación correspondiente a la dimensión país.

Tabla 7 Dimisión país.

Campos de La dimensión	**Campos correspondientes fuente de datos**	**Transformación**
Id_dim_pais: Integer		llave subrogada
pais: varchar	pais: String	
version: Integer		Automático
data_from: timestamp		Automático
data_to: timestamp		Automático

En la tabla 7 se encuentran reflejados los campos de la dimensión país y las transformaciones necesarias a partir de las fuentes de datos. La transformación llave subrogada será común en todas las dimensiones, en ella se genera un id autoincremental para cada una de las tuplas en la dimensión y no será detallada en las siguientes dimensiones. De igual forma las transformaciones marcadas automáticas para la versión, el data_form y el data_to serán común a todas las dimensiones. La versión es un número consecutivo que indica cuántas cargas se han realizado, data_from la fecha en que se realizó la transformación anterior y por último data_to la fecha en que se está realizando la transformación actual. En cuanto al campo país almacenado como un string en la fuente no se requieren transformaciones para que quede recogido como un varchar en la dimensión. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 7.

Figura 7 Transformación de la dimensión país.

Tabla 8 Dimensión persona.

Campos de La dimensión	**Campos correspondientes fuente de datos**	**Transformación**
Id_dim_persona: Integer		llave subrogada
Grupo_etario: varchar	edad: Integer	Discretizacion con 3 valores de igual frecuencia
sexo: Integer	Sexo: string	
raza: varchar	Raza: string	
minoria: boolean	Raza: string	False si “White” sino True
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

La dimensión persona y las transformaciones necesarias para llenar los campos a partir de la fuente de datos se encuentran reflejada en la tabla 8. La transformación grupo etario es la agrupación de las personas en joven, adulto o anciano dependiendo de su edad, una persona será considerada joven si la edad no excede los 35 años, adulta si su edad se encuentra en el rango de 35 a 60 años y anciana si su edad excede los 60 años. La transformación minoría es un campo de clasificación que especifica si la persona dependiendo de su raza es parte de la minoría, teniendo en cuenta la fuente de datos seleccionada. El campo sexo almacenado como un string en fuente no requiere de ninguna transformación y queda recogido como varchar en la dimensión correspondiente. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 8.

Figura 8 Transformación de la dimensión persona.

La tabla 9 recoge la dimensión estudio y las trasformaciones para llenar los campos a partir de la fuente de datos. El campo de estudio no necesita transformación por lo que queda recogido en la dimisión como un varchar proveniente del string correspondiente en la fuente de datos. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 9.

Tabla 9 Dimensión estudio.

Campos de La dimensión	**Campos correspondientes fuente de datos**	**Transformación**
Id_dim_estudio: Integer		llave subrogada
estudio: varchar	estudio: String	
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

Figura 9 Transformación de la dimensión estudio.

La tabla 10 recoge los campos de la dimensión relación y las transformaciones que fueron necesarias para llenarla a partir de la fuente de datos. El campo relación no necesita ninguna transformación por lo que queda recogida en la dimensión como un atributo varchar tomado del string correspondiente de la fuente de datos. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 10.

Tabla 10 Dimensión relación.

Campos de La dimensión	**Campos correspondientes fuente de datos**	**Transformación**
Id_dim_relacion: Integer		llave subrogada
relacion: varchar	estudio: String	
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

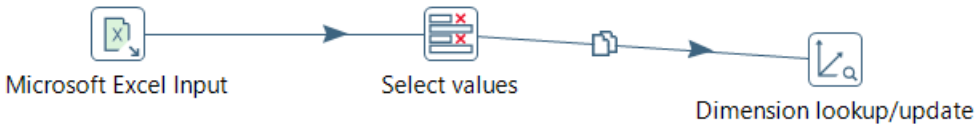

Figura 10 Transformación de la dimensión relación.

La tabla 11 contiene los campos de la dimensión estado_civil y las transformaciones necesarias para llenarla a partir de la fuente de datos. El campo estado civil no necesita ninguna transformación por lo que queda almacenado en la dimensión como un atributo varchar a partir del string que le corresponde en la fuente. El campo actualmente_casado toma valor true si el estado_civil correspondiente tiene alguno de los siguientes valores (Married-civ-spouse, Married-spouse-absent, Married-AF-spouse o Separated) o falso en cualquier otro caso. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 11.

Tabla 11 Dimisión estado civil.

Campos de La dimensión	Campos correspondientes fuente de datos	Transformación
Id_dim_estado_civil: Integer		llave subrogada
Estado_civil: varchar	Estado_civil: String	
Actualmente_casado: boolean	Estado_civil: String	
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

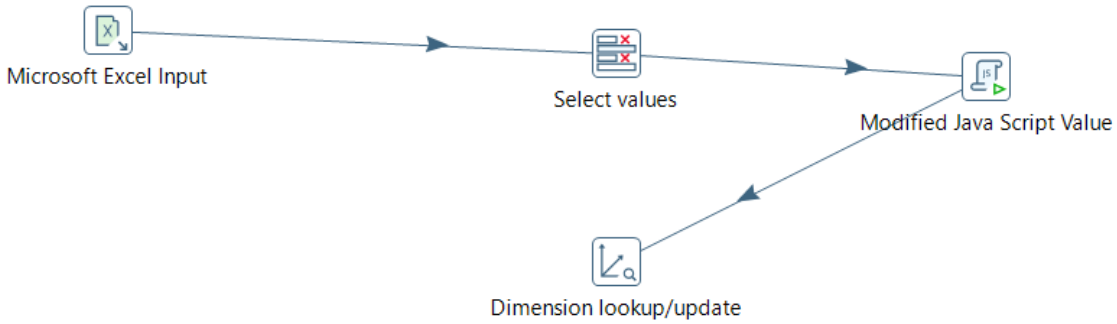

Figura 11 Transformación estado civil.

Tabla 12 Dimensión tipo_trabajo.

La tabla 12 recoje los campos de la dimensión tipo_trabajo y todas las transformaciones que fueron necesarias para llenarla. El campo trabajo no necesita transformación por lo que se almacena en la dimensión como varchar correspondiéndose con el string de la fuente. Los campos trabajo_privado y trabajo_remunerado toman valor true o false correspondiendo con el campo trabajo proveniente de la fuente de datos, trabajo_privado será true solo si trabajo tiene el valor Private y trabajo_remunerado toma valor true solo si trabajo tiene valor Without-pay. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 12.

Campos de La dimensión	Campos correspondientes fuente de datos	Transformación
Id_tipo_trabajo: Integer		llave subrogada
trabajo: varchar	trabajo: String	
Trabajo_privado: boolean	trabajo: String	True if trabajo=”Private”
Trabajo_remunerado: boolean	trabajo: String	True if trabajo !=”Private”
Data_from: timestamp		Automático
Data_to: timestamp		Automático

Figura 12 Transformación tipo_trabajo.

La tabla 13 recoge los campos de la dimensión horas_trabajo y todas las transformaciones que fueron necesarias para llenarla. El campo rango puede tomar valores dependiendo del rango en el que se encuentren las horas trabajadas por las personas durante la semana. Las personas que trabajen menos de 20 horas se considera que trabajan media jornada, las que trabajen entre 20 y 40 horas trabajan jornada completa y las que trabajen más de 40 se les considera que trabajan jornada extra. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 13.

Tabla 13 Dimensión horas_trabajo.

Campos de La dimensión	Campos correspondientes fuente de datos	Transformación
Id_dim_horas_trabajo: Integer		llave subrogada
Estado_civil: varchar		
Rango: String	Horas: integer	
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

Figura 13 Transformación de la dimensión horas_trabajo.

En la tabla 14 se recogen los campos de la dimensión capital_perdido, así como las transformaciones que fueron necesarias para llenarlos. El campo rango clasifica en 3 grupos las pérdidas monetarias de una persona recogido en el campo capital_perdido de la fuente de datos. Se considera una función para discretizar el capital_perdido con 3 categorías e igual distribución. La distribución de los rangos queda de la siguiente forma: las personas con un capital_perdido menor que 1000.00 unidades se considera que fue bajo su capital_perdido, las que su capital_perdido se encuentra en el rango de 1000.00 a 3000.00 se considera que su capital_perdido fue medio y por ultimo las personas con un capital_perdido que exceda las 3000.00 unidades se considera que su capital_perdido fue alto. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 14.

Tabla 14 Dimensión capital_perdido.

Campos de La dimensión	Campos correspondientes fuente de datos	Transformación
Id_dim_capital_perdido: Integer		llave subrogada
Rango: String	Capital_perdido: integer	
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

Figura 14 Transformación de la dimensión capital_perdido.

En la tabla 15 se recogen los campos de la dimensión capital_ganado, así como las transformaciones que fueron necesarias para llenarlos. Se considera una función para discretizar el capital_perdido con 3 categorías e igual distribución para definir el campo capital ganado de la dimensión, para realizar esta transformación se obtienen los datos recogidos en el campo capital_ganado proveniente de la fuente. La distribución de los rangos queda de la siguiente forma: las personas con un capital_ganado menor que 1000.00 unidades se considera que fue bajo su capital_ganado, las que su capital_ganado se encuentra en el rango de 1000.00 a 50000.00 se considera que su capital_ganado fue medio y por último las personas con un capital_ganado que exceda las 50000.00 unidades se considera que su capital_ganado fue alto. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 15.

Tabla 15 Dimensión capital_ganado.

Campos de La dimensión	Campos correspondientes fuente de datos	Transformación
Id_dim_capital_ganado: Integer		llave subrogada
Rango: String	Capital_perdido: integer	Discretizar
version: Integer		Automático
Data_from: timestamp		Automático
Data_to: timestamp		Automático

Figura 15 Transformación de la dimensión capital_ganado.

La tabla 16 recoge los campos correspondientes con el hecho y las trasformaciones que fueron necesarias para llenarlo. En el hecho se recogen los id correspondientes con cada una de las dimensiones, además en el campo cantidad recoge la cantidad de apariciones de cada combinación posible de los id de las dimensiones. Los pasos necesarios para realizar las trasformaciones de la dimensión se muestran en la figura 16.

Tabla 16 Tabla del hecho.

Campos del Hecho:	**Fuente:**	**Transformación:**
Id_dim_capital_ganado: Integer	Id_dim_capital_ganado: Integer	
Id_dim_capital_perdido: Integer	Id_dim_capital_perdido: Integer	
Id_dim_horas_trabajo: Integer	Id_dim_horas_trabajo: Integer	
Id_tipo_trabajo: Integer	Id_tipo_trabajo: Integer	
Id_dim_estado_civil: Integer	Id_dim_estado_civil: Integer	
Id_dim_relacion: Integer	Id_dim_relacion: Integer	
Id_dim_pais: Integer	Id_dim_pais: Integer	
Id_dim_persona: Integer	Id_dim_persona: Integer	
cantidad: Integer	Cantidad: Integer	

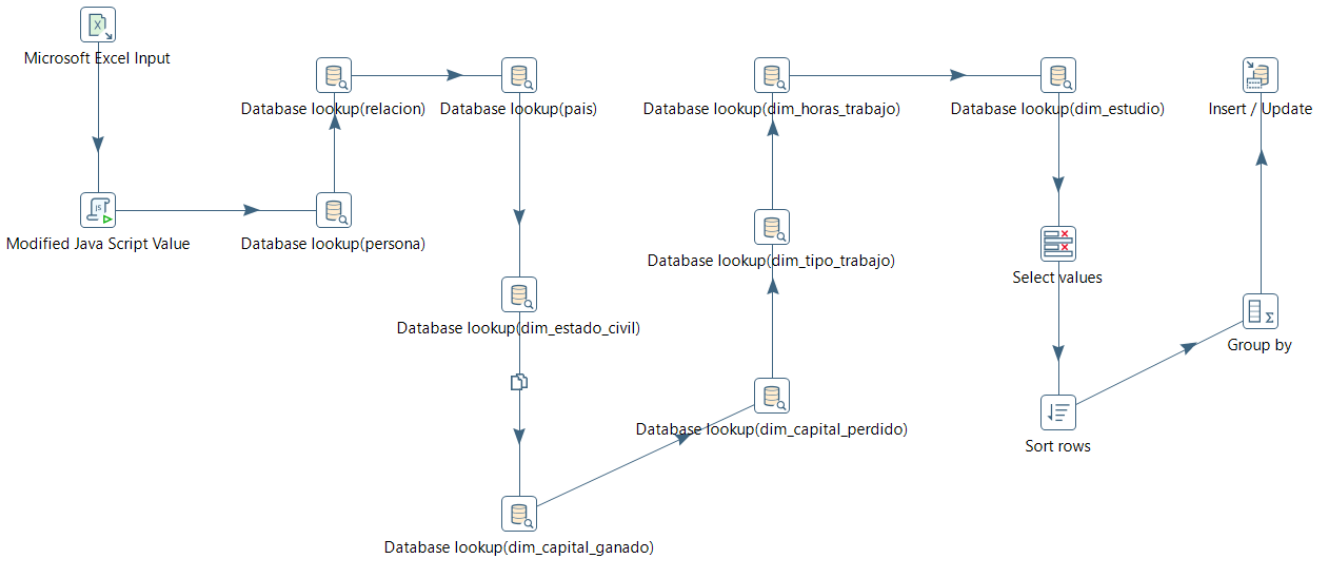

Figura 16 Transformación hecho censo.

2.2.3 Vistas minables

Una de las ventajas de la propuesta está asociada a la obtención de vistas minables a partir de la utilización de las funciones OLAP. En la propuesta las funciones OLAP serán implementadas de la siguiente forma:

Roll-up: esta operación se utiliza para aumentar granularidad, o sea aumentar el nivel de la jerarquía. Se utiliza para resumir el nivel de información y de esta forma se adapta el nivel de detalle del hecho.

Un ejemplo de cubo datos que surgiría al aplicar la operación Roll-up seria cuando el usuario quisiera conocer las reglas de asociación asociadas a las minorías.

```
select id_dim_persona, id_dim_tipo_trabajo, id_dim_estudio, id_dim_horas_trabajo, id_dim_pais, id_dim_capital_ganado, id_capital_perdido,
 id_dim_capital_ganado, id_dim_relacion, sum(fact_censo.cantidad)
 from fact_censo NATURAL join dim_persona
 GROUP BY dim_persona.minoria,id_dim_persona, id_dim_tipo_trabajo, id_dim_estudio, id_dim_horas_trabajo, id_dim_pais, id_dim_capital_ganado,
 id_dim_capital_ganado, id_dim_relacion
```

Figura 17 Código de cubo de datos resultado de la operación Roll-Up.

Drill-Down: Esta operación navega la jerarquía y por tanto la granularidad en el sentido inverso de la anterior. Lo que significa que disminuye el tamaño del grano en el cubo de datos. Es necesario asentar que no puede ir más allá que el nivel mínimo de la granularidad con que han sido almacenados los datos. Esto significa que partiendo del cubo de datos inicial no es posible realizar esta operación.

Dice: selecciona un área del datacube basada en condiciones sobre los valores de las dimensiones. Estas condiciones afectan a los valores sobre los que se aplican y los que están relacionados que pertenecen a otros niveles.

Un ejemplo de un cubo que surgiría a partir de una operación Dice sería cuando el usuario quisiera conocer las reglas de asociación, pero no le interesa la información asociada a la dimensión relación, país, estado civil, capital ganado y capital perdido.

```
select id_dim_persona, id_dim_tipo_trabajo, id_dim_horas_trabajo, id_dim_estudio, cantidad from fact_censo
```

Figura 18 Código de la operación Dice.

Slice: Esta operación reduce la dimensionalidad del almacén o sea se realiza cuando el análisis no depende de todas las dimensiones que contenga el almacén.

Un ejemplo de un cubo de datos que surgiría a partir de una operación Slice sería si el usuario solo le interesara la información de las personas que pertenezcan a la raza blanca.

```
SELECT * from fact_censo, dim_persona WHERE fact_censo.id_dim_persona = dim_persona.id_dim_persona and dim_persona.raza = 'White'
```

Figura 19 Código de la operación Slice.

2.2.4 Minado de reglas de asociación

Dada una base de datos de transacciones, es interesante descubrir asociaciones importantes entre elementos tales que la presencia de algunos elementos en una transacción provoque la presencia de otros elementos en la misma transacción. La formalización del problema de extracción de reglas de asociación requiere la definición de un grupo de conceptos previos.

Sea $I = \{i1, i2, \ldots, im\}$ un conjunto de literales, llamados elementos. Sea D un conjunto de transacciones, donde cada transacción T es un conjunto de elementos tal que T ⊆I. Una transacción T contiene un conjunto de elementos X si y solo si X ⊆ T.

Sea $I = \{i1, i2, \ldots, im\}$ un conjunto de literales y D un conjunto de transacciones definidas sobre I. Una regla de asociación es una implicación de la forma X → Y, donde $X \subset I, Y \subset I \; y \; X \cap Y = \emptyset$. Una regla X → Y tendrá en D una confianza con valor c si el $c\%$ de las transacciones en D que contienen X también contienen Y. La regla tendrá un soporte con valor s en el conjunto de transacciones D si el s% de las transacciones en D contienen a los elementos $X \cup Y$. La extracción de reglas de asociación se ha descompuesto normalmente en dos pasos:

- Descubrir los conjuntos de elementos frecuentes, es decir, los conjuntos de elementos con soporte superior a un umbral prefijado.
- Utilizar los conjuntos de elementos frecuentes para generar reglas de asociación.

El primer punto es el que consume mayores recursos de tiempo y espacio. La generación de reglas una vez se tienen los conjuntos de elementos frecuentes es simple.

Aprovechando las facilidades de cálculo del soporte a partir de métrica reflejada en el cubo de datos la obtención de los itemset frecuentes se limita a una consulta SQL sobre la vista materializada que contiene el cubo restringiendo el soporte.

Por ejemplo, para obtener los itemset frecuentes con soporte mayor que 0.1 sobre la vista materializada cube_censo la consulta seria:

```
select * from cube_censo where soporte> 0.1
```

Figura 20 Restricción de soporte.

Para minar las reglas de asociación se implementó la función PRMA en PLPGSQL.

2.2. Conclusiones parciales

En este capítulo se estableció un modelo teórico para la extracción de reglas de asociación en el modelo multidimensional, seleccionándose los elementos fundamentales para la construcción de un data cubo, igualmente se definieron las operaciones OLAP con las cuales se realizan transformaciones sobre un cubo de datos. Al reunir todos estos elementos se construyó una función llamada PRMA en PLPGSQL para la extracción de reglas de asociación sobre dicho modelo multidimensional. La aplicación de estos elementos sobre el dataset Adult, como prueba de conceptos, permiten asegurar que el procedimiento PRMA es válido para extraer reglas de asociación sobre cubo de datos multidimensionales.

CAPÍTULO 3. VALIDACIÓN DE LA PROPUESTA

3.1. Introducción

La validación de la propuesta se realizará haciendo uso de un método cuasi experimental en el que se medirá el tiempo de respuesta necesario para obtener un modelo de reglas de asociación en dos escenarios, uno utilizando la variante de extracción de reglas de asociación propuesta en esta tesis y dos utilizando la variante de extracción de reglas de asociación.

En un segundo momento se aplican trasformaciones a los datos para simular interacción del usuario y una vez más se ejecutan las tres variantes de extracción de reglas de asociación evaluando el tiempo de respuesta.

3.2 Datasets

Para realizar los experimentos se utilizan 4 datasets disponibles en el repositorio de la UCI (http://www.archive.uci.edu).

Tabla 17 Descripción de los algoritmos.

Datasets	Cantidad de filas	Atributos	Atributos numéricos
Adult	30162	15	6
Brest cáncer	699	10	0
Zoo	101	18	0
Nursery	306	4	2

La tabla 21 muestra la descripción de los dataset que son utilizados para la experimentación, todos los datos fueron convertidos a transacciones, las transacciones vacías fueron eliminadas y los atributos numéricos fueron discretizados, los datos resultantes fueron normalizados.

3.3 Extracción de reglas sobre el cubo general.

Para este experimento se construye un cubo de datos con la información asociada para cada uno de los datasets se extraen las reglas para distintos valores de soporte y confianza y utilizando tres métodos diferentes Apriori, Fp-Growth y PRMA evaluando en cada caso el tiempo de ejecución.

La tabla 18 contiene la información asociada donde cada fila representa una ejecución y las columnas tienen el significado siguiente:

La tabla 18 muestra los resultados del primer experimento:

- La primera columna contiene el nombre de los datasets
- La segunda columna muestra el nombre de los algoritmos utilizados en cada ejecución
- La columna número tres contiene los umbrales de soporte y confianza utilizados para extraer los itemsets frecuentes.
- La columna cuatro contiene la cantidad de reglas generadas por el algoritmo correspondiente
- La última columna contiene los tiempos de ejecución en minutos que demoras la extracción de los itemsets frecuentes por cada algoritmo.

Tabla 18 Resultados del primer experimento.

Datasets	Algoritmo	Soporte confianza	Cantidad reglas	Tiempo ejecución
Adult	Apriori	0.1	7855	15.32
Adult	Fp-growth	0.1	7855	7.58
Adult	PRMA	0.1	7855	6.12
Brest cáncer	A priori	0.1	1300	14.36
Brest cáncer	Fp-growth	0.1	1300	7.16
Brest cáncer	PMRA	0.1	1300	6.15
Zoo	A priori	0.1	8268	17.25
Zoo	Fp-growth	0.1	8268	8.14
Zoo	PMRA	0.1	8268	7.24
Nursery	A priori	0.1	6422	10.33
Nursery	Fp-growth	0.1	6422	6.11
Nursery	PMRA	0.1	6422	5.58
Adult	Apriori	0.2	7822	14.53
Adult	Fp-growth	0.2	7822	7.03
Adult	PMRA	0.2	7822	5.36
Brest cáncer	A priori	0.2	1294	13.44
Brest cáncer	Fp-growth	0.2	1294	6.47
Brest cáncer	PRMA	0.2	1294	5.39
Zoo	Apriori	0.2	8255	16.55
Zoo	Fp-growth	0.2	8255	7.49
Zoo	PRMA	0.2	8255	6.42

Nursery	A priori	0.2	6399	10.03
Nursery	Fp-growth	0.2	6399	5.23
Nursery	PRMA	0.2	6399	4.49
Adult	Apriori	0.3	7750	14.10
Adult	Fp-growth	0.3	7750	5.38
Adult	PRMA	0.3	7750	4.55
Brest cáncer	Apriori	0.3	1270	13.05
Brest cáncer	Fp-growth	0.3	1270	5.04
Brest cáncer	PRMA	0.3	1270	4.10
Zoo	A priori	0.3	8236	16.05
Zoo	Fp-growth	0.3	8236	6.47
Zoo	PRMA	0.3	8236	5.38
Nursery	A priori	0.3	6200	9.54
Nursery	Fp-growth	0.3	6200	4.33
Nursery	PRMA	0.3	6200	3.07
Adult	Apriori	0.4	7615	13.47
Adult	Fp-growth	0.4	7615	4.56
Adult	PRMA	0.4	7615	3.47
Brest cáncer	A priori	0.4	1300	12.58
Brest cáncer	Fp-growth	0.4	1300	4.25
Brest cáncer	PRMA	0.4	1300	3.29
Zoo	Apriori	0.4	8190	15.49
Zoo	Fp-growth	0.4	8190	6.08
Zoo	PRMA	0.4	8190	5.47
Nursery	Apriori	0.4	5860	9.36
Nursery	Fp-growth	0.4	5860	3.43
Nursery	PRMA	0.4	5860	2.14

En la figura 21 se presenta la gráfica de líneas donde se plotean los tiempos de ejecución contra los umbrales de soporte y confianza para el dataset Adult. Las gráficas 22, 23 y 24 presentan la misma información para los datasets Brest cancer, Zoo y Nursery respectivamente.

El análisis de las gráficas permite asegurar que el desempeño de la propuesta realizada es mejor que el desempeño de los algoritmos Apriori y Fp-Growth. Estos resultados coinciden con

lo esperado teóricamente ya que el paso de minado de itemsets frecuentes con una complejidad exponencial tanto en el Apriori como el Fp-Growth se realiza con complejidad lineal en la propuesta realizada.

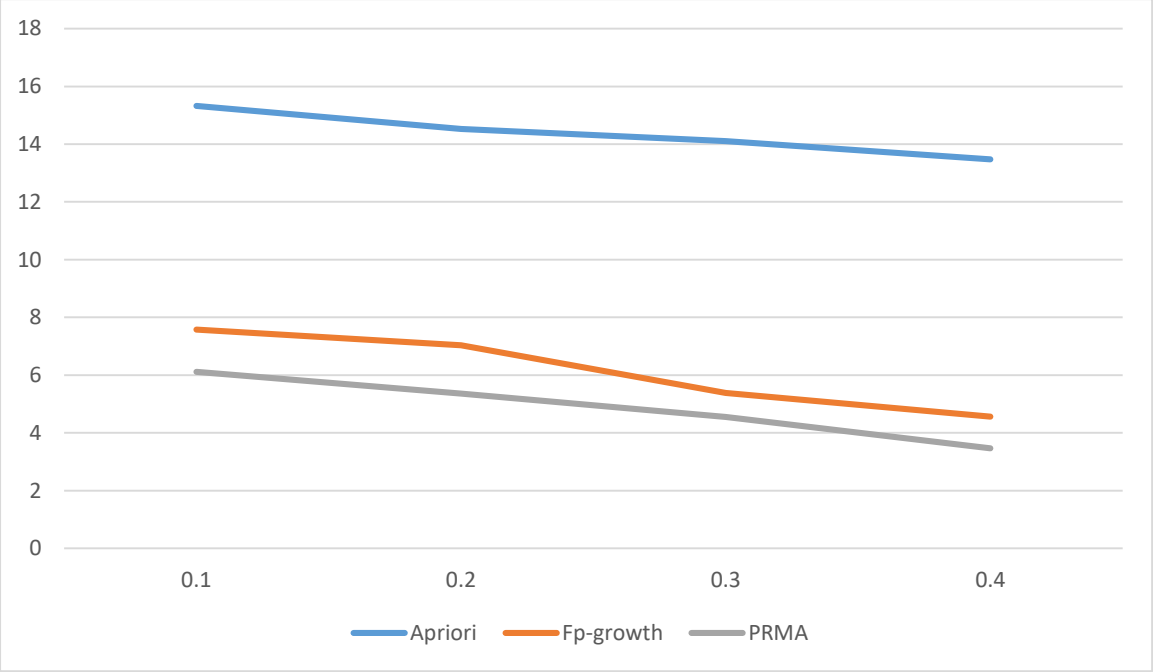

Figura 21 Comparación de los tiempo de ejecución de los algoritmos para el dataset Adult.

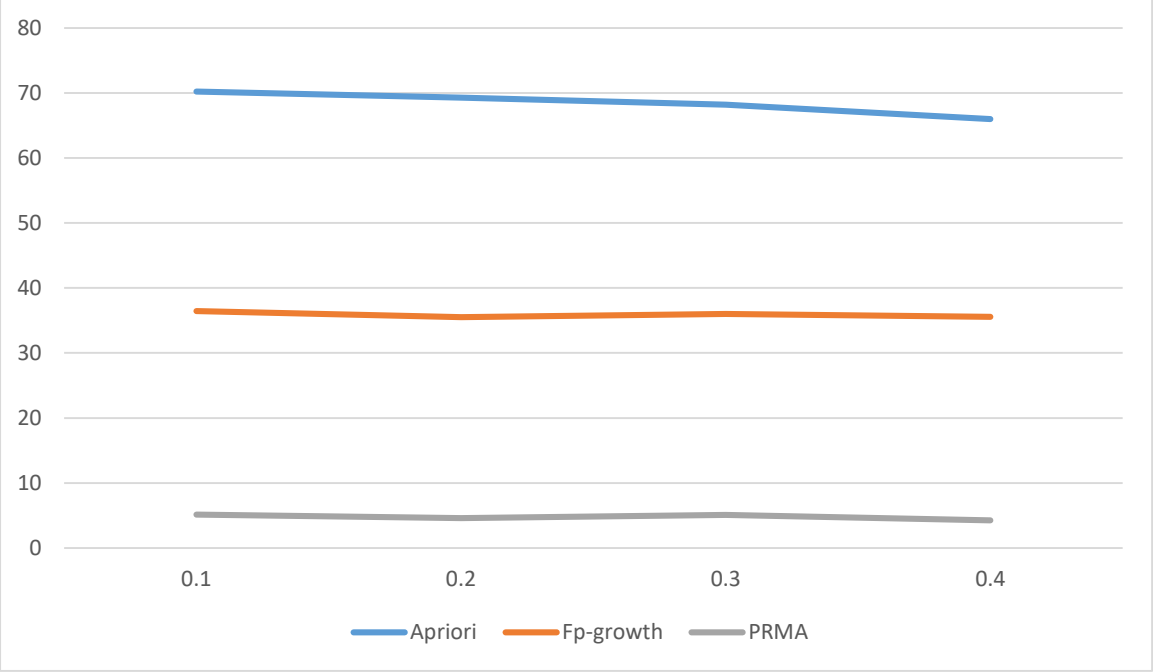

Figura 22 Comparación de los tiempo de ejecución de los algoritmos para el dataset Brest Cancer.

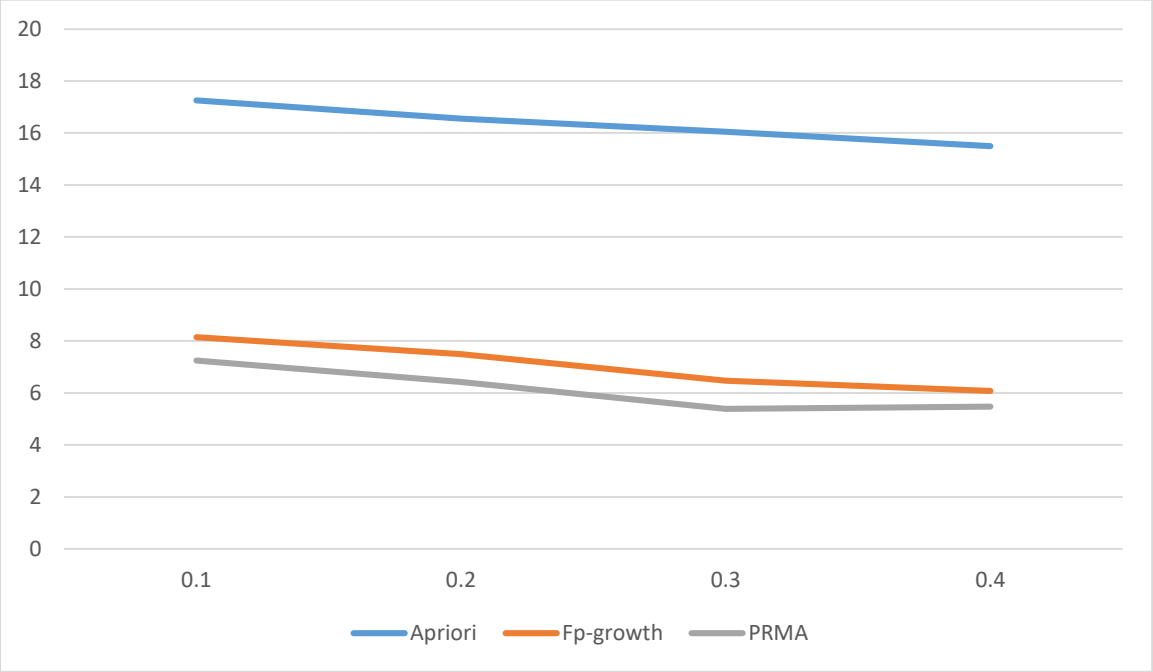

Figura 23 Comparación de los tiempo de ejecución de los algoritmos para el dataset Zoo.

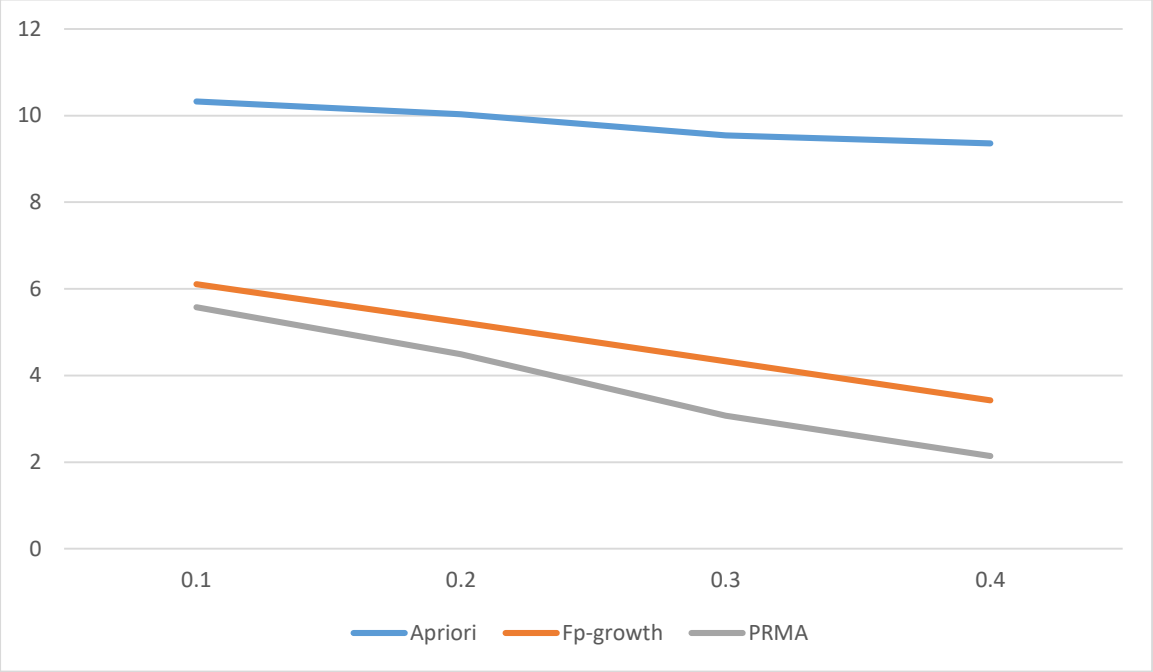

Figura 24 Comparación de los tiempo de ejecución de los algoritmos para el dataset Nursery.

3.4 Extracción de reglas sobre el cubo transformado.

En este experimento se pretende evaluar cómo afecta el rendimiento de los algoritmos de extracción la transformación del dataset de entrada producto de una interacción de experto. Se decidió eliminar

una dimensión de análisis en cada uno de los datasets para ello se construyó un script que procesa los ficheros de texto y elimina una columna y que tomo en cada caso 5 minutos de desarrollo y ejecución. Para el caso de la transformación necesaria en el almacén de datos la consulta asociada a la operación Slice tomo 30 segundos.

Los resultados asociados al experimento se muestran en la tabla 19 que tiene la siguiente estructura:

- La primera columna contiene el nombre de los datasets
- La segunda columna muestra el nombre de los algoritmos utilizados en cada ejecución
- La columna número tres contiene los umbrales de soporte y confianza utilizados para extraer los itemsets frecuentes
- La columna cuatro contiene la cantidad de reglas generadas por el algoritmo correspondiente
- La última columna contiene los tiempos de ejecución en minutos que demoras la extracción de los itemsets frecuentes por cada algoritmo

Tabla 19 Resultados segundo experimento.

Datasets	Método	Soporte y confianza	Cantidad de reglas	Tiempo de ejecución
Adult	Apriori	0.1	7855	72
Adult	Fp-growth	0.1	7855	38.32
Adult	PRMA	0.1	7855	7.21
Brest cáncer	A priori	0.1	1300	70.25
Brest cáncer	Fp-growth	0.1	1300	36.46
Brest cáncer	PMRA	0.1	1300	5.15
Zoo	A priori	0.1	1268	78.58
Zoo	Fp-growth	0.1	1268	39.06
Zoo	PMRA	0.1	1268	8.47
Nursery	A priori	0.1	6422	69.22
Nursery	Fp-growth	0.1	6422	38.17
Nursery	PMRA	0.1	6422	7.45
Adult	Apriori	0.2	7822	71.47
Adult	Fp-growth	0.2	7822	38.00
Adult	PMRA	0.2	7822	7.22
Brest cáncer	A priori	0.2	1294	69.30

Brest cáncer	Fp-growth	0.2	1294	35.52
Brest cáncer	PRMA	0.2	1294	4.59
Zoo	Apriori	0.2	255	77.55
Zoo	Fp-growth	0.2	255	40.46
Zoo	PRMA	0.2	255	9.06
Nursery	A priori	0.2	6399	68.04
Nursery	Fp-growth	0.2	6399	37.33
Nursery	PRMA	0.2	6399	6.02
Adult	Apriori	0.3	7750	69.47
Adult	Fp-growth	0.3	7750	36.26
Adult	PRMA	0.3	7750	6.08
Brest cáncer	Apriori	0.3	1270	68.22
Brest cáncer	Fp-growth	0.3	1270	36.03
Brest cáncer	es	0.3	1270	5.10
Zoo	A priori	0.3	236	76.44
Zoo	Fp-growth	0.3	236	38.23
Zoo	PRMA	0.3	236	7.38
Nursery	A priori	0.3	6200	67.52
Nursery	Fp-growth	0.3	6200	34.12
Nursery	PRMA	0.3	6200	2.02
Adult	Apriori	0.4	7615	68.59
Adult	Fp-growth	0.4	7615	36.10
Adult	PRMA	0.4	7615	4.58
Brest cáncer	A priori	0.4	1300	66.01
Brest cáncer	Fp-growth	0.4	1300	35.57
Brest cáncer	PRMA	0.4	1300	4.24
Zoo	Apriori	0.4	1190	74.03
Zoo	Fp-growth	0.4	1190	37.02
Zoo	PRMA	0.4	1190	5.01
Nursery	Apriori	0.4	5860	74.33
Nursery	Fp-growth	0.4	5860	33.41
Nursery	PRMA	0.4	5860	3.34

En la figura 25 se presenta la gráfica de líneas donde se plotean los tiempos de ejecución contra los umbrales de soporte y confianza para el dataset Adult. Las gráficas 26, 27 y 28 presentan la misma información para los datasets Brest cancer, Zoo y Nursery respectivamente. El aumento de los tiempos de ejecución de los algoritmos Apriori y Fp-Growth se debe a que la transformación realiza demora más de 30 minutos en realizarse.

El análisis de las gráficas permite asegurar que el desempeño de la propuesta realizada es mejor que el desempeño de los algoritmos Apriori y Fp-Growth. Estos resultados coinciden con lo esperado teóricamente ya que el paso de minado de itemsets frecuentes con una complejidad exponencial tanto en el Apriori como el Fp-Growth se realiza con complejidad lineal en la propuesta realizada.

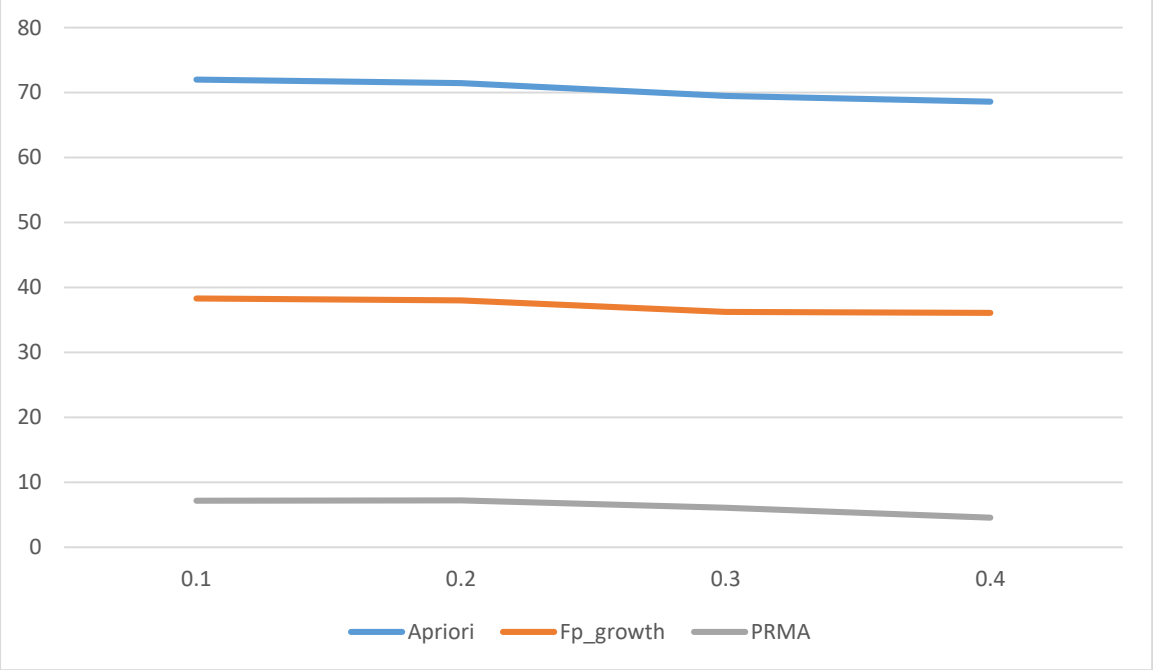

Figura 25 Comparación de los tiempo de ejecución de los algoritmos para el dataset Adult luego de realizada la transformación.

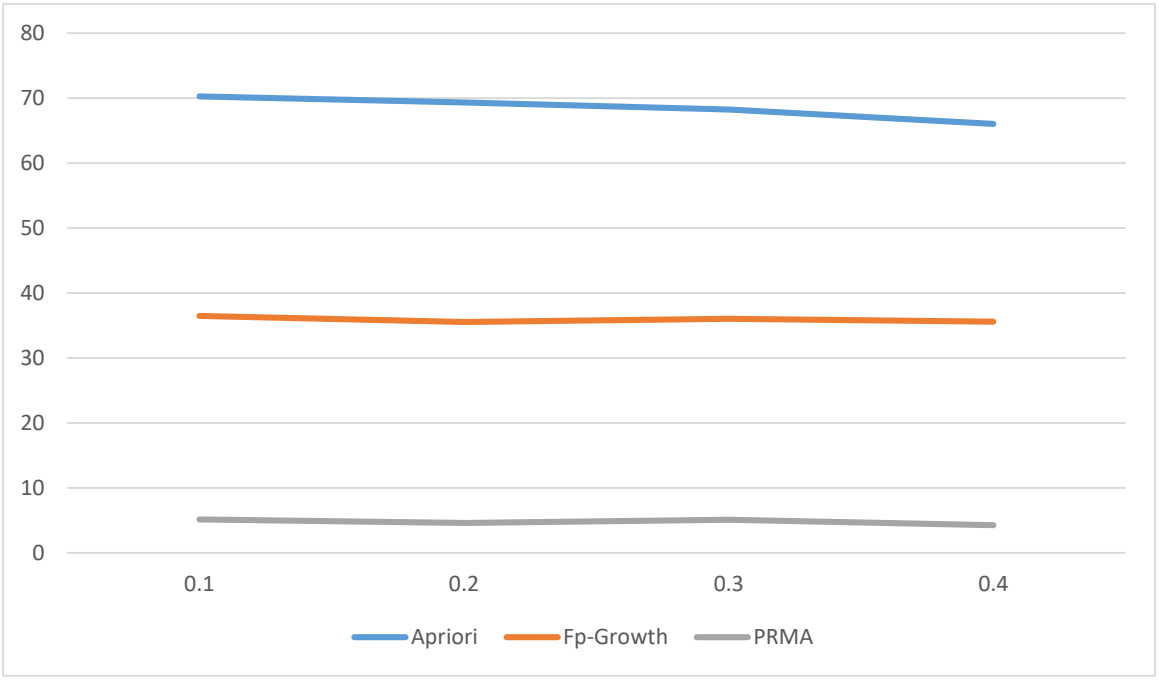

Figura 26 Comparación de los tiempo de ejecución de los algoritmos para el dataset Brest Cancer luego de realizada la transformación.

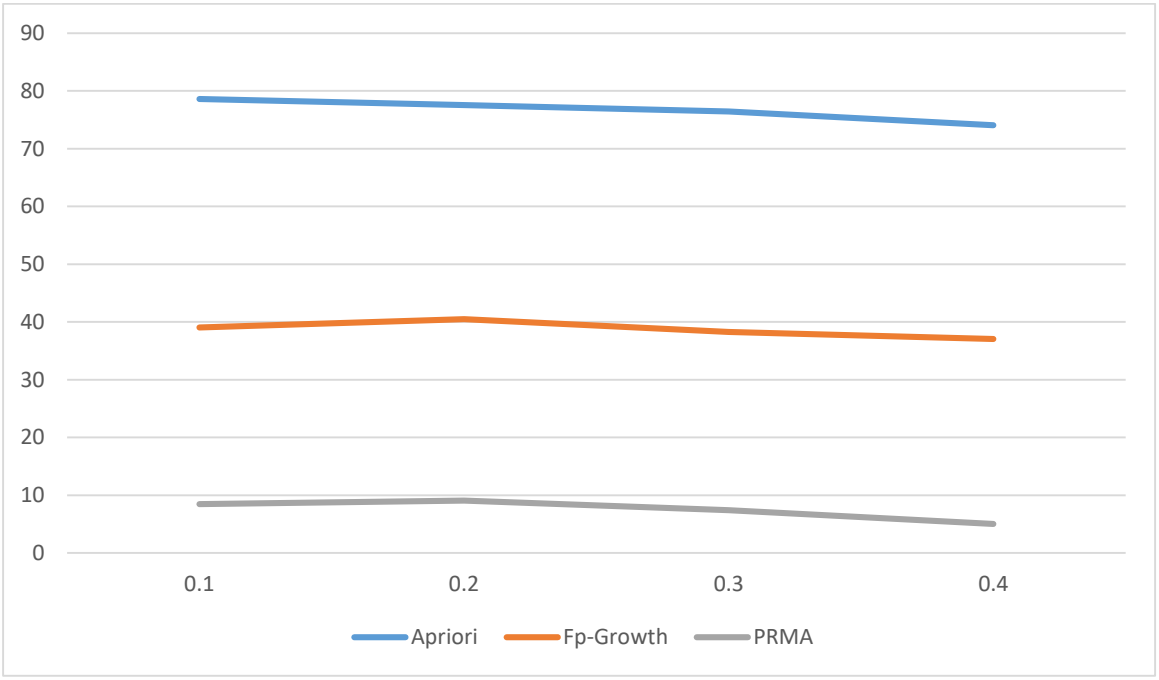

Figura 27 Comparación de los tiempo de ejecución de los algoritmos para el dataset Zoo luego de realizada la transformación.

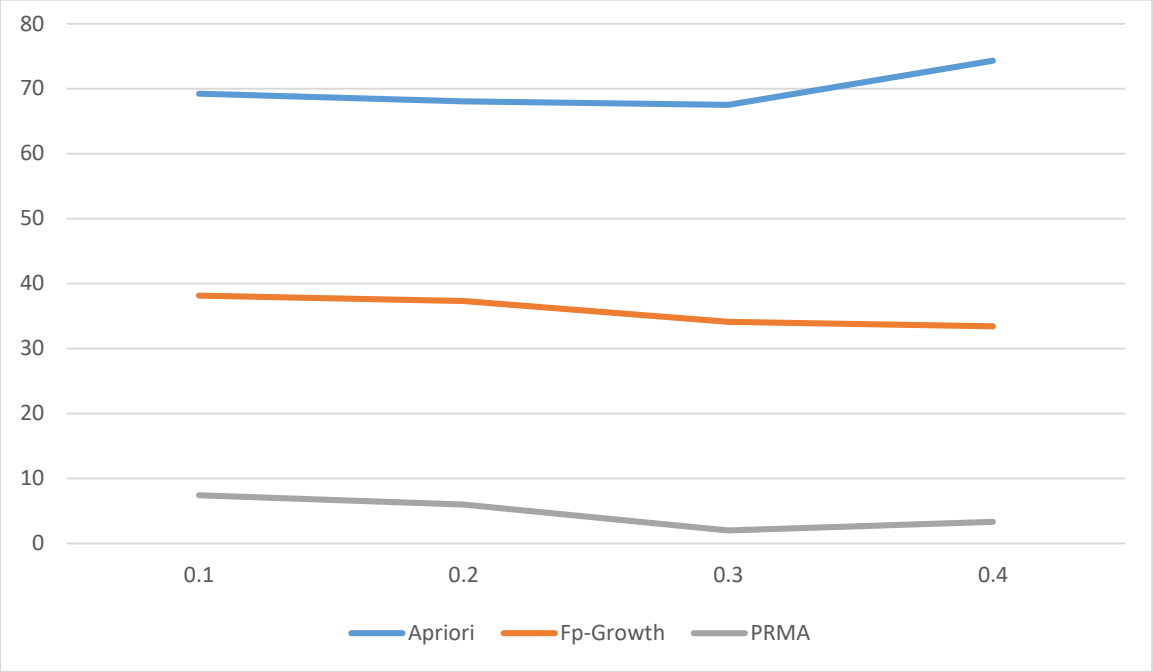

Figura 28 Comparación de los tiempo de ejecución de los algoritmos para el dataset Nursery luego de realizada la transformación.

A partir de los resultados presentados en las figuras de la trece a la veinte se puede comprobar que el método PRMA mejora el rendimiento de las variantes de extracción de reglas de asociación sobre texto plano en todos los casos. Es remarcable la diferencia en los casos en los que se requiere de transformaciones del dataset por intereses del usuario en los que la ventaja del método alcanza varios órdenes de magnitud.

Los resultados alcanzados muestran la ventaja de la utilización de bases de datos y esquemas multidimensionales a la hora de acceder a la información, la incorporación de vistas materializadas con el cómputo del soporte reduce grandemente el tiempo de ejecución para la extracción de reglas de asociación ya que el cálculo de los itemsets frecuentes es la tarea más demandante en términos de recursos.

Se requieren nuevos experimentos, pero los resultados obtenidos son prometedores con vistas a utilizar este procedimiento en el minado de reglas de asociación especialmente en el caso en que los especialistas deseen interactuar con el modelo e introducir restricciones que puedan ser representadas como operaciones OLAP.

3.2. Conclusiones parciales

En este capítulo se desarrolló un experimento para comparar el tiempo de ejecución de la propuesta realizada con la implementación de los algoritmos Apriori y Fp-Growth. El procedimiento PRMA mejora el desempeño de los algoritmos en varias unidades de medidas, con los que se puede concluir que el procedimiento realizado en una opción viable para la extracción de reglas de asociación.

CONCLUSIONES GENERALES

- El estudio realizado para establecer el estado del arte permitió definir que existen un total de 515 artículos en los últimos 5 años orientados al campo de esta investigación de los cuales se analizaron los 15 más relevantes. Los algoritmos de extracción de reglas de asociación más utilizados son Apriori y Fp-Growth y los datasets que utilizan los estudios seleccionados contienen datos reales.
- La aplicación de la prueba de conceptos realizada a la implementación de procedimiento permite asegurar que esta es una aproximación válida para la extracción de reglas de asociación.
- Los experimentos realizados permiten asegurar que el procedimiento PRMA sobrepasa en desempeño a las variantes Apriori y FP-growth.

RECOMENDACIONES

- Extender el modelo multidimensional presentado en este trabajo para que pueda manipular datos difusos.
- Incorporar al modelo otros tipos de reglas de asociación como por ejemplo más reglas de asociación multinivel.

Referencias bibliografía

Abuelyaman, E., & Elgimari, A. (2014). A Prototype for a Data Mining Based Pathfinder to Sudanese Universities. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation* (pp. 119–124). https://doi.org/10.1109/UKSim.2014.65

Arincy, N., & Sitanggang, I. S. (2015). Association rules mining on forest fires data using FP-Growth and ECLAT algorithm. In *2015 3rd International Conference on Adaptive and Intelligent Agroindustry (ICAIA)* (pp. 274–277). https://doi.org/10.1109/ICAIA.2015.7506520

Bawane, G. R., & Deshkar, P. (2015). Integration of OLAP and association rule mining. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1–4). https://doi.org/10.1109/ICIIECS.2015.7193123

Bhavsar, A. R., & Arolkar, H. A. (2014). Multidimensional Association rule based data mining technique for cattle health monitoring using Wireless Sensor Network. In *2014 International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 810–814). https://doi.org/10.1109/IndiaCom.2014.6828074

Bonidia, R. P., Rodrigues, L. A., Avila-Santos, A. P., Sanches, D. S., & Brancher, J. D. (2018). Computational Intelligence in Sports: A Systematic Literature Review. *Advances in Human-Computer Interaction*, *2018*.

Dhanabal, L., & Shantharajah, S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(6), 446–452.

Fisun, M., Kulakovska, I., & Horban, H. (2015). Generation of frequent item sets in multidimensional data by means of templates for mining inter-dimensional association rules. In *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems:*

Technology and Applications (IDAACS) (Vol. 1, pp. 368–375). https://doi.org/10.1109/IDAACS.2015.7340760

Fisur, M., Horban, H., & Dvoretskyi, M. (2018). Methods of Searching for Association Dependencies in Multidimensional Databases. In *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)* (Vol. 2, pp. 88–93). https://doi.org/10.1109/STC-CSIT.2018.8526737

Fournier-Viger, P., Lin, J. C.-W., Vo, B., Chi, T. T., Zhang, J., & Le, H. B. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(4), e1207.

Heaton, J. (2016). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. In *SoutheastCon 2016* (pp. 1–7). IEEE.

Liu, D., Wu, B., Gu, C., Ma, Y., & Wang, B. (2017). A multidimensional time-series association rules algorithm based on spark. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 1946–1952). https://doi.org/10.1109/FSKD.2017.8393066

Liu, F., Zhou, X., Wang, Z., Wang, T., & Zhang, Y. (2018). Identification of Hypertension by Mining Class Association Rules from Multi-dimensional Features. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 3114–3119). https://doi.org/10.1109/ICPR.2018.8545326

Marco-Ruiz, L., Moner, D., Maldonado, J. A., Kolstrup, N., & Bellika, J. G. (2015). Archetype-based data warehouse environment to enable the reuse of electronic health record data. *International Journal of Medical Informatics*, *84*(9), 702–714.

Marinica, C., & Guillet, F. (2010). Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 784–797.

Martínez-Rojas, M., Marín, N., Molina, C., & Vila, M. (2015). Cost analysis in construction projects using fuzzy OLAP cubes. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–8). IEEE.

Noh, B., Son, J., Park, H., & Chang, S. (2017). In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining. *Sustainability*, *9*(11), 2119.

Poli, V. S. R. (2015). Fuzzy data mining and web intelligence. In *2015 International Conference on Fuzzy Theory and Its Applications (iFUZZY)* (pp. 74–79). https://doi.org/10.1109/iFUZZY.2015.7391897

Sarma, H. K. D., & Mishra, S. (2016). Mining time series data with Apriori tid algorithm. In *2016 International Conference on Information Technology (ICIT)* (pp. 160–164). IEEE.

Soni, H. K., Sharma, S., & Jain, M. (2016). Frequent pattern generation algorithms for Association Rule Mining: Strength and challenges. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 3744–3747). IEEE.

Usman, M., Usman, M., & Ahmad, W. (2014). A conceptual model for multi-level mining and visualization of association rules. In *Ninth International Conference on Digital Information Management (ICDIM 2014)* (pp. 175–181). https://doi.org/10.1109/ICDIM.2014.6991409

Usman, Muhammad. (2017). Multi-level mining of association rules from warehouse schema. *Kuwait Journal of Science*, *44*(1).

Wang, H., Zeng, M., Xiong, Z., & Yang, F. (2017). Finding main causes of elevator accidents via multi-dimensional association rule in edge computing environment. *China Communications*, *14*(11), 39–47. https://doi.org/10.1109/CC.2017.8233649

Yokobayashi, R., & Miura, T. (2018). Multidimensional Data Mining Based on Tensor Model. In *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 142–145). https://doi.org/10.1109/AIKE.2018.00031

Yuan, X. (2017). An improved Apriori algorithm for mining association rules. In *AIP conference proceedings* (Vol. 1820, p. 080005). AIP Publishing.

Zeng, Y., Yin, S., Liu, J., & Zhang, M. (2015). Research of improved FP-Growth algorithm in association rules mining. *Scientific Programming*, *2015*, 6.

Zou, Y., Liu, Y., Qin, X., & Ma, S. (2014). Research and application of association rule mining algorithm based on multidimensional sets. In *2014 IEEE 5th International Conference on Software Engineering and Service Science* (pp. 557–560). https://doi.org/10.1109/ICSESS.2014.6933629