# MÉTODO DE REGLAS DE ASOCIACIÓN PARA EL ANÁLISIS DE AFINIDAD ENTRE OBJETOS DE TIPO TEXTO

**Mario Orozco Bohórquez**

**Universidad de la Costa CUC**

**Departamento de ciencias de la Computación y Electrónica**

**Posgrado Maestría en Ingeniería**

**Barranquilla – Colombia**

**2017**

# MÉTODO DE REGLAS DE ASOCIACIÓN PARA EL ANÁLISIS DE AFINIDAD ENTRE OBJETOS DE TIPO TEXTO

**Mario Orozco Bohórquez**

**Notas del autor**

**Trabajo de Investigación presentado para optar por el título de:**

**MAGISTER EN INGENIERÍA**

**Directores**

**PhD. Emiro de la Hoz Franco**

**MSc. Alexis Kevin de la Hoz Manotas**

**Universidad de la Costa CUC**

**Departamento de ciencias de la Computación y Electrónica**

**Posgrado Maestría en Ingeniería**

**Barranquilla – Colombia**

**2017**

**NOTA DE ACEPTACION**

_____

_____

_____

_____

**JURADO**

_____

**JURADO**

_____

**JURADO**

**Barranquilla octubre 2017**

## DEDICATORIA

Principalmente a Dios. Él es el que día a día me permite ver un nuevo amanecer. Una nueva razón para sonreír y vivir. Por enseñarme lo maravilloso que es estar vivo y poder compartir con tu familia y seres queridos cada momento y cada detalle.

A mi madre, por su apoyo incondicional y porque es la responsable del gran porcentaje de mis logros académico-profesionales. Aún hoy está pendiente de mí y no cesa en brindarme sus sabios consejos, manifestarme y demostrarme su apoyo incondicional.

A mi esposa por ser mi columna vertebral, el pilar fundamental del hogar y permanecer a mi lado incondicionalmente tras lluvias y tormentas. Por ser mi bastón y mi paño de lágrimas. Por enseñarme el significado de familia y lo importante que es permanecer unidos a pesar de los tropiezos. Anhelo que la presente tesis signifique para ella, un éxito compartido.

A mis tres hijos, Mario Edgardo, Alberto Mario y Matías. Son mi orgullo y gran motivación y me impulsan a cada día superarme académica y profesionalmente y dar lo mejor de mí día a día. Sin ustedes nada de esto habría sido posible. Son mi vida y mi razón de vivir.

Un muy especial agradecimiento al ingeniero Fabio Enrique Mendoza Palechor por su apoyo incondicional y paciencia en la realización del presente trabajo de investigación.

A mi tutor PhD. Emiro de la Hoz Franco y Co-tutor MSc. Alexis Kevin de la Hoz Manotas por su trabajo y dedicación en este proyecto de investigación de principio a fin.

**Ing. Mario Orozco Bohórquez**

## Contenido

## Índice de Ecuaciones

## Índice de Figuras

# Índice de Tablas

## Resumen

La minería de datos es considerada una herramienta para extraer conocimiento en grandes volúmenes de información. Uno de los análisis realizados en minería de datos son las reglas de asociación, cuyo propósito es buscar co-ocurrencias entre los registros de un conjunto de datos.

Su principal aplicación se encuentra en el análisis de canasta de mercado, donde se establecen criterios para la toma de decisiones a partir del comportamiento de compra de los clientes. Algunos de los algoritmos son Apriori, Frequent Parent Growth, QFP Algorithm, CBA, CMAR, CPAR. Estos algoritmos han sido diseñados para analizar bases de datos estructuradas; en la actualidad, diversas aplicaciones requieren el procesamiento de datos no estructurados, como es el caso de los objetos de tipo texto. La investigación planteada tiene como propósito generar un método que permita establecer la relación existente entre los elementos que componen un objeto de tipo texto, para la adquisición de información relevante a partir del análisis de fuentes masivas de datos del mismo tipo.

**Palabras Clave**

Reglas de Asociación, Minería de Texto, co-occurrencia, Minería de Datos, Bases de Datos Estructuradas, Algoritmos de Asociación, Métodos de Asociación.

## Abstract

Data mining is considered a tool to extract knowledge in large volumes of information. One of the analyzes performed in data mining is the association rules, whose purpose is to look for co-occurrences among the records of a set of data.

Its main application is in the analysis of market basket, where criteria for decision making are established based on the buying behavior of customers. Some of the algorithms are A priori, Frequent Parent Growth, QFP Algorithm, CBA, CMAR, CPAR. These algorithms have been designed to analyze structured databases; At present, various applications require the processing of unstructured data known as text type Objects. The purpose of this research is to generate a method to establish the relationship between the elements that make up an object of text type, for the acquisition of relevant information from the analysis of massive data sources of the same type.

**Key Words**

Association Rules, Text Mining, co-occurrence, Data Mining, Structured Data Base, Algorithms of Association, Methods of Association.

# 1 Introducción

La minería de datos se ha considerado una alternativa para dar solución a diferentes problemáticas que requieren el análisis de grandes volúmenes de datos. Las tareas principales de la minería de datos están orientadas a procesos de clasificación, pronóstico, segmentación y asociación. En este proyecto se aborda específicamente la asociación, también conocida como análisis de afinidad. Este análisis es utilizado en minería de datos para buscar co-ocurrencias entre registros de información. Su principal aplicación se encuentra en el análisis de canasta de mercado, donde se establecen criterios para la toma de decisiones teniendo en cuenta el comportamiento de compra de los clientes. Este análisis es realizado por medio de métodos de reglas de asociación [Agrawal, Imieliński & Swami, 1993], los cuales permiten obtener información del tipo "Clientes que compraron el libro A también compraron el libro B". Algunos de los algoritmos para la creación de reglas de asociación son A priori, Frequent Parent Growth, QFP Algorithm, CBA, CMAR, CPAR, entre otros.

En la literatura, además de encontrar diversas aplicaciones de reglas de asociación a datos estructurados como las compras de clientes, se pueden encontrar aplicaciones de reglas de asociación a datos de tipo texto, logrando la adquisición de un conocimiento útil a partir de fuentes masivas de datos, lo cual permite visualizar la exploración a un nuevo campo de trabajo para el tratamiento de datos de tipo texto.

En la actualidad, diversas aplicaciones de texto requieren el procesamiento de un tipo de dato no estructurado, el cual está constituido por caracteres alfanuméricos. Para el almacenamiento, procesamiento y administración de objetos de tipo texto se han generado diferentes soluciones que permiten buscar eficientemente datos de interés para el usuario final a través de sistemas de

información los cuales analizan grandes volúmenes de datos. Teniendo en cuenta que la semántica de un objeto de tipo texto debe ser expresada por todos los elementos que la componen [Zhuang, Yang & Wu, 2008], en este proyecto se aborda la tarea de analizar la afinidad entre objetos de tipo texto evaluando la heterogeneidad del tipo de información que componen este tipo de objetos.

Un objeto de tipo texto está constituido por caracteres alfanuméricos [Grosky, 1997]. La semántica de un objeto de tipo texto debe ser expresada por todos los elementos que la componen, es decir por el texto tal como lo menciona [Yang, 2008], [Zhuang, Yang & Wu, 2008], [Hunter & Choudhury, 2003].

[Zhuang, Yang & Wu, 2008] manifiesta que, aunque los elementos analizados solo son de tipo texto, existen correlaciones semánticas que permiten relacionarlos con otros. De acuerdo a lo anterior la minería de datos permite el estudio o análisis de diferentes elementos que componen a los Objetos de tipo texto.

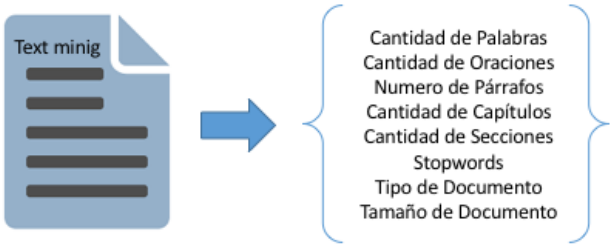
La minería de objetos de tipo texto debe evaluar los diferentes elementos que componen a el objeto, buscando representaciones que sean interpretables por herramientas informáticas. A continuación, se presentan una serie de ilustraciones donde es posible observar las características que se pueden extraer a partir de los diferentes elementos que conforman un objeto de tipo texto.

El texto puede ser representado por un vector de características que contenga información básica del texto como la cantidad de palabras, cantidad de oraciones, numero de párrafos, cantidad de capítulos, numero de secciones, stopwords, tipo de documento de texto, tamaño.

***La Figura 1*** es un elemento de un objeto de tipo texto en el cual se define como características la cantidad de palabras, cantidad de oraciones número de párrafos, cantidad de capítulos, cantidad de secciones, stopwords, tipo de documento, tamaño del documento, lo

anterior genera un vector de características donde se almacena los valores obtenidos en cada uno de los documentos analizados.

$$featureVector = \{cantpalabras, cantoraciones, numparrafos\}.$$




***Figuras 1.*** *Extracción de características de un documento de texto*
Fuente: Elaboración Propia

Debido a su compleja estructura, los objetos de tipo texto requieren un procesamiento complejo para obtener la semántica de sus contenidos. Modelar la estructura del objeto de tipo texto es importante debido a que a partir de dicho modelo es posible por ejemplo medir la similitud entre varios objetos teniendo en cuenta algunas características de los elementos que lo componen [Grosky, 1997].

En la literatura consultada se evidencia que se han tratado de diseñar distintos sistemas de información que permitan el tratamiento de los objetos de tipo texto. Dichos sistemas deben tener como características capacidad para la presentación, almacenamiento y comunicación de los elementos que conforman el objeto de tipo texto, lo anterior teniendo en cuenta condiciones de extrema heterogeneidad en los datos analizados [Little & Ghafoor, 1990].

De acuerdo a lo mencionado por algunos autores se ha generado como necesidad el desarrollo de diferentes aplicaciones, métodos o recursos los cuales permitan un análisis coherente de datos de tipo texto para extraer información relevante,  en [Hu, Xu, Liu, Mei, Chen & Luo, 2014] se

plantea un método inteligente que permita mantener de forma organizada los datos de tipo texto contenidos en bases de datos los cuales son de gran tamaño, en [Zheng, Wang, & Gao, 2006] se plantea una analogía entre recuperación de imágenes y texto, a partir de ello se propone un enfoque visual basado en frases para recuperar imágenes contenidas en objetos deseados, en [Jiang & Tan, 2009] se menciona como problema la fusión de la información que se encuentra alojada en la web, teniendo en cuenta que el proceso de análisis, recopilación y rastreo de la información requiere de un mayor esfuerzo, generalmente las investigaciones en dicho tópico de trabajo se ha enfocado en la generación de resumen multi-documento teniendo en cuenta solo características del texto, mientras que el enfoque propuesto en la investigación hace referencia al tema de la imagen y la asociación de texto, en [Alghamdi, Taileb & Ameen, 2014] se propone la fusión de información de texto y visual, en la investigación planteada se utiliza la combinación de diferentes técnicas de minería de datos se toma como fuente de información 54500 imágenes.

La minería de datos es un área de las ciencias de la computación que se encarga de la exploración de datos con el propósito de obtener información que genere nuevo conocimiento, la minería de datos se enfoca en dos tipos de análisis los cuales corresponden a Análisis Predictivo y Análisis Descriptivo. Dentro del análisis Predictivo es posible hacer tareas de minera de datos tales como clasificación y predicción mientras que en el análisis Descriptivos las tareas de minería de datos a realizar correspondes a segmentación y asociación. Este trabajo se enfoca en la búsqueda de reglas de asociación para objetos de texto la cual hace parte del análisis descriptivo.

Inicialmente las Reglas de Asociación fueron mencionadas por [Agrawal, Imieliński & Swami, 1993], Las reglas de asociación fueron utilizadas para el análisis de las cestas de mercado a partir de dicho estudio se logró establecer criterios para tomar decisiones en base al

comportamiento de compra de los productos por parte de los clientes lo que permitió establecer las reglas del orden o posición de los productos en los diferentes lugares del supermercado obteniendo resultados positivos.

En la literatura consultada autores como [[Agrawal, Imieliński & Swami, 1993], [Agrawal & Srikant, 1994], [Agrawal & Shafer, 1996], [Han, Pei, Yin & Mao, 2004], [Hanguang & Yu, 2012], [Xiang, 2012], [Tsuji, Takizawa, Sato, Ikeuchi, Ikeuchi, Yoshikane & Itsumura, 2014], definen las Reglas de Asociación bajo el mismo fundamento. Una Regla de Asociación se puede definir como $\mathbf{I} = \{\mathbf{i1}, \mathbf{i2}, \mathbf{i3}, \mathbf{i4}\}$ siendo **I** los conjuntos de Ítems. Sea la base de datos un conjunto de transacciones en las que cada **T** es un subconjunto de **I**. Una Regla de Asociación puede ser inferida de la forma $\boldsymbol{X} \rightarrow \boldsymbol{Y}$, donde **X**, **Y** es un subconjunto de **I** y $\mathbf{X} \cap \mathbf{Y} = \boldsymbol{\varphi}$. El conjunto de elementos **X** es llamado antecedente y **Y** es llamado consecuente. Las dos propiedades que generalmente son consideradas en las Reglas de Asociación son el Soporte y la Confianza.

De acuerdo a la definición anterior, dado el conjunto de elementos $\mathbf{A} = \{\mathbf{objeto1}, \mathbf{objeto2}, \mathbf{objeto3}, \mathbf{objeto4}, \mathbf{objeto5}\}$, siendo **A** el conjunto de Ítems. Sea la base de datos un conjunto de transacciones en las que cada **T** es un subconjunto de **A**. En este caso una regla de asociación puede ser inferida de la forma $\boldsymbol{X} \rightarrow \boldsymbol{Y}$, donde **X, Y** es un subconjunto de **A** y $\mathbf{X} \cap \mathbf{Y} = \boldsymbol{\varphi}$**.** El conjunto de elementos **X** es llamado antecedente y **Y** es llamado consecuente. A continuación, se presenta una breve ilustración del ejemplo planteado.

Según [Agrawal, Imieliński & Swami 1993], el problema de descubrir todas las reglas de asociación se puede descomponer en dos sub-problemas. En primer lugar, encontrar para un conjunto de elementos el número de transacciones que contienen dicho conjunto de elementos. Los conjuntos de elementos con el soporte mínimo son llamado gran conjunto de elementos, y

otros pequeños conjuntos de elementos. El segundo problema hace referencia a utilización de los grandes conjuntos de elementos para generar las reglas deseadas.

De acuerdo a lo Mencionado por [Mustafa, Nabila, Evans, Saman & Mamat 2006] las Reglas de Asociación hacen parte de las técnicas de minería de datos, su propósito se basa en el descubrimiento del grado de asociación de la información analizada. La información contenida en la base de datos a veces parece poco frecuente, pero a su vez es muy asociada a un conjunto de datos específicos.

[Xu, Li & Shaw 2011] menciona que la minería de reglas de asociación ha contribuido a muchos avances en el área de descubrimiento de conocimiento. Sin embargo, la calidad de las reglas de asociación descubiertos es una gran preocupación y ha atraído a más y más atención recientemente. Uno de los problemas con la calidad de las reglas de asociación descubiertos es el enorme tamaño del conjunto de reglas extraído.

[Domíngues, 2004] manifiesta que las reglas de asociación tienen como propósito encontrar tendencias que puedan ser utilizadas para entender y explorar patrones de comportamiento en los datos que son objeto de análisis, cabe destacar que no todas las reglas de asociación representan un patrón en los datos estudiados. Una regla representara un patrón siempre y cuando la misma cumpla determinados criterios definidos en los algoritmos de inducción, los cuales también expresan la fiabilidad de las reglas.

# 2 Definición del problema

## 2.1 Problema de investigación

Actualmente el análisis de diferentes tipos de datos se realiza con técnicas supervisadas conocidas, adicionalmente el análisis de esta información en el caso de objetos de tipo texto, se utilizan los metadatos. Es decir, que el análisis de la información se hace sobre los metadatos y no directamente sobre la información. En el presente trabajo de investigación, se está utilizando la información y no los metadatos para encontrar relación entre los diferentes objetos de tipo texto. Lo anteriormente mencionado genera una problemática y es que, al momento de trabajar con el contenido, la cantidad de datos, variables o atributos, aumenta de forma exponencial lo cual se convierte en un problema. A pesar de la gran cantidad de datos que utilizar las reglas de asociación y el algoritmo que utilizas, es capaz de encontrar relación entre los documentos.

La etapa de pre-procesamiento realizada para la creación de la data set es compleja de acuerdo a que se debió realizar una serie de pasos para obtener la información contenida en los documentos y no los metadatos. El primer paso fue seleccionar una data sets que contiene información relacionada al hábitat de cada animal dentro de un zoológico, así como su forma de crecimiento, alimentación, entre otros. Se almacenaron los documentos de la data sets descargados y se extraen las características de los elementos que componen el objeto de texto y se genera un archivo .arff con los atributos o características del texto para encontrar posibles relaciones entre ellos. Se aplica la bolsa de palabras para reducir sustancialmente la cantidad de atributos a tener en cuenta. WEKA es la herramienta utilizada para tales efectos. Es una herramienta de tipo software para el aprendizaje automático y minería de datos.

En la revisión literaria, se ha encontrado la aplicación de reglas de asociación a diferentes formatos de tipo texto. Las reglas de asociación se han aplicado a texto analizando características sobre información básica como la cantidad de palabras, cantidad de oraciones, número de párrafos, cantidad de capítulos, número de secciones, stopwords, tipo de documento, tamaño, entre otras [Tang, Yan & Yuan 2013].

De la exploración literaria se ha encontrado que el análisis de afinidad entre objetos de tipo texto debe afrontar diversos desafíos como:

- Un objeto de texto está compuesto por componentes heterogéneos que deben ser evaluados para determinar la afinidad entre los objetos.
- Las aplicaciones existentes de reglas de asociación a texto, han analizado metadatos más que el contenido de estos formatos.
- El tratamiento de contenido de texto comúnmente está afectado por problemas de alta dimensión, ya que se puede extraer un gran volumen de características de cada formato.

En la literatura consultada autores como [Karabatak, & Ince 2009] [Karabatak, & Ince, 2009a] [Chaves, Ramírez, Górriz, Puntonet, & Alzheimer's Disease Neuroimaging Initiative, 2012] [Dua, Singh, & Thompson, 2009] [Malik & Kender, 2006] [Yin & Li, 2006] entre otros, implementan las reglas de asociación teniendo en cuenta los metadatos de algunos elementos que componen el objeto de texto, lo anterior permite cuestionarse acerca de lo novedoso que puede ser generar un método que contemple el contenido de al menos dos elementos que componen el objeto de tipo texto. Adicionalmente en [Hu, Xu, Liu, Mei, Chen, & Luo, 2014] [Tesic, Newsam, & Manjunath, 2003] [Zheng, Wang, & Gao, 2006] [Jiang & Tan, 2009] [Alghmdi, Taileb, & Ameen, 2014], abordan el estudio de los objetos de tipo texto analizando al menos dos

elementos que componen dicho objeto, en los cuales se puede apreciar un como referente el uso de información de tipo texto para análisis de información compuesta por múltiples-datos, adicionalmente se plantea como solución del análisis de la información la utilización de las reglas de asociación complementadas con algoritmos como por ejemplo clustering, lo cual permite analizar una metodología de trabajo común donde inicialmente los datos son clasificados o asignados a unas categorías y posteriormente se procede a establecer la relación entre los datos de tipo texto.

Debido a su compleja estructura, los objetos de tipo texto requieren un procesamiento complejo para obtener la semántica de sus contenidos. Modelar la estructura del objeto de tipo texto es importante debido a que a partir de dicho modelo es posible por ejemplo medir la similitud entre varios objetos teniendo en cuenta algunas características de los elementos que lo componen [Tesic, Newsam, & Manjunath,, 2003].

Se evidencia que se han tratado de diseñar distintos sistemas de información que permitan el tratamiento de los objetos de tipo texto, dichos sistemas deben tener como características, capacidad para la presentación, almacenamiento y comunicación de los elementos que conforman el objeto de tipo texto, lo anterior teniendo en cuenta condiciones de extrema heterogeneidad en los datos analizados [Yang, Zhuang, Wu, & Pan, 2008].

Teniendo en cuenta la estructura de un objeto de texto, se evidencia que solo existe una metodología específica para el tratamiento de datos de texto e imágenes con diferentes métricas de evaluación ignorando las características de los demás objetos las cuales pueden generar un conocimiento de interés producto del análisis de los datos.

Finalmente los objetos de texto son un recurso disponible y relevantes los cuales son de vital importancia y estudiados de forma generalizada en el ámbito científico, teniendo en cuenta que a

partir de ellos se pueden extraer información que beneficien de forma significativa a la sociedad u individuos que los utilicen, es por ello que es de crucial importancia establecer o formalizar, la manera en como los objetos de tipo texto deben ser tratados, como deben ser aplicados los diferentes algoritmos para obtener el mayor beneficio en cuanto a información de nuevo conocimiento se refiere.

Para aportar a la solución de los problemas detectados en la aplicación de reglas de asociación a objetos de tipo texto, se formula la siguiente pregunta de investigación:

¿Un método de Reglas de Asociación que evalúe el contenido de los formatos existentes en un objeto de tipo texto será más exacto que evaluar los metadatos del objeto?

## 2.2 Justificación

Un objeto de tipo texto está constituido por caracteres alfanuméricos, [Grosky, 1997]. La semántica de un objeto de tipo texto debe ser expresada por todos los elementos que la componen, es decir por el texto, tal como lo menciona [Yang, Zhuang, Wu, & Pan, 2008], [Zhuang, Yang, & Wu, 2008], [Hunter & Choudhury, 2003]. Actualmente el acceso a internet ha crecido de una forma desmedida, de acuerdo a las cifras arrojadas en el año 2014 por el banco mundial BIRF-AIF por cada 100 personas hay 40,7 que son usuarios que acceden a la red mundial, mientras que Internet World Stats informa que en el año 2015 el 46.4% de las personas a nivel mundial tienen acceso a internet, dicho crecimiento ha provocado diferentes áreas de estudio las cuales son abordadas por investigadores que pretenden dar solución a necesidades específicas, en consecuencia el estudio de los Objetos de tipo texto se ha convertido en una de las áreas de estudio de interés a nivel mundial.

Las técnicas de minería de datos se han convertido en una herramienta para la exploración de conocimientos de los diferentes tipos de elementos que componen a un objeto de tipo texto, en la literatura existente es posible evidenciar la aplicación de diferentes métodos los cuales son evaluados de acuerdo a diversas métricas tales como su eficiencia, capacidad de respuesta y capacidad de computo utilizada a la hora de identificar los diferentes tipos de información. El propósito de esta investigación se basa en la utilización de técnicas de minería de datos como lo son las Reglas de Asociación con la finalidad de generar un modelo que permita identificar la afinidad entre diversos Objeto de tipo texto y de esta forma lograr la adquisición de nuevo conocimiento a partir de los análisis realizados teniendo en cuenta la heterogeneidad de los datos y al gran cantidad de características que pueden surgir a partir del contenido de los diferentes objetos de tipo texto.

# 3 Fundamentación del proyecto

## 3.1 Objetivos

### 3.1.1 Objetivo General

Comparar métodos de reglas de asociación para el análisis de afinidad entre objetos de tipo texto.

### 3.1.2 Objetivos Específicos

- Identificar los diferentes algoritmos de inducción de reglas de asociación y sus medidas de interés o métodos de asociación más utilizados.
- Identificar los diferentes métodos de reglas de asociación para objetos de tipo texto
- Aplicar las reglas de asociación para el análisis de afinidad entre objetos de tipo texto con un caso de estudio.

## 3.2 Metodología

Para el proceso de la investigación se plantearon dos etapas que permitirán un seguimiento detallado de la forma en que se desarrollara la temática de investigación que tiene como propósito el diseño de un modelo que permita el análisis de afinidad entre objetos de texto

tomando como base el mejor algoritmo de Reglas de Asociación, a continuación, se realiza una breve descripción de la metodología a seguir.

Fase No 1(Extracción de características): Esta etapa tiene como finalidad realizar el proceso de indagación o búsqueda de las características correspondientes a objetos de tipo texto. A continuación, se describen las actividades que permitirán la culminación de la fase:

- Análisis de características de documentos de tipo texto.
- Preparación del conjunto de datos que se tomara como fuente de información de los objetos de tipo texto.

Fase No 2 (Construcción y Validación de modelo propuesto): Esta etapa tiene como propósito la construcción, entrenamiento y validación del modelo o prototipo planteado para el reconocimiento de Objetos de tipo texto basado en algoritmo de reglas de asociación que arroje mejores resultados los cuales permitan lograr porcentajes confiables al momento de analizar los diferentes elementos del objeto de tipo texto. A continuación, se describen las actividades que permitirán la culminación de la fase:

- Desplegar el método de Reglas de asociación para Objetos de tipo texto.
- Someter a Prueba el método de Reglas de asociación para Objetos de tipo texto.

## 3.3 Alcance y contribuciones

La formulación de un Modelo de Reglas de Asociación para el análisis de afinidad entre objetos de tipo texto es una contribución de nuevo conocimientos en el área de las tecnologías de la información y comunicación. A continuación, se presentan las contribuciones de la investigación propuesta:

- Método de Reglas de Asociación para Objetos de tipo texto: cuyo propósito está relacionado con la generación de las reglas que permita el análisis de afinidad entre Objeto de tipo texto.
- Representación de Objetos de tipo texto: con el propósito de generar la representación de la información presente en los distintos tipos de objetos de tipo texto.
- Criterio de selección de Reglas de Asociación para Objetos de tipo texto: selección del mejor método de reglas de asociación aplicado a información de tipo texto.
- Aplicación en caso de estudio: permite evaluar el método basado en reglas de asociación formulado, dicho método será probado con un caso de estudio.

A continuación, se presentan los productos esperados:

- Estado del arte de los objetos de tipo texto y reglas de asociación el cual será publicado como artículo de investigación de revisión en revista indexada en ISI o SCOPUS.
- Modelo de Minería de Datos propuesto.
- Modelo de Minería de Datos aplicado a un caso de estudio.
- Publicación de resultados finales de la investigación propuesta en revista indexada en ISI o SCOPUS.

## 3.4 Difusión de resultados

Publicaciones en proceso:

- ASSOCIATION RULES METHOD FOR TEXT OBJECTS: REVIEW.
- ASSOCIATION RULES METHOD FOR ANALYSIS OF AFFINITY BETWEEN TEXT TYPE OBJECTS.

# 4 Marco referencial

## 4.1 Marco conceptual

### 4.1.1 Conceptos Iniciales de Asociación

La asociación es una clase de problema de minería de datos en la cual se busca encontrar ítems que aparezcan juntos en transacciones de un determinado conjunto de datos. De esta manera, se establecen reglas que indican dependencias entre los ítems de dicho conjunto de datos, siendo las Reglas de Asociación la forma más natural de representar dichas asociaciones. Actualmente, la inducción de reglas de asociación es una de las técnicas más utilizadas en procesos de descubrimiento del conocimiento, el cual puede ser aplicado en distintas áreas del conocimiento. Además, una regla de asociación posee una formalización que permite su fácil interpretación, incluso por no expertos en minería de datos.

Las reglas de asociación fueron inicialmente introducidas por Agrawal. [Agrawal, Imieliński, & Swami, 1993] Para proporcionar una manera novedosa de obtener información acerca de la compra de productos. El objetivo de los autores fue el de facilitar soporte a la decisión en la determinación de la disposición de productos en los estantes de un supermercado. Así, la información proporcionada por las reglas de asociación acerca de los productos que generalmente eran adquiridos en una misma compra se utilizaba para situarlos en sitios próximos [Harrison, 1998] [Gonçalves & Léren, 1999]. Por lo tanto, a la tarea de descubrir reglas de asociación en transacciones comerciales se denominó “análisis de la cesta de compra” (market basket analysis).

[Agrawal, Imieliński, & Swami, 1993] mostraron que una regla de asociación expresa, en un conjunto de datos, la probabilidad de que la ocurrencia de un conjunto de ítems implique la ocurrencia de otro conjunto de ítems. Los autores definieron la siguiente formalización: considérese un conjunto de ítems (también llamados de atributos) $I = \{i1,i2,i3,...,in-1,in\}$, donde cada elemento "i" perteneciente a I puede asumir valores binarios 1 o 0 (verdadero o falso) que expresan respectivamente su presencia o ausencia en el conjunto. Además, sobre los elementos de I, se tiene un conjunto de transacciones $T = \{t1,t2,t3,...,tn-1,tn\}$, donde cada elemento "ti" perteneciente a T corresponde a un conjunto de ítems presentes en I, tal que $t \subseteq I$. Un ítem es considerado como una instancia de un atributo, o sea, su valor en un determinado registro (u objeto) que represente una transacción.

Así mismo, [Agrawal, Imieliński, & Swami, 1993] se considera que si todos los elementos pertenecientes a un conjunto de ítems A en una transacción "t", entonces el conjunto A es subconjunto de "t", o sea, $A \subseteq t$. Una regla de asociación puede tener una representación equivalente a $A \rightarrow B$, es decir, la existencia de los ítems que pertenecen al conjunto A (antecedente), en una transacción, implican la existencia de los ítems que pertenecen al conjunto B (consecuente), donde $A \subseteq I$ y $B \subseteq I$. Es importante señalar que los ítems de A son distintos de los ítems de B, o sea, $A \cap B = \emptyset$.

[Agrawal, Imieliński, & Swami, 1993] también consideran la existencia de variables cuantitativas en un conjunto de ítems. Para ello, establecieron intervalos de valores para dichas variables y utilizaron dichos intervalos como nuevos atributos con valores binarios. Si la variable numérica toma un valor de un intervalo en la transacción, el atributo binario correspondiente a ese intervalo será verdadero, mientras que los atributos correspondientes al resto de los intervalos serán falsos.

[Domingues, 2004] La inducción de reglas de asociación, tiene el objetivo de encontrar tendencias que puedan ser utilizadas para comprender y explorar patrones de comportamiento en los datos. Sin embargo, no todas las reglas de asociación representan un patrón en los datos. Una regla representará un patrón solamente si la misma cumple determinados criterios definidos en los algoritmos de inducción, los cuales también expresan la fiabilidad de las reglas. Dichos criterios, o medidas de interés, van a ser descritos en el siguiente apartado.

### 4.1.2 Algoritmos de Reglas de Asociación

En la literatura consultada se pueden apreciar diversos algoritmos diseñados para generar Reglas de asociación a partir de distintas fuentes masivas de datos, dentro de los algoritmos identificados es posible destacar la relevancia que ha tenido el algoritmo Apriori [Agrawal, & Srikant, 1994] para la generación de las reglas de asociación, debido a que a partir de dicho método ha sido posible generar nuevos métodos eficientes tales como ECLAT [Zaki, Parthasarathy, Ogihara, & Li, 1997] y FPGrowth [Han, Pei, & Yin, 2000] entre otros. A continuación, se presentan los algoritmos de reglas de asociación consultados en las fuentes literarias.

#### *4.1.2.1 Apriori*

El funcionamiento del algoritmo Apriori empieza con la obtención de los llamados “conjuntos de ítems frecuentes”, los cuales son aquellos conjuntos cuyos ítems superan un umbral que define un valor mínimo para la medida de soporte. Debido al amplio uso del algoritmo Apriori,

desde que se formalizó la inducción de reglas de asociación, la obtención de los conjuntos de ítems frecuentes es una tarea común en dichos algoritmos.

[Agrawal & Srikant 1994] Determinaron una propiedad fundamental al proponer el algoritmo Apriori, mediante la cual se puede afirmar que todo subconjunto de un conjunto de ítems frecuentes también va a ser un conjunto de ítems frecuentes. Por lo tanto, el algoritmo Apriori obtiene en primer lugar los conjuntos de ítems frecuentes de tamaño 1 y, luego, los de tamaño 2 y así sucesivamente hasta que no se encuentren más conjuntos.

Apriori es un algoritmo diseñado para descubrir grandes conjuntos de elementos a través de la realización de varias pasadas sobre los datos [Agrawal & Srikant 1994]. El primer paso hace referencia al conteo de las ocurrencias del elemento con el propósito de determinar el tamaño de 1-itemset (conjunto de elementos). Posteriormente se procede a llamar a la interacción k la cual está compuesta de dos fases. Primero, el tamaño del conjunto de elementos $L_{k-1}$ es encontrado en la interacción (K-1) este paso es usado para generar el candidato del conjunto de elementos $C_k$, el paso anterior se realiza utilizando la función de generación de candidatos de a priori (A priori-gen) la cual recibe como argumentos $L_{k-1}$. Luego, la base de datos es escaneada y se procede a realizar el conteo del soporte “support” de los candidatos en $C_k$. Para lograr un conteo eficiente es necesario determinar los candidatos de $C_k$ que son contenidos en la transacción dada t.

De acuerdo a lo mencionado por [Narvekar & Syed, 2015], Apriori hace parte de los métodos de Reglas de Asociación, el cual genera una cantidad de reglas consideradas como buenas al momento de analizar pequeñas bases de datos, por el contrario se considera como desventaja del algoritmo que a medida que aumenta el tamaño de la base de datos el rendimiento de dicho algoritmo disminuye, esto se debe a que debe escanear toda la base de datos cada vez que se escanea una transacción, es por ello que se considera que el principal problema del algoritmo

Apriori es la utilización de grandes cantidades de datos, problema que también es mencionado en [Bhandari, Gupta & Das, 2015] quien manifiesta que el rendimiento de este algoritmo será muy bajo e ineficiente cuando la capacidad de memoria es limitada con un gran número de transacciones, otro de los problemas del algoritmo Apriori es que adolece de algunas deficiencias a pesar de ser claro y sencillo, la principal limitación es la pérdida de tiempo para mantener gran número de conjuntos de candidatos con mucho conjuntos de elementos frecuentes.

#### *4.1.2.2 Frequent Pattern Growth*

Con la intención de solventar las fragilidades expuestas anteriormente del algoritmo Apriori [Han, Pei & Yin, 2000] propusieron el algoritmo FP-Growth. Para construir un algoritmo de mejor rendimiento que el Apriori, dichos autores propusieron una estructura de datos alternativa, llamada FP-Tree (árbol de patrones frecuentes). Un FP-Tree almacena información acerca de conjuntos de ítems frecuentes de forma compacta y así permite realizar consultas de manera más eficiente. Haciendo uso de dicha estructura de almacenamiento se reduce el coste de computación del proceso de obtención de reglas de asociación, pues no es necesario generar los conjuntos de ítems candidatos a frecuentes y tampoco verificar si ellos superan a un determinado umbral.

[Han, Pei, Yin & Mao, 2004] las transacciones de conjuntos de datos suelen compartir ítems frecuentes y, por lo tanto, el tamaño del FP-Tree correspondiente suele ser mucho menor que el del conjunto de datos original. Además, a diferencia de los métodos basados en el algoritmo Apriori, en ningún caso se generaría un FP-Tree con un número exponencial de nodos.

El algoritmo Frequent Parent Growth, utiliza la filosofía de divide y vencerás, para lo cual construye un árbol de patrones frecuentes y posteriormente calcula los conjuntos de elementos frecuentes. De acuerdo a lo mencionado en [Han, Pei & Yin, 2000][Han, Pei, Yin & Mao, 2004]. La estructura del algoritmo FP-Tree está dada a la existencia de un nodo llamado raíz el cual posee una etiqueta que a su vez posee un valor nulo, a partir del nodo raíz se desprende un conjunto de subárboles los cuales almacenan los valores de las transacciones o registros que están siendo procesados frecuentemente. Adicionalmente existe una tabla cabecera la cual sirve como respaldo para guardar los elementos a los cuales se accede con mayor frecuencia, en cada entrada de la tabla se ingresa el nombre del elemento y el apuntador para el primer nodo del árbol que posee el elemento.

#### *4.1.2.3 QFP Algorithm*

El algoritmo QFP propuesto por [Juan & De-ting, 2010], está compuesta básicamente por dos etapas básicas, para comenzar, se procede a escanear las transacciones de la base de datos, las cuales son transformadas en forma de árbol, como en el algoritmo FP-tree, proceso llamado QFP-tree, En este punto toda la información relacionada entre los ítems de la base de datos es guardada. Para finalizar, el algoritmo QFP-tree se dedica a localizar todas las posibles reglas de asociación. De acuerdo a lo mencionado por [Juan & De-ting, 2010] al comparar los algoritmos QFP-tree y FP-Tree, el resultado es: QFP es más eficiente debido a que solo escanea la base de datos una sola vez, lo que aumenta la capacidad de respuesta del algoritmo propuesto.

#### *4.1.2.4 CBA*

CBA fue el primer algoritmo en el que se formalizó de manera efectiva la integración de conceptos de clasificación y asociación en una única implementación. El algoritmo suele ser utilizado como referencia para el desarrollo de otros trabajos más recientes relativos a la implementación de algoritmos de clasificación basada en asociación, donde se realizan estudios de casos para evaluar la eficiencia de dichos algoritmos.

El algoritmo Clasification Based Asociation CBA, es propuesto por [Liu, Hsu & Ma, 1998] con el propósito de realizar procesos de clasificación tomando como base las reglas de asociación, CBA está compuesto por dos fases, la primera llamada generación de reglas CBA-RG la cual toma como base el algoritmo a priori mencionado por [Agrawal & Srikant, 1994] y la segunda fase es llamada constructor del clasificador CBA-CB. Existen dos versiones del algoritmo CBA-CB, las cuales son denominadas M1 y M2, la versión M2 suele utilizarse cuando la cantidad de datos a procesar y requiere gran capacidad de computo de tal manera que la capacidad de almacenamiento en memoria es superada, mientras que la versión M1 debe ser utilizada cuando la cantidad de datos es menor y en consecuencia la capacidad de memoria a utilizar es menor.

#### *4.1.2.5 CMAR*

Con el objetivo de lograr mejoras en relación a algunas limitaciones existentes en el algoritmo CBA, [Li, Han & Pei, 2001] desarrollaron el algoritmo CMAR(Classification based on Multiple Association Rules -Clasificación basada en Múltiples Reglas de Asociación). El nombre del

algoritmo se refiere a una concepción implantada en el mismo, en la cual no se utiliza una, sino un conjunto de reglas para determinar el atributo etiqueta de un determinado registro.

CMAR también es un algoritmo compuesto por dos etapas, pues la obtención del clasificador y de las reglas de clasificación se realiza en etapas distintas. En la primera etapa se generan reglas a partir del conjunto de entrenamiento, donde se definen umbrales de confianza y soporte mínimo. Para ello, CMAR utiliza una adaptación del algoritmo FP-Growth, en el cual también se construye una estructura adaptada de FP-tree para almacenar los patrones frecuentes durante el proceso de generación de las reglas.

El algoritmo CMAR mencionado por [Li, Han & Pei, 2001], es considerado como una extensión del algoritmo FP-growth,. Cuya principal característica es la utilización de un conjunto de reglas para poder determinar el atributo etiqueta de un determinado registro. La estructura que almacena los conjuntos de ítems frecuentes, en el CMAR, también es análoga a un FP-Tree, excepto que se asignan los valores del atributo etiqueta en los últimos nodos de las ramas del árbol que representan cada conjunto frecuente.

#### *4.1.2.6 CPAR*

El algoritmo CPAR (*Classification based on Predictive Association Rules* - Clasificación Basada en Reglas de Asociación Predictivas), por sus siglas en inglés, fue desarrollado por [Yin & Han, 2003] con el propósito de brindar un método de clasificación basada en reglas de asociación que aportara mayor eficiencia en menor tiempo para procesar conjuntos de datos de mayores proporciones. Los autores se fundamentaron en la realización de experimentos, y en que

la generación y selección de reglas en los algoritmos CBA y CMAR consumían mucho tiempo de procesamiento en bases de datos que poseen un número muy grande de líneas y/o columnas.

Para alcanzar dichos objetivos, implementaron un algoritmo integrado, donde se realiza una sola pasada, en el conjunto de datos, para generar las reglas y obtener el clasificador.

El algoritmo Clasification Based on Predictive Association Rules CPAR, propuesto por [Yin & Han, 2003] es un algoritmo diseñado con el propósito de realizar procesos de clasificación los cuales tomen como base las reglas de asociación predictivas. Una de las ventajas del algoritmo CPAR hace referencia a que genera conjuntos pequeños de gran calidad de reglas predictivas para el conjunto de datos, también evita la generación de reglas redundantes. Para generar las reglas de clasificación CPAR toma como base el algoritmo FOIL [Quinlan & Cameron-Jones, 1993] el cual es un algoritmo de clasificación que se destina a diferenciar registros positivos de registros negativos.

Se puede decir que el FOIL es un algoritmo bastante agresivo. Realiza repetidas búsquedas por la mejor regla para el momento y, al encontrarla, descarta los registros positivos del conjunto de datos que son cubiertos por dicha regla. Un algoritmo agresivo, se caracteriza por intentar obtener una solución óptima para un problema a través de una secuencia de elecciones locales, donde cada elección busca una solución óptima para el momento [Brassard, 1997]. Los algoritmos agresivos pueden no lograr una solución óptima, pues se basan en una estrategia de elecciones locales en las que no se considera el contexto global.

Para componer las reglas, FOIL trata cada regla separadamente, seleccionando cada atributo del conjunto de datos a la vez y evaluando la ganancia que el ítem referente a tal atributo aporta para la regla en análisis. Para calcular la ganancia aportada por un ítem a una regla, se utiliza una

función aritmética que considera el número de registros positivos y negativos que fueron obtenidos antes y después de añadir el ítem a la regla.

Utilizando básicamente los mismos principios de FOIL, [Yin & Han, 2003] desarrollaron el algoritmo PRM (*Predictive Rule Mining* – Minería de Reglas Predictivas) que constituye un aporte a la implementación del algoritmo CPAR. PRM difiere de FOIL respecto a la eliminación de los registros positivos ya clasificados por una regla. En PRM dichos registros no son descartados, sino que se decrementa un factor utilizado para contabilizar la relevancia de cada registro. Dicha versión "ponderada" de FOIL, produce más reglas y cada registro positivo generalmente es cubierto más de una vez.

#### *4.1.2.7 Rapid Association Rule Mining (RARM)*

[Das, Ng, & Woon, 2001] propone un algoritmo llamado Minería de Reglas de Asociación Rápida (RARM), cuya característica hace referencia a el aumento de velocidad en los procesos de aplicación de las reglas de asociación. Para conseguir el aumento de velocidad teniendo en cuenta las reglas cuyos valores soporte y umbral sean bajos, RARM construye una nueva estructura denominada Soporte-Ordenado Trie Ztemset - **SOTrielT**, la estructura diseñada tiene forma de árbol el cual almacena el soporte contando todos los ítems de la base de datos, tomando como inicio el 1-itemsets y 2-itemsets, la información extraída a partir del conjunto de datos se utiliza para actualizar el SOTrielT, finalmente la estructura se ordena de forma descendente de acuerdo a los recuentos del soporte.

#### *4.1.2.8 Partition*

El algoritmo partition en propuesto por [Savasere, Omiecinski & Navathe, 1995], se basa en la idea de recordar que el análisis de la base de datos se debe realizar en una gran cantidad de tiempo debido a que el número de posibles conjuntos de elementos que permitan someter a probar el soporte es exponencialmente grande, teniendo en cuenta que dicho proceso hay que hacerlo en una sola exploración de la base de datos. El algoritmo partition se ejecuta en dos fases, la primera fase consiste en la división lógica de la base de datos en un numero de particiones que no estén superpuestas. Las particiones consideran el tiempo y todo el gran conjunto de elementos de las particiones que fueron generadas. La segunda fase consiste en la genera el soporte para el conjunto de elementos, los cuales posteriormente son identificados.

#### *4.1.2.9 MMAC*

El algoritmo MMAC propuesto por [Thabtah, Cowling & Peng, 2004], consta de tres fases las cuales consisten a la generación de las reglas, aprendizaje recursivo y por último el proceso de clasificación. En la primera fase se procede a análisis de los datos de entrenamiento los cuales permiten generar reglas, posteriormente la segunda fase tiene como propósito el descubrimiento de más reglas que superen los valores mínimos de soporte y confianza, en la fase número tres los conjuntos de reglas derivan en cada iteración los cuales serán fusionados para formar una etiqueta global multi-clase. La mayoría de los trabajos de investigación relativos a clasificación no se dedican al estudio de problemas que presentan múltiples etiquetas.

#### *4.1.2.10 ECLAT*

Algoritmo ECLAT propuesto por [Zaki, Parthasarathy, Ogihara & Li, 1997], utiliza el formato de base de datos verticales a diferencia de los algoritmos Apriori y RARM los cuales funcionan bajo el formato de datos horizontal. Eclat usa un conjunto de elementos bajo un esquema de clustering con el propósito de generar candidatos potenciales a partir del conjunto de elementos. Cada uno de estos clústeres induce una subred, que se recorre mediante la búsqueda de abajo hacia arriba para generar todos los conjuntos de elementos frecuentes. Cada cluster es procesado por completo para poder pasar al siguiente grupo. Una de las principales características del algoritmo Eclat es el bajo consumo de memoria debido a que solo los K-Conjuntos de elementos frecuentes es cargado en memoria.

### 4.1.3 Evaluación de Métodos de Asociación

Después de haber descrito o abordado los algoritmos de inducción de reglas de asociación es necesario describir algunas de las principales medidas de interés o métodos de asociación que son utilizadas, en general, como umbral para obtener reglas de asociación válidas. Las reglas generadas por un algoritmo puede ser un número muy elevado, especialmente si hay muchos ítems en las transacciones, lo cual no facilita que se haga el proceso de KDD (*Knowledge Discovery in Databases*, por sus siglas en inglés) de manera eficiente y fiable. Las reglas de asociación permiten identificar tendencias o relaciones entre los datos almacenados en una base de datos y a partir de las coincidencias encontradas se generan reglas, de acuerdo a lo anterior es importante determinar la calidad de las reglas de tal manera que permitan garantizar una solución

óptima para el análisis del comportamiento de los datos. También se hace necesario establecer una manera de reducir el número de reglas generadas por el algoritmo de inducción. Las medidas de interés o métodos de asociación son utilizadas justamente para solventar este inconveniente, pues las mismas definen criterios para evaluar la calidad de las reglas de asociación y desechar aquellas que no cumplen dichos criterios.

Los métodos de asociación o medidas de interés, en general, son utilizados como parámetros de entrada por algoritmos de inducción de reglas de asociación. Además, se puede hacer una lista ordenada de reglas en la cual se indique la fiabilidad de las mismas basándose en alguna de dichas medidas. Según [Brin, Motwani & Silverstein, 1997], las medidas de soporte y de confianza son las más utilizadas por los algoritmos de inducción de reglas de asociación. Teniendo en cuenta lo expuesto anteriormente, a continuación, se presentan las medidas de evaluación de las reglas de asociación y algunas medidas que surgen como alternativas para evaluar las Reglas de Asociación.

#### *4.1.3.1 Soporte*

El soporte es una medida que contabiliza la frecuencia en la cual los términos de una regla de asociación se encuentran en los datos, es decir, el número de transacciones en las cuales los ítems presentes en una regla ocurren juntos en los datos en relación con el número total de transacciones.

En el caso del Soporte de la regla denominado “Supp”, es una medida de interés para medir la calidad de las reglas de asociación. Dado que $\mathrm{Freq}(x)$ es el número de filas que contienen el Ítem X en la base de datos dada. El soporte del conjunto de elementos de X se define como la

fracción de todas las filas que contienen el conjunto de elementos expresado de la siguiente manera $\mathrm{Freq(X)/D}$. El soporte en Reglas de Asociación está representado por la unión del antecedente X y el consecuente Y.

$$\mathbf{Supp} = (\mathbf{X} \rightarrow \mathbf{Y}) = \frac{\mathbf{X} \cup \mathbf{Y}}{\mathbf{D}}$$

Ecuación 1: Soporte de una Regla de Asociación

[Pinho, 2010], [Kotsiantis & Kanellopoulos, 2006] Definen el soporte como una medida que contabiliza la frecuencia en la cual los términos de una regla de asociación se encuentran en los datos, es decir, el número de transacciones en las cuales los ítems presentes en una regla ocurren juntos en los datos en relación con el número total de transacciones.

**Ejemplo para el Cálculo del Soporte de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo del soporte de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación $(\text{llanta} \rightarrow \text{gps})$ el cálculo del soporte se realiza por medio de la formula $\text{Supp} = (\text{X} \rightarrow \text{Y}) = \frac{\text{X} \cup \text{Y}}{\text{D}}$ donde $X$ es representado por el artículo "llanta" y $Y$ es representado por el artículo "gps", el valor de D corresponde al total de transacciones registradas en la base de datos. De acuerdo a lo anterior se procede al cálculo del soporte el cual se observa a continuación:

$$supp(X \rightarrow Y) = \frac{X \cup Y}{D} = \frac{(llantas \cup gps)}{5} = \frac{3}{5} = 0{,}6 = 60\%$$

[Agrawal, Imieliński & Swami, 1993] menciona que el Soporte es una norma que se define como la fracción de las transacciones en T que satisfacen la unión del elemento Consecuente y Antecedente de la regla, no se debe confundir con la Confianza debido a que la confianza es una medida que permite evaluar la fuerza de la regla, mientras que el soporte corresponde a la significación estadística.

#### *4.1.3.2 Confianza*

La medida de confianza se refiere a un valor de correspondencia entre los ítems que componen una regla, es decir, la medida denota el porcentaje de transacciones que contienen conjuntamente el término antecedente y el término consecuente en relación al número de transacciones que contienen la parte antecedente.

[Kotsiantis & Kanellopoulos, 2006] mencionan que la **confianza** de una regla de asociación se define como el porcentaje / fracción del número de transacciones que contiene X ∪ Y con el número total de registros que contiene X. Los rangos de los valores que puede tomar la confianza esta dado entre 0 y 1 de acuerdo a lo mencionado por [Azevedo & Jorge, 2007].

$$\mathbf{Conf} = (\mathbf{X} \rightarrow \mathbf{Y}) = \frac{\mathbf{Soporte}(\mathbf{X} \cup \mathbf{Y})}{\mathbf{Soporte}\ (\mathbf{X})}$$

Ecuación 2: Confianza de una Regla de Asociación

**Ejemplo para el Cálculo de la Confianza de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo de la confianza de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación $(cables \rightarrow gps)$ el cálculo de la confianza se obtiene a través de fórmula $\mathrm{Conf} = (\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\mathrm{Soporte}(\mathrm{X} \cup \mathrm{Y})}{\mathrm{Soporte}\ (\mathrm{X})}$ donde $X$ es representado por el artículo “cables” y $Y$ es representado por el articulo “gps”. A continuación, se describen los pasos para el cálculo de la confianza:

- Inicialmente se calcula el valor del soporte de $X$

$$supp(cables) = \frac{4}{5} = 0.8 = 80\%$$

- Posteriormente se calcula el soporte de $X \rightarrow Y$

$$\text{supp}(\text{cables} \rightarrow \text{gps}) = \frac{3}{5} = 0.6 = 60\%$$

- Con los valores anteriores se procede al cálculo de la confianza

$$\text{Conf} = (\text{X} \rightarrow \text{Y}) = \frac{\text{Soporte}(\text{X} \cup \text{Y})}{\text{Soporte (X)}} = \frac{supp\ (cables\ \cup gps)}{supp(cables)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

#### *4.1.3.3 Lift*

Lift o "sustentación" mencionada por [Pinho, 2010], es una medida de interés que tiene como propósito analizar el grado de dependencia entro los elementos que conforman una regla. Dada una regla de $X \rightarrow Y$, la medida lift representa en qué grado Y tiende a ser frecuente cuando X ocurre, lo anterior es representado por la siguiente notación:

$$\boldsymbol{Lift = (X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{soporte(Y)}}$$

Ecuación 3: Lift de una regla de Asociación

[Geng & Hamilton, 2006] mencionan que la medida de **lift** alcanza un valor de 1 en lugar de 0 en el caso de que los atributos sean independientes, adicionalmente Un valor mayor que 1 indica una correlación positiva entre los datos, y un valor de menos de 1 indica una correlación negativa.

**Ejemplo para el Cálculo de Lift de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo de la medida lift de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación $(cables \rightarrow gps)$ el cálculo de lift se obtiene a través de fórmula $Lift = (X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{soporte(Y)}$ donde $X$ es representado por el articulo “cables” y $Y$ es representado por el articulo “gps”. De acuerdo a lo anterior se procede al cálculo de la confianza de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $Y$.

  $supp(Y) = \frac{4}{5} = 0.8 = 80\%$

- Se procede al cálculo de la confianza de $X \rightarrow Y$, para lo cual es necesario saber el valor del soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

$$conf(X \rightarrow Y) = \frac{0.6}{0.8} = 0.75 = 75\%$$

- Una vez hallados los valores anteriores se procede al cálculo de lift tal como se observa a continuación:

$$lift(X \rightarrow Y) = \frac{0.75}{0.8} = 0.93 = 93\%$$

#### *4.1.3.4 Conviction*

**Conviction**, o convicción. Mencionado por [Pinho, 2010] es una medida que tiene como propósito, evaluar el grado en que el antecedente influye en la ocurrencia del consecuente de una regla de asociación. A diferencia del lift, la medida conviction es una medida unidireccional, lo cual quiere decir que el resultado de conviction $(x \rightarrow Y)$ es diferente de $(Y \rightarrow X)$. [Berzal, Blanco, Sanchez & Vila, 2002] mencionan que en la medida de conviction cuando se expresa " ¬ C" significa la ausencia de C. Los resultados de la conviction se encuentran dentro del dominio de (0, ∞), donde 1 significa independencia y los valores generados entre (0, 1) significan dependencia negativa. La medida de conviction es representada mediante la siguiente denotación:

$$\boldsymbol{conv(X \rightarrow Y) = \frac{soporte(X)Soporte(\neg Y)}{soporte(X \cup \neg Y)}}$$

Ecuación 4: Convicción de una regla de Asociación

**Ejemplo para el Cálculo de Conviction de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo de la medida conviction de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada las reglas de asociación $(cables \rightarrow gps)$ el cálculo de conviction se obtiene a través de fórmula $conv(X \rightarrow Y) = \frac{soporte(X)Soporte(\neg Y)}{soporte(X \cup \neg Y)}$ donde $X$ es representado por el artículo "cables" y $Y$ es representado por el artículo "gps". De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente calculamos el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Se procede a calcular el soporte de $\neg Y$ donde el símbolo de "¬" representa ausencia de Y

$$supp(\neg Y) = \frac{1}{5} = 0.2$$

- Posteriormente se calcula el soporte de $soporte(X \cup \neg Y)$

$$supp(X \cup \neg Y) = \frac{0}{5} = 0$$

- Finalmente se calcula el valor de conviction

$$\text{conv} = \frac{(0.8)(0.2)}{0} = 0$$

#### *4.1.3.5 Leverage*

**Leverage**, mencionado por [Azevedo & Jorge, 2007] permite medir cómo se obtiene el recuento de la co-ocurrencia del antecedente y el consecuente. A continuación, se representa la medida leverage a través de la siguiente denotación:

$$\boldsymbol{leverage(X \rightarrow Y) = soporte(X \cup Y) - soporte(X) * soporte(Y)}$$

Ecuación 5: Laverage de una regla de Asociación

**Ejemplo para el Cálculo de Leverage de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo de la medida leverage de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación ($cables \rightarrow gps$) el cálculo de leverage se obtiene a través de fórmula $leverage(X \rightarrow Y) = soporte(X \cup Y) - soporte(X) * soporte(Y)$ donde $X$ es representado por el artículo “cables” y $Y$ es representado por el articulo “gps”. De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Posteriormente se calcula el soporte de $Y$

$$supp(Y) = \frac{4}{5} = 0.8 = 80\%$$

- Luego se calcula el soporte de de $soporte(X \cup Y)$

$$supp(X \cup Y) = \frac{3}{5} = 0.6 = 60\%$$

- Finalmente se calcula la medida leverage

$$leverage(X \rightarrow Y) = 0.6 - (0.8 * 0.8) = 0.6 - 0.64 = 0.04$$

#### *4.1.3.6 Coeficiente de Jaccard*

El **Coeficiente de Jaccard** es una medida que evalúa el grado de coincidencia entre los casos analizados, como resultado los valores que se pueden tomar se encuentran en el rango de [0, 1] y evalúa la distancia entre antecedente y consecuente. Los valores altos indican que X y Y tienden a cubrir los mismos casos [Azevedo & Jorge, 2007]. A continuación, se presenta la medida del coeficiente de Jaccard de acuerdo a la siguiente denotación:

$$\boldsymbol{jacc(X \rightarrow Y) = \frac{soporte(X \rightarrow Y)}{Soporte(X) + soporte(Y) - soporte(X \rightarrow Y)}}$$

Ecuación 6: Coeficiente de Jaccard de una regla de Asociación

**Ejemplo para el Cálculo de Jaccard de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo del coeficiente de Jaccard de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación ($cables \rightarrow gps$) el cálculo del coeficiente de Jaccard se realiza teniendo en cuenta que $X$ es representado por el artículo "cables" y $Y$ es representado por el articulo "gps". De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Posteriormente se calcula el soporte de $Y$

$$supp(Y) = \frac{4}{5} = 0.8 = 80\%$$

- Luego se calcula el soporte de $soporte(X \cup Y)$

$$supp(X \cup Y) = \frac{3}{5} = 0.6 = 60\%$$

- Finalmente se calcula el coeficiente de jaccard

$$jacc(X \rightarrow Y) = \frac{0.6}{0.8 + 0.8 - 0.6} = \frac{0.6}{1} = 0.6$$

#### *4.1.3.7 Cosine*

**Cosine** mencionado por [Azevedo & Jorge, 2007] permite medir la distancia entre antecedente y consecuente cuando éstos son vistos como dos vectores binarios, los valores de resultados de la medida cosine se encuentra entre el rango [0,1], cuando Cosine toma el valor de 1 significa que los vectores coinciden, mientras que si el valor resultante es 0 significa que los vectores no coinciden. El valor de cero sólo sucede cuando el antecedente y el consecuente tienen ninguna superposición. A continuación, se presenta la denotación de la medida Cosine:

$$\cos(X \rightarrow Y) = \frac{soporte(X \rightarrow Y)}{\sqrt{soporte(X) + soporte(Y)}}$$

Ecuación 7: Cosine de una Regla de Asociación

**Ejemplo para el Cálculo de Cosine de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo la medida cosine de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación $(cables \rightarrow gps)$ el cálculo de cosine se realiza teniendo en cuenta que $X$ es representado por el artículo "cables" y $Y$ es representado por el articulo "gps". De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Posteriormente se calcula el soporte de $Y$

$$supp(Y) = \frac{4}{5} = 0.8 = 80\%$$

- Luego se calcula el soporte de $soporte(X \cup Y)$

$$supp(X \cup Y) = \frac{3}{5} = 0.6 = 60\%$$

- Finalmente se procede a calcular la medida cosine

$$\cos(X \rightarrow Y) = \frac{0.6}{\sqrt{0.8+0.8}} = \frac{0.6}{\sqrt{1.6}} = 0.47$$

#### *4.1.3.8 Coeficiente de Pearson*

De acuerdo a [Azevedo & Jorge, 2007] el **Coeficiente de Pearson** es una medida que permite evaluar el grado de asociación entre el antecedente y el consecuente, los rangos de valores que puede tomar están dados entre [-1,1], cuando el Coeficiente de Pearson toma el valor de -1 significa que el antecedente y consecuente cubren casos opuestos, mientras que si toma el valor de 1 significa que tanto el antecedente como el consecuente cubren los mismos casos. A continuación, se presenta la denotación del Coeficiente de Pearson:

$$\boldsymbol{coefp(X \rightarrow Y)} = \frac{\boldsymbol{leverage(X \rightarrow Y)}}{\sqrt{\boldsymbol{soporte(X) * soporte(Y) * \big(1 - soporte(X)\big) * (1 - soporte(Y))}}}$$

Ecuación 8: Coeficiente de Pearson de una Regla de Asociación

**Ejemplo para el Cálculo de Coeficiente de Pearson de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de

asociación y de esta manera poder determinar el cálculo del Coeficiente de Pearson de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación $(cables \rightarrow gps)$ el cálculo del Coeficiente de Pearson se realiza teniendo en cuenta que $X$ es representado por el artículo "cables" y $Y$ es representado por el articulo "gps". De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Posteriormente se calcula el soporte de $Y$

$$supp(Y) = \frac{4}{5} = 0.8 = 80\%$$

- Luego se calcula el soporte de $(X \cup Y)$ el cual permitirá calcular la medida leverage de $X \rightarrow Y$

$$supp(X \cup Y) = \frac{3}{5} = 0.6 = 60\%$$

$$leverage(X \rightarrow Y) = 0.6 - (0.8 * 0.8) = 0.6 - 0.64 = 0.04$$

- Finalmente se procede al cálculo del Coeficiente de Pearson

$$coefp(X \rightarrow Y) = \frac{0.04}{\sqrt{0.8 * 0.8 * (1 - 0.8) * (1 - 0.8)}} = \frac{0.04}{\sqrt{0.0256}} = \frac{0.04}{0.16} = 0.25$$

#### *4.1.3.9 Coverage*

La medida de cobertura o (coverage), expresa la proporción de objetos (o registros) en determinado conjunto de datos que son cubiertos por los ítems del término antecedente de una determinada regla.

**Coverage** es una medida también conocida como soporte del Antecedente, dicha medida se encarga de verificar cuando la regla $X \rightarrow Y$ es aplicable en la base de datos. Se denota de la siguiente manera:

$$\boldsymbol{coverage(X \rightarrow Y) = soporte(X)}$$

Ecuación 9: Coverage de una Regla de Asociación

**Ejemplo para el Cálculo de Coverage de las reglas de asociación:**

En la base de datos de una tienda de ventas de accesorios para vehículos se almacenan los registros de las compras realizadas por los clientes lo cual permite realizar tareas de reglas de asociación y de esta manera poder determinar el cálculo del coverage de las reglas generadas a partir del tratamiento de la información.

| ID | Artículos |
|---|---|
| 1 | llanta, gps, radio |
| 2 | llanta, batería, cables |
| 3 | gps, batería, cables |
| 4 | llanta, gps, batería, cables |
| 5 | llanta, gps, radio, cables |

Dada la reglas de asociación ($cables \rightarrow gps$) el cálculo del coverage realiza teniendo en cuenta que $coverage(X \rightarrow Y) = soporte(X)$ donde $X$ es representado por el artículo "cables" y $Y$ es representado por el articulo "gps". De acuerdo a lo anterior se procede al cálculo de conviction de acuerdo a los siguientes pasos:

- Inicialmente se calcula el soporte de $X$

$$supp(X) = \frac{4}{5} = 0.8 = 80\%$$

- Finalmente él $coverage$ es igual al soporte de $X$ lo que significa que: $coverage = 0.8$

# 5 Estado del arte

## 5.1 Reglas de asociación aplicadas a datos estructurados

Las reglas de asociación juegan un papel muy importante para el desarrollo de distintas temáticas de investigación es por ello que podemos apreciar en la literatura existente múltiples aplicaciones de las mismas en diferentes escenarios tales como:

- Validación de grado de asociación de grandes volúmenes de datos.
- Análisis del comportamiento de los clientes en compras de mercado.
- Análisis del comportamiento de los clientes en préstamos de libros.
- Diseño de Herramientas para prueba de resistencia de choques en automóviles.
- Análisis de calidad y cantidad de reglas de asociación.
- Utilización de Computación en Paralelo para el análisis de grandes volúmenes de datos.
- Creación de nuevos algoritmos optimizados que permitan analizar de forma rápida y eficiente grandes volúmenes de datos.
- Creación de nuevas tecnologías para la detección de intrusiones basadas en reglas de asociación.
- Métodos para la extracción de conocimiento en el área de la biología.

Las reglas de asociación han sido implementadas desde la década de los 90, lo cual ha producido grandes avances a lo largo de los años producto de los grandes esfuerzos de los investigadores. A continuación, se presenta una breve reseña de la utilización de las reglas de asociación:

[Agrawal, Imieliński & Swami, 1993] Utilizan las Reglas de Asociación para el análisis de las cestas de mercado a partir de dicho estudio se logró establecer criterios para tomar decisiones en base al comportamiento de compra de los productos por parte de los clientes lo que permitió establecer las reglas del orden o posición de los productos en los diferentes lugares del supermercado obteniendo resultados positivos. La investigación fue realizada utilizando las transacciones de una empresa minorista la cual contiene 63 departamentos en los cuales hay un total de 46873 registros de transacciones de las compras realizadas por los clientes.

[Agrawal & Srikant, 1994] Consideran que los algoritmos existentes de reglas de asociación no son lo suficientemente rápidos para el análisis de grandes cantidades de datos es por ello que se somete a experimentación utilizando datos sintéticos, producto de la experimentación realizada se propone un algoritmo llamado AprioriHybrid el cual se basa en los algoritmos anteriormente mencionados.

[Agrawal & Shafer, 1996] Manifiestan que uno de los problemas de la minería de reglas de asociación está relacionado con no compartir nada a través de multiprocesos, en consecuencia, la investigación realizada presenta un algoritmo que explora el espectro del balance entre computación, comunicación y uso de memoria, sin embargo, solo requiere una sobrecarga mínima en comparación con los mejores algoritmos de serie. Producto de la investigación realizada se implementó el algoritmo Apriori utilizando múltiples procesos al tiempo lo cual se conoce como computación en paralelo.

[Han, Pei, Yin & Mao, 2004] Mencionan que la minería patrones frecuentes en las bases de datos de transacciones, bases de datos de series de tiempo, y muchos otros tipos de bases de datos se ha estudiado popularmente en la investigación de minería de datos. La mayoría de los estudios previos asumen un enfoque de generación de conjunto de candidatos semejante al

realizado por el algoritmo Apriori. Sin embargo, la generación de conjunto candidato es todavía costoso, especialmente cuando existe un gran número de patrones y / o patrones de gran tamaño. Lo anterior género como resultado la generación de un método de patrón árbol frecuente (FP-árbol), que es una estructura de árbol extendido para el almacenamiento de información crucial comprimido sobre los patrones frecuentes.

[Xu, Li & Shaw, 2011] Utilizan como objeto de análisis el conjunto de datos Mushroom el cual tiene como propósito identificar el tipo de hongos son comestibles y cuales son venenosos , para el análisis del conjunto de datos mencionado anteriormente se utilizan las reglas de asociación , y producto de dicho estudio se pretende tener la menor cantidad de reglas construidas, debido a que se pretende superar uno de los errores más comunes al utilizar las reglas de asociación que hace referencia a la gran cantidad de reglas generadas. Adicionalmente como resultado de la investigación planteada se propone utilizar el factor de seguridad como el criterio para medir la fuerza de las reglas de asociación descubiertas.

[Xiang, 2012] Implementa las reglas de asociación para análisis de la variedad de los datos de accidentes automóviles a través de la implementación del algoritmo Apriori, el resultado puede mostrar las relaciones que afectan las pruebas de choque entre los coches relacionados, y puede proporcionar una referencia para el diseño de automóviles.

[Hanguang & Yu, 2012] Manifiestan que la detección de intrusiones es una de las partes importantes del sistema de seguridad, es por ello que se ha convertido en un área de interés de investigación. Actualmente existe una variedad de nuevos métodos de ataque, lo cual ha generado una demanda de sistemas de detección de ataques los cuales se basen en algoritmos eficientes. Mediante el análisis de la tecnología del sistema de detección de intrusiones y la minería de datos está en capacidad de generar tecnologías eficientes para la detección de ataques,

en consecuencia, se utiliza el algoritmo Apriori que es el clásico de las reglas de asociación en el sistema de detección de intrusiones basado en la Web y se aplica la base de reglas generadas por el algoritmo Apriori para identificar una variedad de ataques. Para la experimentación planteada se utilizó el Data-Ser NSLKDD.

[Tsuji et al., 2014] Utilizan las reglas de asociación para el análisis de préstamos de libros en los cuales se quiere observar el comportamiento de una persona que presta un libro A,B y C lo cual puede generar una cantidad de reglas interesantes. Se presenta el concepto de Antecedente y consecuente los cuales son denominados como premisa y conclusión, adicionalmente se menciona las medidas de interés conocidas como soporte y confianza.

[Nagata, Washio, Kawahara & Unami, 2014] Manifiestan que la reciente llegada de las nuevas tecnologías en la biología como los Microarray de ADN y secuenciador de última generación ha dado a los investigadores un gran volumen de datos que representan las respuestas biológicas de todo el genoma, sin embargo, no es fácil de obtener conocimiento a partir de dichas fuentes de datos. En la experimentación planteada, se aplicó clasificación basada en el algoritmo de asociación (CBA), una de las técnicas de minería de reglas de asociación de clase, a la base de datos TG-Gates, donde se almacenan los datos tanto toxico genómicos y toxicológicos de más de 150 compuestos en ratas y humanos. Se procedió a realizar una comparación entre el algoritmo generado CBA y análisis discriminante lineal (LDA) y demostró que CBA es superior a la LDA de acuerdo a los siguientes resultados (precisión: 83% para la CBA vs. 75% para LDA, la sensibilidad: 82% para CBA vs. 72% para LDA, especificidad: 85\% para la CBA vs. 75% para LDA).

[Soysal, 2015] Menciona que el problema de la minería de datos estructurados para encontrar patrones potencialmente útiles de la minería de reglas de asociación. Se propone un método

heurístico para extraer patrones mayormente asociados (MASPS). Este enfoque utiliza una restricción máxima y de asociación para generar patrones sin tener que buscar todo el entramado de combinaciones de elementos. El enfoque propuesto requiere menos recursos computacionales en términos de requisitos de tiempo y de la memoria, mientras que la generación de una larga secuencia de patrones que tienen la mayor concurrencia.

[Sahoo & Goswami, 2015] Manifiestan que tradicional minería de reglas de asociación basada en el Soporte y confianza proporcionan la medida objetiva de las normas que son de interés para los usuarios. Sin embargo, no refleja la medida semántica entre los elementos. La medida semántica de un conjunto de elementos se caracteriza con valores de utilidad que normalmente se asocian con los objetos de transacción, donde un usuario se interesara a un conjunto de elementos solo si se satisface una restricción utilidad dada. En la investigación planteada, primero se define el problema de encontrar reglas de asociación utilizando el framework utilityconfidence, que es una generalización de la medida de la cantidad en sí mismo. El uso de este concepto semántico de las reglas, genera una representación comprimida de reglas de asociación que tienen antecedente como valor mínimo y la consecuente como valor máximo. Producto de la investigación realizada se proponen los algoritmos para generar las reglas de asociación de utilidad basada no redundantes y métodos para la reconstrucción de todas las reglas de asociación. Además, se describen los algoritmos que generan conjuntos de elementos de servicios públicos (HUI). Estos algoritmos propuestos se implementan utilizando conjuntos de datos tanto sintéticos y reales. Los resultados demuestran mejor eficiencia y eficacia del algoritmo HUCI-Miner propuesto en comparación con otros algoritmos existentes bien conocidas. Además, los resultados experimentales muestran una mejor calidad en la representación comprimida de todo el conjunto de reglas en el marco considerado.

[Narvekar & Syed, 2015] mencionan que la minería de reglas de asociación ha contribuido a muchos avances en el área de descubrimiento de conocimiento. Sin embargo, la calidad de las reglas de asociación descubiertos es una gran preocupación y ha atraído a más y más atención recientemente. Uno de los problemas con la calidad de la asociación descubierto es el enorme tamaño del conjunto de reglas extraído. La solución planteada hace referencia a generar un método que permita disminuir la cantidad de reglas de poca relevancia y las que se encuentren redundantes.

## 5.2 Reglas de asociación aplicadas a texto

Las reglas de asociación son aplicadas para el estudio de diferentes situaciones lo que se aprecia en los relatos de distintos autores quienes analizan bases de datos de texto a través de la aplicación de las Reglas de asociación.

Los autores [Cherfi, Napoli & Toussaint, 2006] proponen una metodología para la minería de texto basándose en el circuito de descubrimiento de conocimiento clásico, con una serie de adaptaciones. Inicialmente los textos se indexan y son preparan para ser procesados. Las reglas de asociación se extraen y se interpretan, con respecto a un conjunto de medidas de calidad y conocimiento del dominio, bajo el control de un analista.

Los autores [Herawan & Deris, 2011] presentan un enfoque o alternativa para la minería de reglas de asociación “Regulares” y reglas de asociación “máxima” a partir de Data-Set transaccionales utilizando teoría de conjuntos, el enfoque propuesto de forma inicial se encarga de transformar el Data-Set Transaccional en un sistema de información con valores booleanos, adicionalmente se utiliza el concepto de parámetro de co-ocurrencia de una transacción el cual

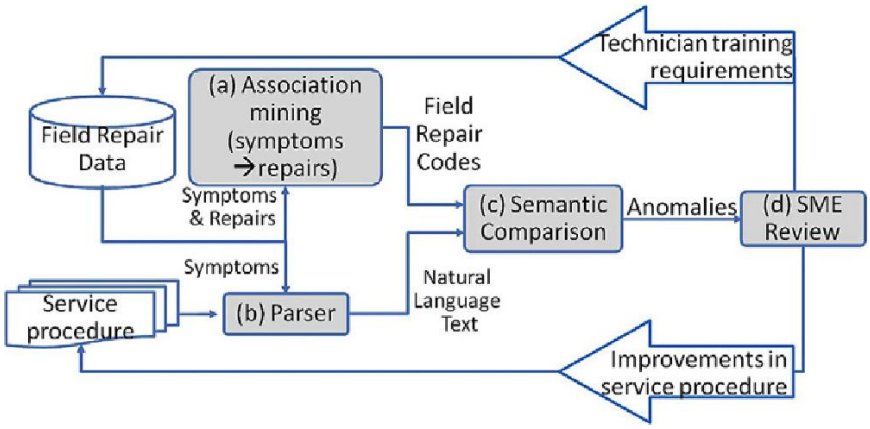
permite definir la noción de regla de asociación “máxima” y “Regulares” entre dos conjuntos de parámetros como resultado se puede observar que el software de reglas de asociación proviene de reglas idénticas que son comparadas.

[Chen, Tseng & Liang, 2010] presentan un estudio tomando como base la Integración de Wordnet y Minería de Reglas de Asociación Difusas Para el Documento de Agrupación Multi Etiquetas, debido a que en la actualidad se presenta un rápido crecimiento de los documentos de texto, lo cual ha causado que la agrupación de documento se convierta en una de las principales técnicas para organizar gran cantidad de documentos en un pequeño número de grupos significativos. Sin embargo, aún existen varios retos para la agrupación de documentos, como la alta Dimensionalidad, la Escalabilidad, la Precisión, Clusters solapados, y la extracción de la semántica de los textos. Con el fin de mejorar la calidad de los resultados de los clustering de documentos, proponen un enfoque eficaz basado en un efectivo clustering de documentos multi etiqueta basado en Reglas de Asociaciones Difusas, que integra minería las reglas de asociación difusas con la ontología existente WordNet para aliviar estos problemas. En el estudio realizado, los términos clave se extraen del conjunto de documentos, y la representación inicial de todos los documentos se enriquece aún más mediante el uso de hiperónimos de WordNet, ya que permite explotar las relaciones semánticas entre los términos. Luego, se emplea un algoritmo de minería de reglas de asociación difusa para textos, para descubrir un conjunto de elementos frecuentes difusos relacionados, que contienen términos clave a ser considerado como las etiquetas de los clusters candidatos. Finalmente, cada documento se distribuye en más de un cluster haciendo referencia a los clusters candidatos, y entonces los clusters objetivos altamente similares se fusionan. Los resultados experimentales demostraron que el enfoque propuesto supera a los métodos de clustering de documentos influyentes con una alta precisión.

[Brin, Motwani & Silverstein, 1997] mencionan que uno de los problemas más estudiados en el medio en la minería de datos es la minería por reglas de asociación. Las Reglas de Asociación cuya importancia se mide a través del soporte y la confianza, tienen la intención de identificar las reglas del tipo "Un cliente que compra el Ítem A, a menudo también compra Ítem B" Motivado por el objetivo de generalizar más allá del mercado y por las reglas de asociación utilizadas con ellos, el concepto de reglas de minería que identifican correlaciones (asociaciones generalizadas), y se considera tanto la ausencia y la presencia de artículos como base para la generación de reglas. En la investigación realizada se pretende medir la importancia de las asociaciones a través de la prueba de chi-cuadrado para correlacionarla con la estadística clásica. Esto conduce a una medida de cierre en el alza en el conjunto de elementos, lo que permite reducir el problema de minería a la búsqueda de una frontera entre conjuntos de elementos correlacionados y no correlacionadas en la red. Posteriormente se desarrollan estrategias y se diseñan un algoritmo eficiente para el problema resultante. Demostramos su eficacia probándola en los datos del censo y en la búsqueda de términos dependientes en el cuerpo del documento de texto, así como en los datos sintéticos.

[Khare & Chougule, 2012] presentan un sistema de soporte a la toma de decisiones Dominio basado en Minería de texto consciente y minería de Asociación (Datam), el cual fue desarrollado para mejorar el servicio de post-venta y reparaciones de automóviles, en consecuencia, se propone un nuevo enfoque que compare datos textuales y no textuales para la detección de anomalías. Esto combina asociación y ontología basada en la minería de texto. Minería de Asociación se ha empleado para identificar las reparaciones realizadas en el campo por un determinado síntoma, mientras que, la minería de texto es utilizada para inferir las reparaciones de las instrucciones de texto mencionadas en los documentos de servicio para el mismo síntoma.

Estos a su vez son comparados y contrastados para identificar casos anómalos. El enfoque desarrollado se ha aplicado a los datos de campo automotriz. Usando el top 20 de síntomas más frecuentes, observados en un automóvil sedan de tamaño medio construido y vendido en América del Norte, demuestra que los datos pueden identificar todos los síntomas anómalos - combinaciones de códigos de reparación (con una tasa de falsos positivos del 0,04). Este conocimiento, en forma de anomalías, puede posteriormente ser utilizado para mejorar el servicio/procedimiento de solución de problemas e identificar las necesidades de formación o entrenamiento de técnicos.




Figuras 2. *Framework para la detección de anomalías entre las reparaciones previstas en los manuales de servicio*. *Fuente: [Khare, 2012].*

[Holt & Chung, 2001] proponen un algoritmo nuevo llamado Multipass-Apriori y Multipass-DHP para minería de reglas de asociación entre palabras que se encuentran en bases de datos de texto. Las características de las bases de datos de texto son bastante diferentes de aquellas bases de datos de transacciones de ventas al por menor y detal, existen algoritmos de minería que no pueden manejar de forma eficiente las bases de datos de texto por el numero largo o cantidad de ítems. En este artículo dos algoritmos de minería conocidos Apriori y Direct Hashing and

Pruning son evaluados bajo el mismo contexto que en este caso corresponde a la minería de texto.

[Song, Song, Hu & Allen, 2007] presentan la Integración de Reglas de Asociación y Ontologías Para La Expansión de Consultas Semánticas como un nuevo método de expansión de consultas semánticas que combina las Reglas de Asociación con la ontología y las técnicas de procesamiento del lenguaje Natural. La técnica desarrollada es novedosa debido a que utiliza la semántica explicita, así como otras propiedades lingüísticas del cuerpo del texto no estructurado. Hace uso de las propiedades contextuales de importantes términos descubiertos por las Reglas de Asociación, y las entradas de la ontología son agregadas a la consulta por desambiguación del significado de las palabras. Utilizando TREC consultas Ad Hoc se logró alcanzar resultados entre 13,41 y 32,39% de mejora para precisión y de 8,39% a 14,22% para la Media Armónica.

[Li & Wu, 2014] proponen un Estudio Para Interpretación de Reglas de Asociación en las Estructuras Multi-Nivel, lo cual permite enfrentar la gran cantidad de datos que resultan de la minería de reglas de asociación debido a que es considerado un gran reto. El inconveniente esencial es como proporcionar métodos eficientes para resumir y representar el conocimiento descubierto de las bases de datos. En consecuencia, el estudio realizado propone un enfoque llamado minería granulada multi-nivel la cual permite mejorar de una forma considerable el rendimiento de la minería de reglas de asociación, en lugar de usar patrones se utilizan gránulos los cuales permiten representar el conocimiento que esta implícitamente contenido en las bases de datos relacionales. Este enfoque también utiliza las estructuras multi-nivel y mapeos de asociación para interpretar las reglas de asociación en forma granulada. En consecuencia, las reglas de asociación pueden ser evaluadas de forma rápida y las reglas de asociación sin sentido

pueden justificarse de acuerdo con estos mapeos de asociación. Los resultados experimentales indican que el enfoque propuesto es alentador.

[Amir, Aumann, Feldman & Fresko, 2005] proponen en su investigación la utilización de Reglas de Asociación Máxima Como Herramienta Para Minería de Asociaciones en Texto, en el cual se describe una novedosa herramienta para la minería de reglas de asociación, el cual tiene un valor especial en la minería de texto. La nueva herramienta, llamada asociaciones máximas, se orienta hacia el descubrimiento de las asociaciones que se pierden con frecuencia cuando usan las reglas de asociación regulares. Las reglas de Asociaciones Máximas permiten el descubrimiento de las asociaciones pertenecientes a los elementos que más a menudo no aparecen solos, sino con los temas estrechamente relacionados, y por lo tanto las asociaciones relevantes solo a estos elementos tienden a obtener una confianza baja. Proporcionamos una descripción formal de las reglas de asociación máximas y algoritmos eficientes para descubrir todas esas asociaciones. Producto de la experimentación realizada se presentan los resultados de la aplicación de reglas de asociación máximas a dos corpus de texto.

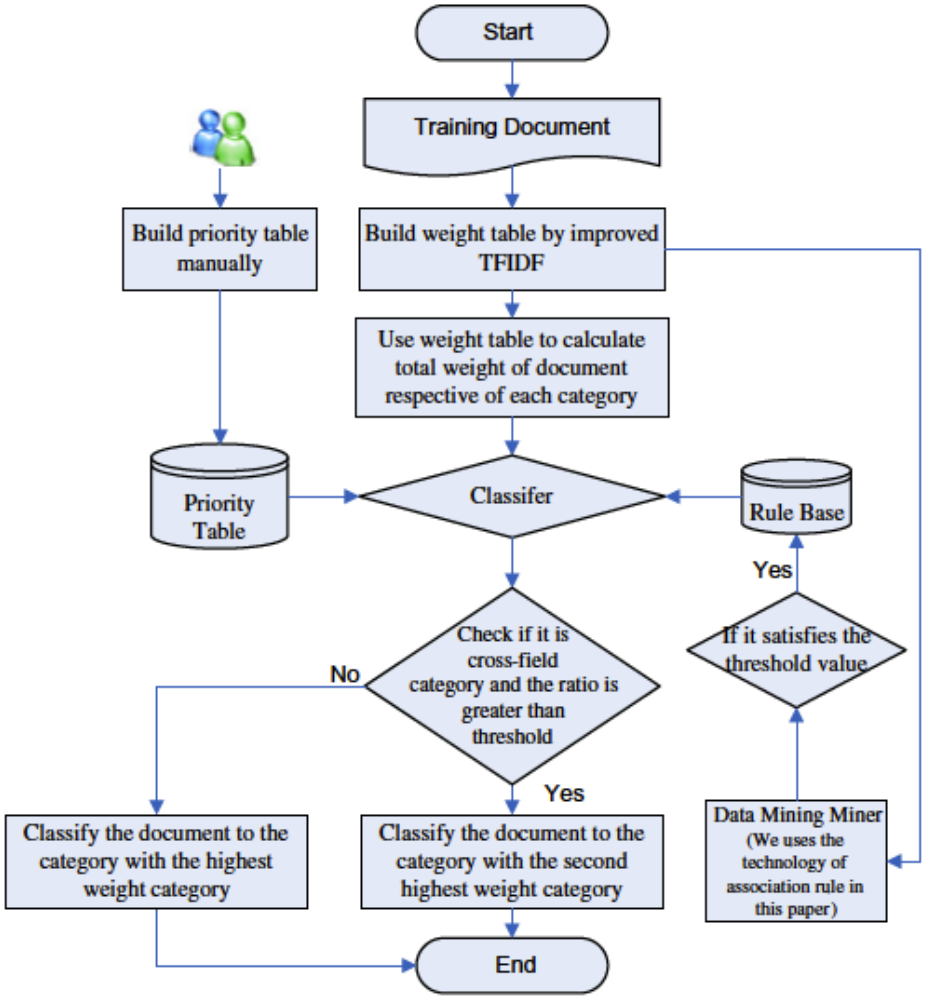
[Huang, Liao, Yang, Chang, & Luo, 2010] proponen la Realización de un Agente de Diseminación de Noticias Basado en Reglas de Asociación Ponderadas y Técnicas de Minería de Texto, en el cual un agente titular de noticias financieras propone asesorar a los inversores en la decisión de comprar y vender acciones en el mercado de Taiwán luego de recibir por parte del agente el titular de la noticia de la primera plana en tiempo real diseminada. Las reglas de asociación ponderadas y las técnicas de minería de texto se utilizan para determinar el grado de importancia de cada titular de noticia en la fluctuación del índice de precio de la bolsa de Taiwán el siguiente día de negociación. Los resultados experimentales revelan que esta propuesta de trabajo de hecho logra un rendimiento significativo y demuestra su viabilidad en las aplicaciones

de difusión de la información en tiempo real, como titulares de noticias financieras a través de Internet.

[Tang, Yan & Yuan, 2013] proponen un método basado en diccionario semántico para clasificación de texto corto, manifiestan que la precisión de la clasificación de textos cortos tradicional generalmente depende de la selección de características estadística. Debido al hecho que los textos cortos tienen defectos inherentes tales como la longitud corta, señal débil y menos funciones. Es difícil evitar las palabras irrelevantes cuando se hace la extensión característica que tendrá una gran influencia en la precisión de la clasificación. El método construye un conjunto de diccionario de dominio mediante el análisis de las características específicas en cierto campo. Como el peso de cada palabra en el diccionario está diseñado de acuerdo a la correlación entre la palabra y la categoría, la precisión de clasificación ha mejorado en cierta medida. Entonces, con el fin de mejorar la cobertura del diccionario de vocabulario, se utilizan reglas de asociación para extender automáticamente diccionario semántico. Finalmente, para el proceso de validación se llevó a cabo un experimento basado en datos de micro-blog que demuestra que el método tiene un buen resultado.

[Chiang, Keh, Huang & Chyr, 2008] mencionan que el proceso de categorización de texto implica algunos conocimientos del contenido de los documentos de forma previa, adicionalmente se requiere algunos conocimientos previos de las categorías. Para el contenido de los documentos se utilizan medidas de filtros para la selección de características en el sistema de categorización del texto chino. En el artículo de investigación se modifica la fórmula de TFIDF para fortalecer los pesos de las palabras claves importantes y reducir los pesos de las palabras claves que no son importantes. Para el conocimiento de las categorías, se utilizan las reglas de asociación para mejorar la clasificación de texto y usa prioridad de categoría para representar la relación entre 2

categorías diferentes. Consecuentemente los resultados de la investigación realizada muestran que el método no solo puede reducir el ruido, sino que incrementa el radio de precisión y cobertura de la categorización del texto.




**Figuras 3.** ***Proceso del Sistema de categorización de texto***
*Fuente: [Chiang, 2008]*

[Lopes, Pinho, Paulovich & Minghim, 2007] menciona en su investigación que, en muchas situaciones, individuales o grupos individuales se enfrentan con la necesidad de examinar conjuntos de documentos o archivos para entender su estructura y para localizar la información

relevante, en dicho contexto, la investigación realizada presenta un Framework para minería de texto visual, soportada en la exploración de ambas estructuras generales y temas relevantes dentro de una colección de documentos de texto. El enfoque presentado dentro de la investigación inicia con la construcción de un Data-Set de texto para la visualización, adicionalmente se presenta una técnica que genera reglas de filtro de asociación para detectar y mostrar temas para un grupo de documentos.

[Wong, Whitney & Thomas, 1999] investigan sobre la visualización de reglas de asociación para minería de texto, donde uno de los principales objetivos es el descubrimiento de reglas de asociación importantes dentro de un corpus de tal manera que la presencia de un conjunto de temas en un artículo implica la presencia de otro tema. Una regla de asociación en minería de datos es una implicación donde X es un conjunto de ítems antecedentes, Y es el ítem consecuente. Durante años, los investigadores han desarrollado muchas herramientas para visualizar las reglas de asociación. Sin embargo, algunas de estas herramientas pueden manejar más de decenas de normas, y ninguno de ellos puede gestionar eficazmente reglas con varios antecedentes, lo cual indica que es extremadamente difícil de visualizar y entender la información de asociaciones en un gran conjunto de datos. La investigación realizada presenta una técnica de visualización novedosa para abordar muchos de los problemas anteriormente mencionados, por lo cual aplicamos la tecnología para un estudio de la minería de texto con un amplio corpus. Los resultados indican que el diseño realizado puede manejar fácilmente cientos de múltiples reglas de asociación antecedente en una pantalla tridimensional con mínima interacción humana, bajo porcentaje de oclusión, y ningún intercambio pantalla. La investigación realizada utiliza un corpus de reportaje obtenido de fuentes abiertas, el cual tiene un tamaño 9 MB y es almacenado como un archivo ASCII con más de 3.000 artículos recopilados durante 20

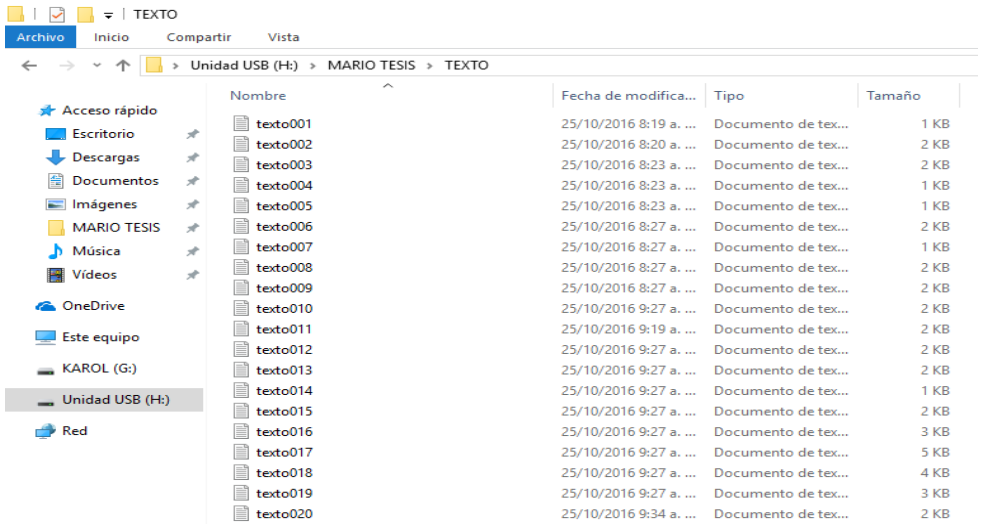
a 26 abril de 1995. Este corpus tiene un tema fuerte asociado con el atentado contra el Edificio Federal en Oklahoma.

# 6 Aplicación de reglas de asociación para el análisis de afinidad entre objetos de tipo texto

En este capítulo, socializaremos el proceso realizado para la implementación de las reglas de asociación sobre objetos de tipo texto. A continuación, se describen los pasos que fueron realizados en dicha experimentación.

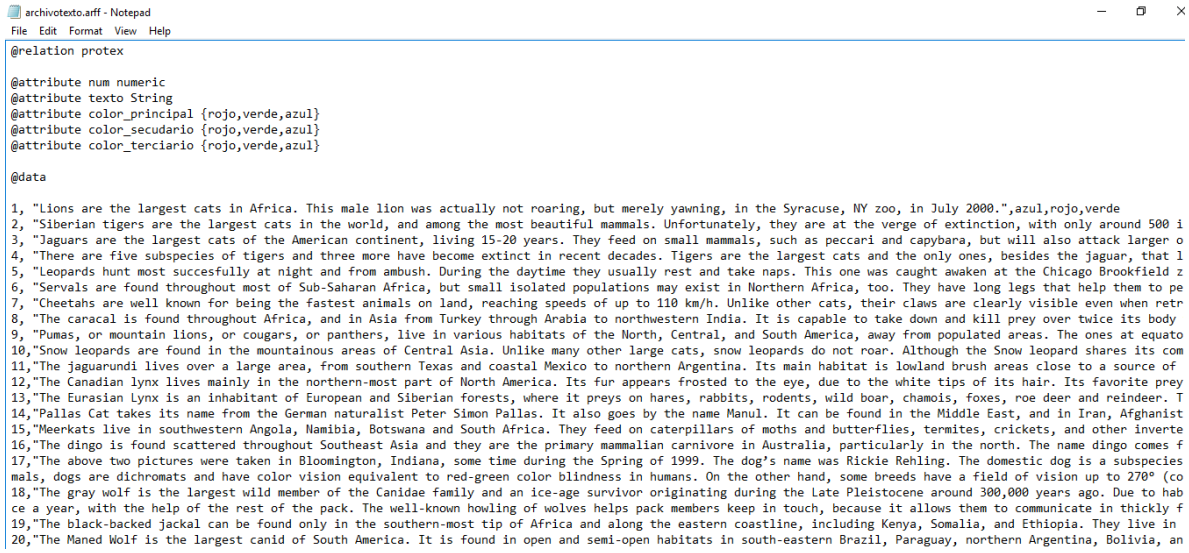
En primera instancia, es importante tener seleccionado y claramente identificado el conjunto de datos que se va a utilizar. Para esta selección, fue necesario realizar una búsqueda a través de las fuentes de información o bases de datos especializadas en las cuales se encuentran los artículos de investigación. Se analizaron cada uno de los Data sets y teniendo en cuenta lo anterior se tomó como conjunto de datos la información descrita en la publicación [Malik & Kender, 2006] el cual corresponde a los datos almacenados en una página web que contiene elementos de tipo imagen y de tipo texto. En dicho repositorio, se pueden apreciar una serie de imágenes de animales los cuales tienen la descripción en forma de texto de la forma de hábitat de cada animal, así como su forma de crecimiento, alimentación entre otros. Se hizo un análisis solamente de la parte de contenido de texto de dicho Data Set.

Para la construcción de manera eficiente y lógica del conjunto de datos fue necesario almacenar los documentos de texto que se encuentra en la página web del repositorio, en total fueron descargados 100 documentos de texto, a continuación, se pueden observar algunos de los datos almacenados.




Figuras 4. *Pantalla de Repositorio de documentos de texto*
*Fuente: Elaboración propia*

Posteriormente almacenados los documentos, se realiza el conjunto de datos, lo cual es estrictamente necesario tener en cuenta los atributos a extraer de cada elemento. Posterior a la extracción de características de los elementos que componen el objeto de texto, se procede a la generación del archivo arff el cual debe contener los atributos o características del texto para encontrar posibles relaciones entre ellos.




**Figuras 5. *.: Pantalla de Fragmento de archivo arff con atributos de texto***
*Fuente: elaboración propia*

Finalmente, el documento de texto al aplicar bolsa de palabras y al unir dicha información con los atributos extraídos de las imágenes se genera un conjunto de datos que contiene 100 registros y 2556 atributos aproximadamente, los cuales deben ser reducidos teniendo en cuenta que en la aplicación de bolsas de palabras se generar algunos errores o palabras que no infieren o son irrelevantes. Esto disminuye de forma sustancial la cantidad de atributos a tener en cuenta.

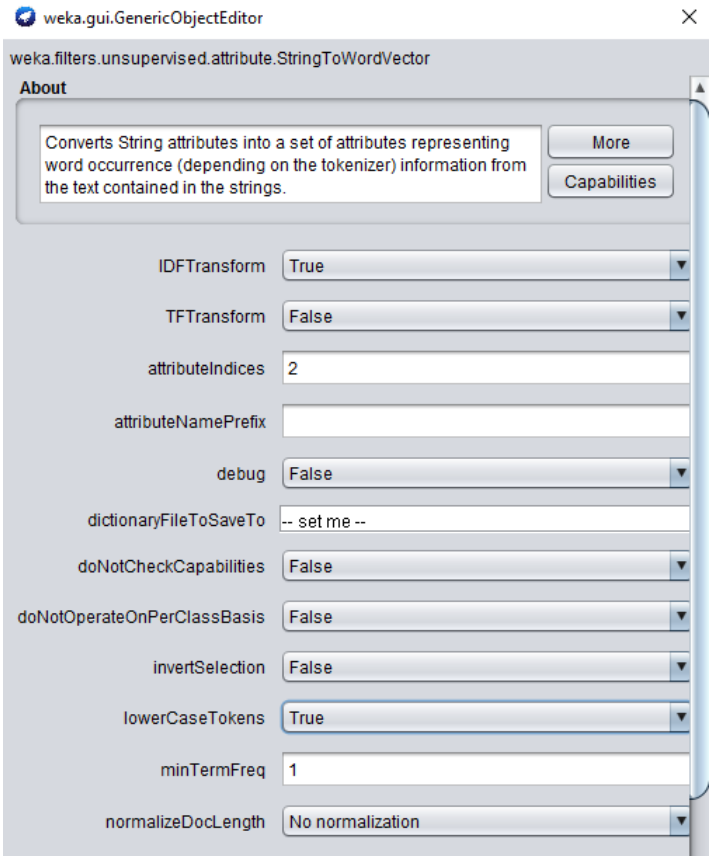
En este caso se hizo una reducción de aproximadamente el 96% de palabras irrelevantes que no tienen inferencia o son simplemente conectores lógicos, símbolos, caracteres especiales, números (arábigos o romanos), abreviaturas, entre otros. De lo anterior quedaron 95 atributos significativos.

## 6.1 Configuración de string to word vector

Para la realización del proceso descrito a continuación, se utilizó la herramienta Weka. Esta es una herramienta de tipo software para el aprendizaje automático y minería de datos. Contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, también tiene herramientas para la visualización de estos datos, además provee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición. De esta herramienta obtendremos como resultado el conjunto de datos al cual se le implementara el algoritmo de reglas de asociación.

El objetivo principal de esta etapa es generar la matriz de la bolsa de palabras (Stops Word), en la que se pueda representar mediante un dato continuo, la cantidad de veces que aparece una palabra o un atributo en los documentos que pertenece al conjunto de datos. Para lograr obtener los resultados esperados, se debieron aplicar algunos métodos de stemmer, discretizacion, se aplicó la metodología de bolsa de palabras y se realizó un proceso de verificación de los atributos

generados a partir de los documentos de texto esto con el propósito de realizar el análisis de la información con datos correctos y no con datos que contengan ruido. A continuación, se presenta la configuración utilizada para generar la matriz de la bolsa de palabras.

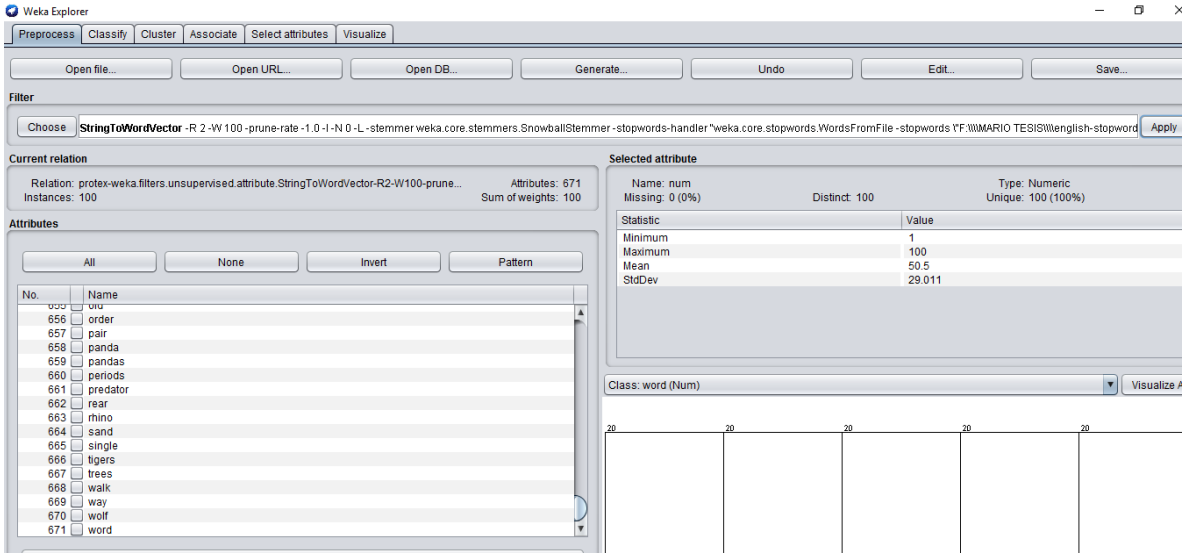



**Figuras 6.** ***Pantalla de Configuración de método string to word vector (1)***
*Fuente: Elaboración propia*


**Figuras 7.** ***Pantalla de Configuración de método string to word vector (2)***
*Fuente: Elaboración propia*

## 6.2 Cantidad de atributos generados

Posterior a la aplicación del filtro String to Word Vector se obtuvo una cantidad importante de atributos los cuales fue necesario depurar para la posterior aplicación del algoritmo de reglas de asociación llamado Apriori. A continuación, se presenta una ilustración en la que se puede apreciar la cantidad de atributos generados.

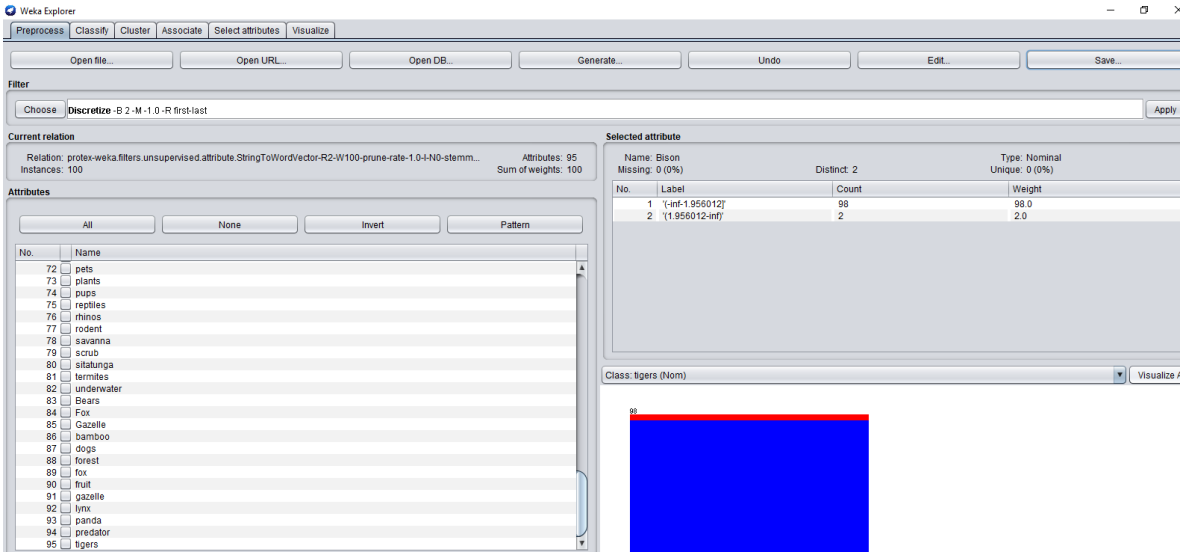



**Figuras 8.** ***Pantalla de Cantidad de atributos generados***
*Fuente: Elaboración propia*

## 6.3 Cantidad de atributos posterior a eliminación de datos irrelevantes

Una vez generados los atributos es necesario realizar un proceso de limpieza de datos con la finalidad de reducir la mayor cantidad de atributos que generar ruido para la implementación de las reglas de asociación, en dicho proceso se procede a comprobar de forma manual los datos contenidos en la bolsa de palabras, de esta manera eliminando los datos redundantes, datos que no tengan un valor semántico, abreviaturas, palabras en plural. Se obtuvo un total de 95 atributos

que utilizaremos para la aplicación del algoritmo A priori, esperando resultados importantes. A continuación, se presenta una ilustración de la cantidad de atributos posterior al proceso de eliminación de atributos irrelevantes.



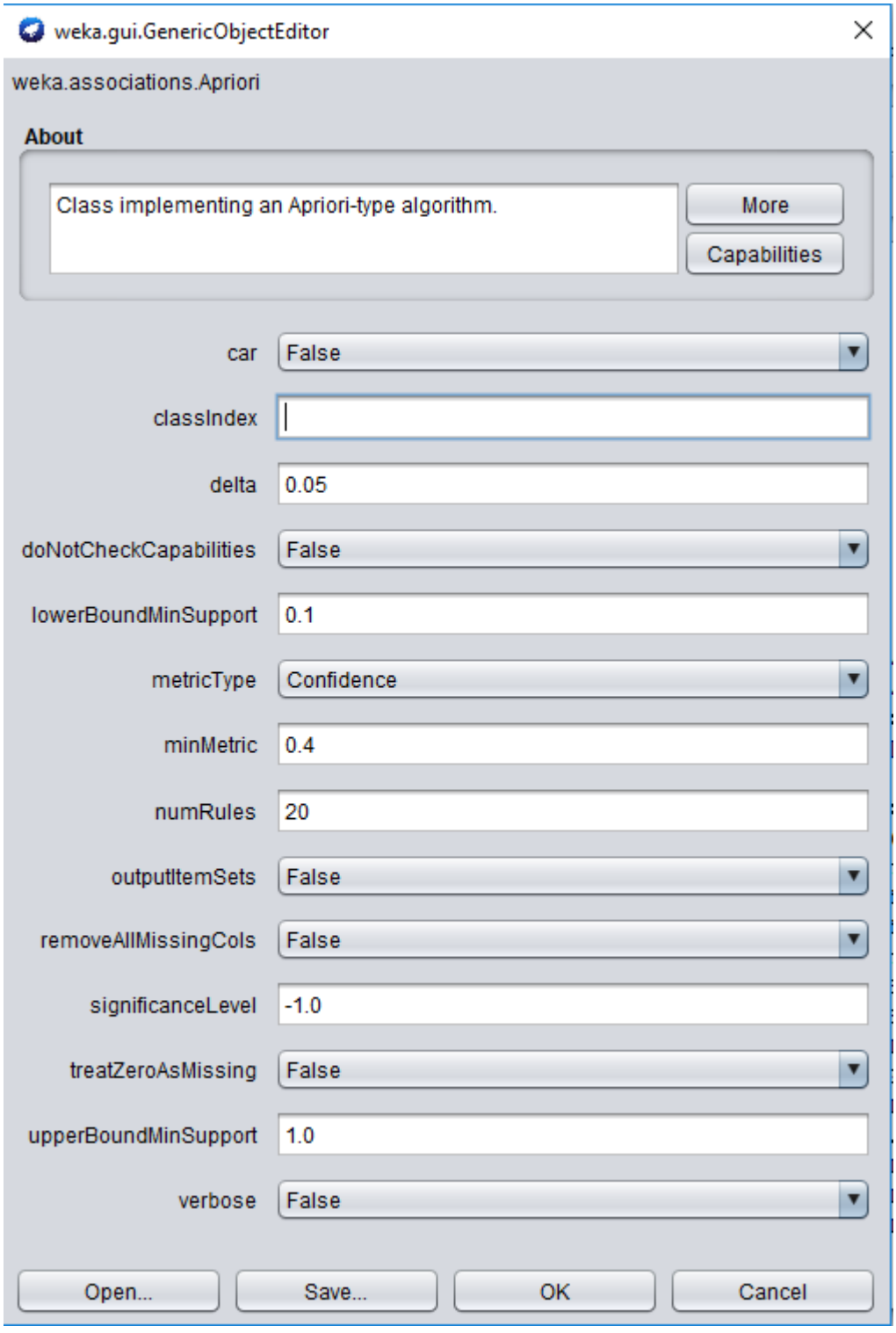

**Figuras 9.** ***Pantalla de Cantidad de atributos posterior a la limpieza de datos***
*Fuente: Elaboración propia*

## 6.4 Configuración de método a priori

Una vez los datos se encuentren preparados se procede a la aplicación del método de a priori. Apriori es un algoritmo diseñado para descubrir grandes conjuntos de elementos a través de realizar varias pasadas sobre los datos [Yin & Li, 2006]. El primer paso hace referencia al conteo de las ocurrencias del elemento con el propósito de determinar el tamaño de 1-itemset (conjunto de elementos). Posteriormente se procede a llamar a la interacción k la cual está compuesta de dos fases. Primero, el tamaño del conjunto de elementos Lk-1 es encontrado en la interacción (K-1) este paso es usado para generar el candidato del conjunto de elementos Ck, el paso anterior se realiza utilizando la función de generación de candidatos de a priori (A priori-gen) la cual recibe como argumentos Lk-1. Luego, la base de datos es escaneada y se procede a realizar el conteo

del soporte “support” de los candidatos en Ck. Para lograr un conteo eficiente es necesario determinar los candidatos de Ck que son contenidos en la transacción dada t.

De acuerdo a lo mencionado por [Narvekar & Syed, 2015], Apriori hace parte de los métodos de Reglas de Asociación, el cual genera una cantidad de reglas consideradas como buenas al momento de analizar pequeñas bases de datos, por el contrario se considera como desventaja del algoritmo que a medida que aumenta el tamaño de la base de datos el rendimiento de dicho algoritmo disminuye, esto se debe a que debe escanear toda la base de datos cada vez que se escanea una transacción, es por ello que se considera que el principal problema del algoritmo Apriori es el tratamiento de los datos en grandes volúmenes de datos, problema que también menciona [Huang, Liao, Yang, Chang & Luo, 2010] quien manifiesta que el rendimiento del algoritmo de Apriori será muy bajo e ineficiente cuando la capacidad de memoria es limitada con gran número de transacciones, otro de los problemas del algoritmo Apriori es que adolece de algunas deficiencias a pesar de ser claro y sencillo, la principal limitación es costosa pérdida de tiempo para mantener gran número de conjuntos de candidatos con mucho conjuntos de elementos frecuentes. A continuación, se presenta la configuración utilizada para el algoritmo a priori.




*Figuras 10. Pantalla de Configuración de Método A priori*
*Fuente: Elaboración propia*

## 6.5 Dificultades durante proceso de ejecución

A continuación, se describen las dificultades obtenidas durante el proceso de ejecución del algoritmo Apriori:

• Para la configuración del algoritmo Apriori se debe elegir en cuantos grupos se van a dividir los atributos los cual es un factor de incertidumbre teniendo en cuenta que no es posible tener de forma certera cual es el número de divisiones correctas para generar los intervalos de valores de los atributos analizados.

• Otro de los inconvenientes presentados es la representación de la información del objeto de tipo texto, ya que la preparación de datos en los documentos de tipo texto requiere de mucho cuidado. Se debe, necesariamente eliminar información que no es útil para el análisis, como, por ejemplo, caracteres especiales, abreviaturas, conectores, palabras irrelevantes, entre otros.

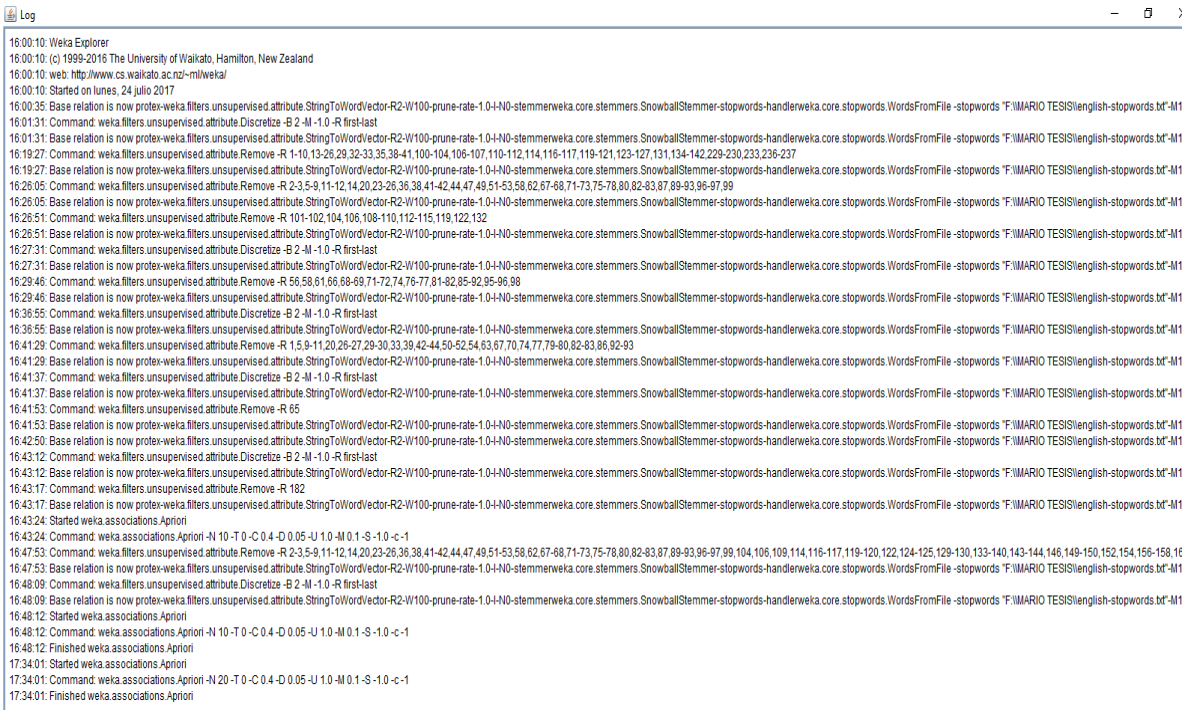
• Durante el procesamiento de los documentos de tipo texto se debe tener cuidado especial al momento de generar las bolsas de palabras, lo anterior debido a que si es configurado de forma incorrecta no se tendrá control acerca de la cantidad de atributos que se van a generar, en este caso se deberá tener presente el campo denominado “words tokeep” el cual se debe parametrizar de tal manera que se generen pocos atributos en la bolsa de palabra. La situación antes mencionada representa la toma de decisión acerca de la cantidad de atributos a generar teniendo en cuenta que de ello depende el contenido de la bolsa de palabras.

• El método Apriori para poder ser implementado necesita de una variable de clase, lo cual es de vital importancia al momento de implementarlo en la herramienta Weka, teniendo en cuenta que si la información del objeto de tipo texto no presenta una variable de clase es necesario

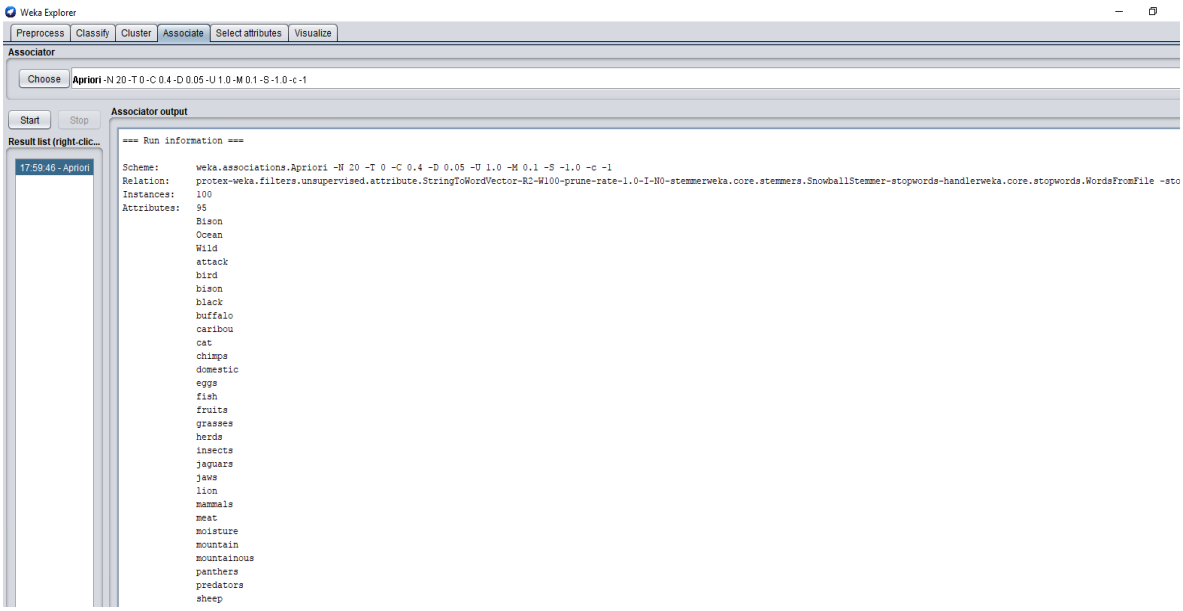
agregarla lo cual dificulta la preparación de los datos teniendo en cuenta que los valores que pueden ser tenidos en cuenta como variable de clase pueden ser muy diferentes.

## 6.6 Resultados obtenidos

Finalmente, en la etapa de resultados no se pueden apreciar las reglas generadas, la cantidad de atributos manejados para la implementación del método Apriori corresponde a 100 registros los cuales contienen 95 atributos que corresponden a los datos extraídos de texto. A continuación, se presentan una serie de imágenes las cuales permiten visualizar los pasos realizados para la implementación del método Apriori sobre los datos objeto de análisis.




*Figuras 11. Pantalla del Log generado previo a la implementación*
*Fuente: Elaboración propia*




*Figuras 12. Pantalla de Ejecución del método A priori*
*Fuente: Elaboración propia*

```
Best rules found:

 1. panda='(-inf-1.956012]' 98 ==> bamboo='(-inf-1.956012]' 98    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.96)
 2. bamboo='(-inf-1.956012]' 98 ==> panda='(-inf-1.956012]' 98    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.96)
 3. bison='(-inf-1.956012]' 97 ==> Bison='(-inf-1.956012]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 4. squirrels='(-inf-2.302585]' 97 ==> caribou='(-inf-2.302585]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 5. jaguars='(-inf-2.302585]' 97 ==> panthers='(-inf-2.302585]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 6. herd='(-inf-2.302585]' 97 ==> Forest='(-inf-2.302585]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 7. rhinos='(-inf-1.753279]' 97 ==> lip='(-inf-2.302585]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 8. savanna='(-inf-2.302585]' 97 ==> scrub='(-inf-2.302585]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
 9. hibernate='(-inf-2.302585]' panda='(-inf-1.956012]' 97 ==> bamboo='(-inf-1.956012]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
10. hibernate='(-inf-2.302585]' bamboo='(-inf-1.956012]' 97 ==> panda='(-inf-1.956012]' 97    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.94)
11. stripes='(-inf-1.956012]' 96 ==> zebra='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
12. fox='(-inf-1.609438]' 96 ==> Fox='(-inf-1.609438]' 96    <conf:(1)> lift:(1.04) lev:(0.04) [3] conv:(3.84)
13. Fox='(-inf-1.609438]' 96 ==> fox='(-inf-1.609438]' 96    <conf:(1)> lift:(1.04) lev:(0.04) [3] conv:(3.84)
14. bison='(-inf-1.956012]' Forest='(-inf-2.302585]' 96 ==> Bison='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
15. bison='(-inf-1.956012]' cow='(-inf-1.956012]' 96 ==> Bison='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
16. bison='(-inf-1.956012]' mating='(-inf-2.302585]' 96 ==> Bison='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
17. Bison='(-inf-1.956012]' herd='(-inf-2.302585]' 96 ==> Forest='(-inf-2.302585]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
18. Bison='(-inf-1.956012]' Bears='(-inf-1.956012]' 96 ==> mating='(-inf-2.302585]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
19. Bison='(-inf-1.956012]' panda='(-inf-1.956012]' 96 ==> bamboo='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
20. Bison='(-inf-1.956012]' bamboo='(-inf-1.956012]' 96 ==> panda='(-inf-1.956012]' 96    <conf:(1)> lift:(1.02) lev:(0.02) [1] conv:(1.92)
```

**F***iguras 13. Pantalla de Ejecución del método A priori*
*Fuente: Elaboración propia*

Las reglas, como se puede observar, son generadas con los 4 tipos de métricas. Confidence, lift, leverage y conviction. Las reglas generadas podrán ser utilizadas para establecer relación

entre los elementos que componen el objeto de tipo texto. A continuación, se describen algunas de las reglas creadas por el método apriori:

• Entre los elementos de texto que componen el objeto, cada vez que se encuentra escrita la palabra Panda también se encuentra la palabra bamboo. Esta regla es evaluada a través de la métrica denominada confianza la cual arroja un valor de 100%

• Entre los elementos de texto que componen el objeto, cada vez que se encuentra escrita la palabra Squirrels también se encuentra la palabra Caribou. Esta regla es evaluada a través de la métrica denominada confianza la cual arroja un valor de 100%

• Entre los elementos de texto que componen el objeto, cada vez que se encuentra escrita la palabra jaguars también se encuentra la palabra panthers. Esta regla es evaluada a través de la métrica denominada confianza la cual arroja un valor de 100%.

## 6.7 Tabla comparativa de métricas

Al realizar la ejecución del algoritmo A priori, teniendo en cuenta la generación de 20 reglas de asociación y teniendo en cuenta 2 diferentes tipos de métricas (confidence y lift), se obtuvieron los siguientes resultados:

**Tabla 1**
***Comparación de Métricas***

| Método | Métricas | |
|---|---|---|
| Apriori | Confidence | Lift |
| | 1 | 1.02 – 1.04 |

Datos obtenidos al aplicar el método A priori en la herramienta Weka. Fuente: Elaboración propia

En la métrica confidence, se determinaron valores de 1 en todas las reglas de asociación. Quiere decir que en un 100% de las veces se cumplen las reglas de asociación extraídas por el software.

En el caso de la métrica lift, podemos notar que los valores arrojados entre 1.02 y 1.04, indican que hay una correlación positiva entre los datos. Es decir, indica que ese conjunto aparece una cantidad de veces superior a lo esperado bajo condiciones de independencia (por lo que se puede deducir que existe una relación que hace que los productos se encuentren en el conjunto más veces de lo normal). En los casos en donde el valor es menor que uno, la correlación sería negativa.

# 7 Conclusión

En conclusión, se da por finalizada este trabajo de investigación, con un resumen de las conclusiones obtenidas al utilizar el Método de clasificación basado en asociación "A priori", así como un relato de las principales contribuciones alcanzadas por el mismo. Asimismo, en el apartado 7.2 se describen algunas direcciones futuras que pueden ser llevadas a cabo.

Para realizar este trabajo, en primer lugar, se consideró la revisión bibliográfica respeto a la clasificación basada en asociación en documentos de tipo texto y a las principales limitaciones inherentes a cada categoría de métodos de recomendación. A partir de dicha revisión bibliográfica se pudo concluir que, para lograr recomendaciones eficaces, los clasificadores asociativos, así como otros métodos de minería de datos, necesitan ser adaptados de acuerdo a las limitaciones inherentes, al tipo y tamaño del conjunto de datos y al tipo de método de recomendación aplicado.

Los objetos de tipo texto, son un recurso disponible en constante crecimiento, actualmente son muy utilizados para suplir diferentes necesidades, la minería de datos permite el análisis de fuentes masivas de datos, por lo tanto, representa una alternativa para la exploración de los objetos de tipo texto en búsqueda de la adquisición de conocimiento relevante.

El algoritmo Apriori es un método de las reglas de asociación el cual es utilizado de acuerdo a la literatura consultada para dar solución a múltiples problemáticas, es por ello que para el análisis de la afinidad entre objetos de tipo texto es una alternativa que puede lograr resultados de gran nivel.

Las reglas generadas a partir del método Apriori poseen valores altos en relación a la métrica denominada confianza, logrando valores del 100% de efectividad al momento de identificar relación entre los elementos que componen documentos u objetos de tipo texto.

La experimentación realizada permite plantear las reglas de asociación como método para el análisis de todos los elementos que hacen parte de un conjunto de objetos de tipo texto, logrando la adquisición de nuevo conocimiento a partir de su estructura o atributos.

## 7.1 Trabajos futuros

En trabajos futuros es posible la implementación de las reglas de asociación para el análisis de elementos diferentes a objetos de tipo texto. Por ejemplo, imágenes, audio y video los cuales conformen el llamado objetos multimedia.

Adicional es posible utilizar otros métodos para el análisis y posterior estudio de objetos de tipo texto, identificando su comportamiento y comparando los resultados obtenidos en comparación con los ya estudiados.

# 8 Bibliografía

[Agrawal, 1993] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

[Agrawal, 1994] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

[Agrawal, 1996] Agrawal, R., & Shafer, J. C. (1996). Parallel mining of association rules. IEEE Transactions on Knowledge & Data Engineering, (6), 962-969.

[Alghmdi, 2014] Alghamdi, R. A., Taileb, M., & Ameen, M. (2014, April). A new multimodal fusion method based on association rules mining for image retrieval. In Mediterranean Electrotechnical Conference (MELECON), 2014 17th IEEE (pp. 493-499). IEEE.

[Amir, 2005] Amir, A., Aumann, Y., Feldman, R., & Fresko, M. (2005). Maximal association rules: A tool for mining associations in text. Journal of Intelligent Information Systems, 25(3), 333-345.

[Azevedo, 2007] Azevedo, P. J., & Jorge, A. M. (2007). Comparing rule measures for predictive association rules. In Machine Learning: ECML 2007 (pp. 510-517). Springer Berlin Heidelberg.

[Berzal, 2002] Berzal., F., Blanco., I., Sanchez., D., & Vila., M. (2002). Measuring the accuracy and interest of association rules: A new

framework. In Journal Intelligent Data Analysis, Vol 6(3), p.p. 221-235.

[Bhandari, 2015] Bhandari, A., Gupta, A., & Das, D. (2015). Improvised Apriori Algorithm using frequent pattern tree for real time applications in data mining. Procedia Computer Science, 46, 644-651.

[Breese. 1997] John S. Breese, David Heck and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, *ACM SIGMOD International Conference on Management of Data*, pages 255–264. ACM Press, 05 1997.

[Brin, 1997] Brin, S., Motwani, R., & Silverstein, C. (1997, June). Beyond market baskets: Generalizing association rules to correlations. In ACM SIGMOD Record (Vol. 26, No. 2, pp. 265-276). ACM.

[Chaves, 2012] Chaves, R., Ramírez, J., Górriz, J. M., Puntonet, C. G., & Alzheimer's Disease Neuroimaging Initiative. (2012). Association rule-based feature selection method for Alzheimer's disease diagnosis. Expert Systems with Applications, 39(14), 11766-11774.

[Chen, 2007] Chen, M., Chen, S. C., & Shyu, M. L. (2007, April). Hierarchical temporal association mining for video event detection in video databases. In Data Engineering Workshop, 2007 IEEE 23rd International Conference On (pp. 137-145). IEEE.

[Chen, 2010] Chen, C. L., Tseng, F. S., & Liang, T. (2010). An integration of WordNet and fuzzy association rule mining for multi-label

document clustering. Data & Knowledge Engineering, 69(11), 1208-1226.

[Cherfi, 2006] Cherfi, H., Napoli, A., & Toussaint, Y. (2006). Towards a text mining methodology using association rule extraction. Soft Computing, 10(5), 431-441.

[Chiang, 2008] Chiang, D. A., Keh, H. C., Huang, H. H., & Chyr, D. (2008). The Chinese text categorization system with association rule and category priority. Expert Systems with Applications, 35(1), 102-110.

[Das, 2001] Das, A., Ng, W. K., & Woon, Y. K. (2001, October). Rapid association rule mining. In Proceedings of the tenth international conference on Information and knowledge management (pp. 474-481). ACM.

[Domingues, 2004] Domingues., M. (2004).Generalization of association rules (Tesis de Maestria). Escola de Engenharia de São Carlos, Brasil.

[Dua, 2009] Dua, S., Singh, H., & Thompson, H. W. (2009). Associative classification of mammograms using weighted rules. Expert systems with applications, 36(5), 9250-9259.

[Geng, 2006] Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38(3), 9.

[Goncalves, 1998] Léren P. F Gonçalves. Mineração de dados em supermercados: O caso do supermercado "tal". Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 1999.

[Grosky, 1997] Grosky, W. I. (1997). Managing multimedia information in database systems. Communications of the ACM, 40(12), 72-80.

[Han, 2000] Han, J., Pei, J., & Yin, Y. (2000, May). Mining frequent patterns without candidate generation. In ACM SIGMOD Record (Vol. 29, No. 2, pp. 1-12). ACM.

[Han, 2004] Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery, 8(1), 53-87.

[Hanguang, 2012] Hanguang, L., & Yu, N. (2012). Intrusion detection technology research based on apriori algorithm. Physics Procedia, 24, 1615-1620.

[Harrison, 1998] Thomas Harrison. *Intranet Data Warehouse*. 1998

[Herawan, 2011] Herawan, T., & Deris, M. M. (2011). A soft set approach for association rules mining. Knowledge-Based Systems, 24(1), 186-195.

[Holt, 2001] Holt, J. D., & Chung, S. M. (2001). Multipass algorithms for mining association rules in text databases. Knowledge and Information Systems, 3(2), 168-183.

[Hu, 2014] Hu, C., Xu, Z., Liu, Y., Mei, L., Chen, L., & Luo, X. (2014). Semantic link network-based model for organizing multimedia big data. IEEE Transactions on Emerging Topics in Computing, 2(3), 376-387.

[Huang, 2010] Huang, C. J., Liao, J. J., Yang, D. X., Chang, T. Y., & Luo, Y. C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications, 37(9), 6409-6413.

[Hunter, 2003] Hunter, J., & Choudhury, S. (2003). Implementing preservation strategies for complex multimedia objects. In Research and Advanced Technology for Digital Libraries (pp. 473-486). Springer Berlin Heidelberg.

[Jiang,2009] Jiang, T., & Tan, A. H. (2009). Learning image-text associations. IEEE Transactions on Knowledge and Data Engineering, 21(2), 161-177.

[Juan, 2010] Juan, L., & De-ting, M. (2010, October). Research of an association rule mining algorithm based on FP tree. In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on (Vol. 1, pp. 559-563). IEEE.

[Karabatak, 2009] Karabatak, M., & Ince, M. C. (2009). A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases. Expert Systems with Applications, 36(10), 12500-12505.

[Karabatak, 2009a] Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2), 3465-3469.

[Khare, 2012] Khare, V. R., & Chougule, R. (2012). Decision support for improved service effectiveness using domain aware text mining. Knowledge-Based Systems, 33, 29-40.

[Kotsiantis, 2006] Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering, 32(1), 71-82.

[Li, 2001] Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on (pp. 369-376). IEEE.

[Li, 2014] Li, Y., & Wu, J. (2014). Interpretation of association rules in multi-tier structures. International Journal of Approximate Reasoning, 55(6), 1439-1457.

[Little, 1990] Little, T. D., & Ghafoor, A. (1990). Synchronization and storage models for multimedia objects. Selected Areas in Communications, IEEE Journal on, 8(3), 413-427.

[Liu, 1998] Liu., B., Hsu., W., & Ma., Y. (1998, August). Integrating classification and association rule mining. In Proceedings of the fourth international conference on knowledge discovery and data mining.

[Lopes, 2007] Lopes, A. A., Pinho, R., Paulovich, F. V., & Minghim, R. (2007). Visual text mining using association rules. Computers & Graphics, 31(3), 316-326.

[Malik, 2006] Malik, H. H., & Kender, J. R. (2006, July). Clustering web images using association rules, interestingness measures, and hypergraph partitions. In Proceedings of the 6th international conference on Web engineering (pp. 48-55). ACM.

[Mustafa, 2006] Mustafa, M. D., Nabila, N. F., Evans, D. J., Saman, M. Y., & Mamat, A. (2006). Association rules on significant rare data using second support. International Journal of Computer Mathematics, 83(1), 69-80.

[Nagata, 2014] Nagata, K., Washio, T., Kawahara, Y., & Unami, A. (2014). Toxicity prediction from toxicogenomic data based on class association rule mining. Toxicology Reports, 1, 1133-1142.

[Narvekar, 2015] Narvekar, M., & Syed, S. F. (2015). An Optimized Algorithm for Association Rule Mining Using FP Tree. Procedia Computer Science, 45, 101-110.

[Pinho, 2010] Pinho., J. (2010). Métodos de Clasificación basados en asociación aplicados a sistemas de Recomendación (Tesis de Doctorado). Universidad de Salamanca, España.

[Quinlan, 1993] Quinlan, J. R., & Cameron-Jones, R. M. (1993, January). FOIL: A midterm report. In Machine Learning: ECML-93 (pp. 1-20). Springer Berlin Heidelberg.

[Sahoo, 2015] Sahoo, J., Das, A. K., & Goswami, A. (2015). An efficient approach for mining association rules from high utility itemsets. Expert Systems with Applications, 42(13), 5754-5778.

[Savarese, 1995] Savasere, A., Omiecinski, E. R., & Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases.

[Song, 2007] Song, M., Song, I. Y., Hu, X., & Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. Data & Knowledge Engineering, 63(1), 63-75.

[Soysal, 2015] Soysal, Ö. M. (2015). Association rule mining with mostly associated sequential patterns. Expert Systems with Applications, 42(5), 2582-2592.

[Tang, 2013] Tang, H. J., Yan, D. F., & Yuan, T. I. A. N. (2013). Semantic dictionary based method for short text classification. The Journal of China Universities of Posts and Telecommunications, 20, 15-19.

[Tesic, 2003] Tešic, J., Newsam, S., & Manjunath, B. S. (2003). Mining image datasets using perceptual association rules. In Proc. SIAM Sixth Workshop on Mining Scientific and Engineering Datasets in conjunction with SDM.

[Thabtah, 2004] Thabtah, F., Cowling, P., & Peng, Y. (2004, November). MMAC: A new multi-class, multi-label associative classification approach. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 217-224). IEEE.

[Tsuji, 2014] Tsuji, K., Takizawa, N., Sato, S., Ikeuchi, U., Ikeuchi, A., Yoshikane, F., & Itsumura, H. (2014). Book Recommendation Based on Library Loan Records and Bibliographic Information. Procedia-Social and Behavioral Sciences, 147, 478-486.

[Wong, 1999] Wong, P. C., Whitney, P., & Thomas, J. (1999). Visualizing association rules for text mining. In Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on (pp. 120-123). IEEE.

[Xiang, 2012] Xiang, L. I. (2012). Simulation System of Car Crash Test in C-NCAP Analysis Based on an Improved Apriori Algorithm*. Physics Procedia, 25, 2066-2071.

[Xu, 2011] Xu, Y., Li, Y., & Shaw, G. (2011). Reliable representations for association rules. Data & Knowledge Engineering, 70(6), 555-575.

[Yang, 2008] Yang, Y., Zhuang, Y. T., Wu, F., & Pan, Y. H. (2008). Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. Multimedia, IEEE Transactions on, 10(3), 437-446.

[Yin, 2003] Xiaoxin Yin and Jiawei Han. Cpar: Classification based on predicttive association rules. In *SIAM International Conference on Data Mining (SDM03)*, pages 331–335, 2003.

[Yin, 2003] Yin, X., & Han, J. (2003, May). CPAR: Classification based on Predictive Association Rules. In SDM (Vol. 3, pp. 369-376).

[Yin, 2006] Yin, P. Y., & Li, S. H. (2006). Content-based image retrieval using association rule mining with soft relevance feedback. Journal of Visual Communication and Image Representation, 17(5), 1108-1125.

[Zaki, 1997] Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997, August). New Algorithms for Fast Discovery of Association Rules. In *KDD* (Vol. 97, pp. 283-286).

[Zheng, 2006] Zheng, Q. F., Wang, W. Q., & Gao, W. (2006, October). Effective and efficient object-based image retrieval using visual phrases. In Proceedings of the 14th ACM international conference on Multimedia (pp. 77-80). ACM.

[Zhuang, 2008] Zhuang, Y. T., Yang, Y., & Wu, F. (2008). Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. Multimedia, IEEE Transactions on, 10(2), 221-229.