UNIVERSIDAD DE PAMPLONA

Procedimiento de Minería de Datos para el Análisis Cienciometrico del OJS(Open Journal System)

AUTOR

JORGE ANDRES SANCHEZ CARRILLO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DEPARTAMENTO DE ELECTRÓNICA, ELÉCTRICA, SISTEMAS Y TELECOMUNICACIONES

PAMPLONA

2016

UNIVERSIDAD DE PAMPLONA

Procedimiento de Minería de Datos para el Análisis Cienciometrico del OJS(Open Journal System

AUTOR

JORGE ANDRES SANCHEZ CARRILLO

DIRECTOR

PH.D NELSON FERNANDEZ PARADA

CO-DIRECTOR

M.G LUIS ALBERTO ESTEBAN VILLAMIZAR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DEPARTAMENTO DE ELECTRÓNICA, ELÉCTRICA, SISTEMAS Y TELECOMUNICACIONES

PAMPLONA

2016

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

INTRODUCCIÓN

1.1. Resumen

Las técnicas de minería de datos en la actualidad son una potente herramienta que apoya procesos vitales en diversas áreas del conocimiento y de la vida cotidiana. Aplicada al ámbito de la producción académica permite identificar patrones y comportamientos de los elementos involucrados en la elaboración de material científico como artículos y revistas científicas. En el presente trabajo se muestra la construcción de un procedimiento de minería de datos para el análisis cienciometrico del *OJS* (Open Journal System), el cual es la plataforma que almacena la información las revistas y artículos de la universidad de Pamplona. Para el la construcción de este procedimiento se tuvo como fundamento la metodología *CRISP-DM* para el proceso de minería de datos y procesos de Descubrimiento de conocimiento *(KDD)* cuyas fases permiten un manejar de manera ordenada lo que tiene que ver con la definición del conjunto de datos a estudiar, su correcta adaptación para el proceso de minería de datos y su correspondiente interpretación. Palabras clave: (Minería de datos, procedimiento, cienciometria).

1.2. Justificación

En los registros documentales de las instituciones y en las fuentes bibliográficas hay una enorme cantidad de datos que dan cuenta de los modos de producción de conocimientos y del trabajo en colaboración de sus investigadores[1] Se dispone de la información correspondiente a las revistas publicadas por la universidad de Pamplona en la herramienta *OJS*, pero no se conoce que tipo de relación existe allí implícitamente en lo que tiene que ver con temas de investigación, investigadores y sus correspondientes productos de investigación. Por lo tanto se definira un procedimiento para la aplicación de técnicas de minería de datos que permita obtener nuevo conocimiento y poder determinar el comportamiento que se presenta en relación a la producción cientifica en base a los productos de investigación y sus correspondientes autores. Los procesos de recopilación y almacenamiento de información de las revistas y documentos científicos son tareas que tienen una gran importancia en las instituciones educativas, ya que permite manejar de forma ordenada toda la producción de conocimiento con que se cuenta, en la actualidad se cuenta con herramientas que gestionan esta información bibliográfica y es el caso de la universidad de Pamplona institución educativa que cuenta con el sistema *OJS* para llevar a cabo esta labor, pero la información simplemente alojada no permite apreciar que patrones o relaciones se presentan entre temas de investigación y los autores de las diferentes investigaciones. En la presente propuesta aprovechando de recursos informáticos se busca aplicar técnicas de minería de datos para determinar estos patrones o comportamientos que estén inmersos en la base de datos del *OJS* que aloja las diferentes revistas de la universidad de Pamplona.

1.3. Objetivos

1.3.1. General

- Definir un procedimiento de minería de datos para el análisis cienciometrico en el *OJS* (Open Journal System)

1.3.2. Específicos

- Realizar un estudio bibliográfico de técnicas de minería de datos aplicable al modelo de datos del software *OJS*.
- Establecer un procedimiento para la aplicación de las técnicas identificadas con la información disponible de las revistas digitales de la Universidad de Pamplona.
- Verificar el procedimiento establecido mediante el procesamiento de la información disponible en el *OJS*.

Capítulo 2

MARCO TEORICO

2.1. Minería de datos

La minería de datos es un proceso muy importante que constituye un campo interdisciplinario que emerge de áreas tales como los sistemas de bases de datos, los data warehose (repositorios de datos), la estadística, el aprendizaje automático, la visualización de datos, la búsqueda y recuperación de la información y de la computación de alta ejecución, además de otras contribuciones procedente de los modelos de redes neuronales, el reconocimiento de patrones, el análisis de datos espaciales, entre otras. Constituye una tarea de descubrimiento de patrones interesantes a partir de grandes volúmenes de datos almacenados en bases de datos y otros repositorios de datos como los data warehouse [2] .

Otros autores aclaran que existen muchos términos que se relacionan o utilizan como sinónimos de la minería de datos, una de ellas es el *KDD* se define como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. A diferencia de la minería de datos es un proceso más complejo que lleva no solo a obtención de modelos o patrones, que es el objetivo de la minería de datos, sino que incluye además una evaluación y una posible interpretación de los mismos [3].

FIGURA 2.1: Proceso de extracción de conocimiento. Hernandez (2005)

2.2. Técnicas de minería de datos de datos

Las técnicas de minería de datos tienen como objetivo explorar una agrupación de datos utilizando herramientas de cómputo como algoritmo de inteligencia artificial, visualizaciones de datos, computación en paralelo, etc. que se encargan de procesar este volumen de datos para poder generar nuevo conocimiento a raíz del ya existente. Según la tarea de moderación las técnicas de minería de datos se clasifican en técnicas descriptivas y predictivas

2.2.1. Técnicas predictivas

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de Minería de Datos, antes de aceptarlo como válido [4]. Técnicas predictivas constan de un conjunto de tareas las cuales son: Clasificación: La clasificación es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir [5] Predicción: se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones [6]. Regresión: El objetivo es predecir los valores de una variable continua a partir de la evolución de otra variable continua, generalmente el tiempo.

2.2.2. Técnicas descriptivas

No se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones [4].
Las técnicas anteriores están compuestas por técnicas auxiliares como lo son:

- Clasificación: Predicen varias variables discretas, teniendo en cuenta los demás atributos del conjunto de datos.
- Regresión: Predicen una o más variables numéricas continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos.
- Agrupamiento: Dividen los datos en grupos, o clusteres, de elementos que tienen propiedades similares.
- Asociación: Con esta técnica se buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas

2.3. Minería de datos supervisada y no supervisada

Los algoritmos pupervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (Atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados que descubren patrones y tendencias en los datos actuales [7].

2.4. Bases de datos

El término base de datos surgió en 1963, en la informática una base de datos consiste en una colección de datos interrelacionados y un conjunto de programas para acceder a dichos de datos. En otras palabras, una base de datos no es más que un conjunto de información (un conjunto de datos) relacionada que se encuentra agrupada o estructurada [8].

2.5. Sistemas de gestión de bases de datos

Consiste en un conjunto de programas utilizados para definir, administrar y procesar una base de datos y sus aplicaciones. A los sistemas de administración de bases de datos también se les llama Sistemas de Gestión de Bases de Datos *(SGBD)*. Un sistema de administración de bases de datos es una herramienta de propósito general que permite crear bases de datos de cualquier tamaño y complejidad y con propósitos específicos distintos [8]

2.6. Algoritmos de minería de datos

AGRUPAMIENTO (cluster): Región continua del espacio que contiene una densidad relativamente alta de puntos, y que se encuentra a su vez separada de otras regiones de alta densidad por regiones cuya densidad de puntos es relativamente baja [9].

2.6.1. Algoritmos de cluster

El clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional y muchas otras.

De forma general, las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y

una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases [10].

K-Means

Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Los pasos básicos para aplicar el algoritmo son muy simples. Primeramente se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones. Luego se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:

- Determina las coordenadas del centroide.
- Determina la distancia de cada objeto a los centroides.
- Agrupa los objetos basados en la menor distancia.

Finalmente quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo [10].

2.6.2. Algoritmos Asociación

Mediante algoritmos de asociación podemos realizar la búsqueda automática de reglas que relacionan conjuntos de atributos entre sí. Son algoritmos no supervisados, ya que no existen relaciones conocidas a priori con las que contrastar la validez de los resultados, sino que se evalúa si esas reglas son estadísticamente significativas. El principal algoritmo es A priori, el cual sólo busca reglas entre atributos simbólicos, por lo cual todos los atributos numéricos deberían ser discretizados previamente [11].

2.6.3. Algoritmos jerárquicos

Los llamados métodos jerárquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración

o división, se minimice alguna distancia o bien se maximice alguna medida de similitud. Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes [12]

2.6.4. Redes Neuronales Artificiales

Las redes neuronales artificiales *(RNA)* esta basada en el sistema neuronal biológico, es un sistema computacional que permite realizar un mapeo de un conjunto de datos o patrones de entrada a un conjunto de salida. Kohonen describe: "Las redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico" [12]

2.7. Bibliomentría

El uso de la minería de datos en el ámbito bibliotecario viene de la mano con el advenimiento de las nuevas tecnologías en las bibliotecas, con la adopción de catálogos automatizados paralelamente se mejoraron las técnicas y los métodos estadísticos de la bibliometría y de la visualización para localizar patrones no comunes inmersos en grandes cantidades de datos. Por consiguiente el bibliomining se refiere al uso de estas técnicas que permiten sondear las enormes cantidades de datos generados por las bibliotecas automatizadas [13]. Un nuevo conocimiento producto de la investigación científica adquiere valor cuando se publica y posteriormente, aplicado en el campo específico, contribuye al desarrollo de la sociedad. La bibliometría juega un papel crucial, ya que le da valor medible al resultado de dicha actividad científica; por consiguiente, se puede situar o comparar la creación de 'X' institución, grupo investigativo, país, etc. Aunque uno de los puntales de la ciencia es el uso de técnicas cuantitativas, hasta tiempos relativamente cercanos no comenzó a aplicarse para estudiar su naturaleza y realidad social. La medida de magnitudes sociales como: los presupuestos científicos, la cantidad de investigadores, las publicaciones científicas, etc., precisan de una técnica de análisis sociológico cuantitativa que corresponden a la disciplina de la cienciometría, aunque no existe unanimidad en el uso de tal término. La bibliometría, por su parte, se centra esencialmente en el cálculo y en el análisis de los valores de lo que es cuantificable en la producción y en el consumo de la información científica [14].

2.8. Cienciometría

La cienciometría estudia los aspectos cuantitativos de la ciencia como disciplina o actividad económica, forma parte de la sociología de la ciencia y encuentra aplicación en el establecimiento de las políticas científicas, donde incluye entre otras las de publicación. Ella emplea, al igual que las otras dos disciplinas estudiadas, técnicas métricas para la evaluación de la ciencia (el término ciencia se refiere, tanto a las ciencias naturales como a las sociales), y examina el desarrollo de las políticas científicas de países y organizaciones. Los análisis cienciométricos analizan a la ciencia como una disciplina o actividad económica, comparan las políticas de investigación desarrolladas por distintos países y sus resultados desde una perspectiva económica y social. Los temas de estudio de la cienciometría incluyen, entre otros [14].

- El crecimiento cuantitativo de la ciencia.
- El desarrollo de las disciplinas y subdisciplinas.
- La relación entre ciencia y tecnología.
- La obsolescencia de los paradigmas científicos.
- La estructura de comunicación entre los científicos.
- La productividad y creatividad de los investigadores.
- Las relaciones entre el desarrollo científico y el crecimiento económico.

2.9. Redes de coautoría

la firma conjunta de un trabajo científico por dos autores, el término sajón clúster en su acepción relacionada con los modelos de grafos para referir el conjunto de nodos o vértices (autores) altamente conectados entre sí mediante arcos o enlaces (relaciones de coautoría), pero con conexiones esporádicas hacia el exterior y el término umbral o intensidad de colaboración, el valor utilizado para formar los clústeres de autores, que hace referencia a la frecuencia de coautoría entre las parejas de autores y que refleja las relaciones más o menos consolidadas entre los mismos a la hora de publicar los resultados de sus investigaciones de forma conjunta [15]. Este valor ha sido utilizado en diversos estudios bibliométricos como

criterio para considerar los clústeres identificados como grupos de investigación. Cuando en lugar de autores se hace referencia a la firma conjunta de un trabajo por dos o más instituciones se utiliza el término colaboración institucional, siendo igualmente aplicables en este caso los términos clúster y umbral o intensidad de colaboración, representando en este caso los nodos las instituciones y los enlaces las relaciones de colaboración [16].

2.10. Redes de cocitación

Es una técnica que se usa para medir la semejanza o similitud entre documentos es la basada en la cocitación. La cocitación no es sino el hecho posible de que dos artículos científicos aparezcan simultáneamente en las referencias de uno tercero. La frecuencia de cocitación se define como la frecuencia con la que dos artículos científicos son citados conjuntamente y es una medida cambiante que puede crecer a medida que transcurre el tiempo. Si se realiza correctamente un análisis de cocitación, se posibilita descubrir los autores o los trabajos más relevantes de una disciplina mediante el consenso empírico establecido por los cientos de citantes de esos autores o trabajos y no sólo por las meras impresiones de los investigadores individuales. Refleja, a diferencia de los análisis de co-palabras, aspectos tanto cognitivos como de vínculos y de relaciones sociales [17]

Capítulo 3

PROCEDIMIENTO

3.1. Modelos de procesos para proyectos de mineria de datos

Son diversos los modelos de proceso que han sido propuestos para el desarrollo de proyectos de minería de datos tales como *SEMMA* (Sample, Explore, Modify, Model, Assess), *DMAMC* (Definir, Medir, Analizar, Mejorar, Controlar), o *CRISP-DM* (Cross Industry Standard Process for Data Mining), sin embargo uno de los modelos principalmente utilizados en los ambientes académico e industrial es el modelo *CRISP-DM*.
A continuación se describen cada una de las fases en que se divide CRISP-DM

3.1.1. Fase de comprensión del negocio o problema

La primera fase de la guía de referencia *CRISP-DM*, denominada fase de comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados [18].

3.1.2. Comprensión de datos

La segunda fase, fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de minería de datos, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a labase de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas [18].

3.2. Fase de preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato [18].

3.2.1. Fase de minería de datos

En esta fase de *CRISP-DM*, se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de Los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.

- Conocimiento de la técnicaFase de implementación.

En esta fase se puede decir que el modelo ha sido validado correctamente y se ha transformado en una funcionalidad que genera un nuevo conocimiento el cual se ejecuta dentro de los proceso de la empresa [18].

3.2.2. Fase de evaluacón

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados [18].

3.2.3. Fase de implementación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados [18].

3.3. Metodología

Para la construcción del procedimiento que contemple la aplicación de técnicas de mimería de datos al sistema ojs para el análisis cienciometrico se tuvo en cuenta la metodología *CRISP-DM* (Cross-Industry Standard Process for Data Mining), que da soporte a los procesos de mineria de datos mediante las etapas que la componen.

3.4. Materiales

3.4.1. R

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de [19]:

- almacenamiento y manipulación efectiva de datos
- operadores para cálculo sobre variables indexadas (Arrays), en particular matrices
- una amplia, coherente e integrada colección de herramientas para análisis de datos
- posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y
- un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

3.4.2. Librerías de R

- Bibliometrix: Proporciona varias rutinas para importar datos bibliográficos de SCOPUS y Thomson Reuters 'Bases de datos de ISI Web of Knowledge, perfomando análisis bibliométrico y matrices de datos de Co-citación, acoplamiento y análisis de colaboración científica [20].
- Igraph: Los objetivos principales de la biblioteca de igraph son proporcionar un conjunto de tipos de datos y funciones para 1) Implementación de algoritmos de gráficos, 2) manejo rápido de gráficos grandes, con millones de vértices y Bordes, 3) permitiendo el prototipado rápido a través de lenguajes de alto nivel como R [21].
- DBI: El paquete DBI define una interfaz común entre el R y los sistemas de gestión de bases de datos (DBMS). La interfaz define un pequeño conjunto de clases y métodos similares en espíritu a DBI de Perl, JDBC de Java, DB-API de Python y ODBC de Microsoft [22].

- Rmysql: Es una interfaz de base de datos y controlador MySQL para R. Esta versión cumple con la definición de interfaz de base de datos implementada en el paquete DBI [23].
- Rattle:Es un paquete escrito en R que proporciona una interfaz gráfica de usuario a muchos otros paquetes R que proporcionan funcionalidad para la minería de datos [24].

3.5. Procedimiento propuesto

Con el objetivo de analizar el comportamiento de la producción científica que existe en la información aportada por el sistema OJS, en este estudio se aplican técnicas de minería de datos que permitan apreciar las relaciones que poseen los elementos de este conjunto de datos, utilizando la metodología CRISP-DM que sienta las bases para el desarrollo de proyectos de minería de datos dando un soporte en todas las fases que componen esta metodología. Este estudio está orientado a la descripción de un conjunto de información aplicando minería de datos con técnicas descriptivas como visualización y agrupamientos. Como parte de la definición del procedimiento se plantea que aplicar la presente metodología que permitaran llegar a un modelo de minería de datos que dara una descripción apropiada del conjunto de datos estudiado.

El siguiente diagrama describe el procedimiento que se plantea, teniendo en cuenta la metodóloga CRISP-DM, para identificar los elementos que se requieren en el proceso deconstrucción del modelos de minería de datos.

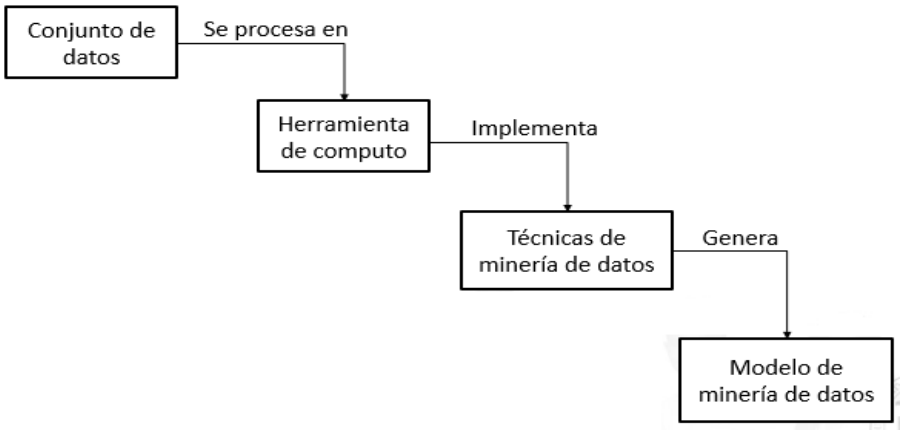


FIGURA 3.1: Flujo de procesos

3.5.1. Comprensión del negocio

3.5.2. Contexto

Open Journal Systems *(OJS)* es una solución de código abierto para la gestión y publicación de revistas académicas en línea. Ofrece un sistema de gran flexibilidad para la gestión y publicación de revistas académicas que puede descargarse sin costo e instalarse en un servidor local, y su funcionamiento queda en manos del equipo editorial de cada institución. El diseño del OJS facilita la reducción del tiempo y la energía que implican las tareas administrativas y de supervisión propias de la edición de revistas académicas, mejora también la conservación de registros y la eficacia de los procesos editoriales [25].

3.5.3. Objetivos del OJS

- Almacenar datos de artículos y revistas.
- Realizar revisiones a los diferentes documentos que se carguen.
- Generar reportes de los registros.
- Realizar consultas de autore, titulos y años por revista.

3.5.4. Evaluación de la situación

El sistema de información OJS es una herramienta de carácter libre que consta de una base de datos la cual puede ser accedida sin ningún problema permitiendo extraer la información necesaria para los estudios de minería de datos. El sistema se puede instalar en cualquier equipo que cuente con el respectivo motero de datos Mysql [26], servidor Apache [27] y soporte para PHP [28].

3.5.5. Criterios de éxito

Obtener una descripción del comportamiento de los datos del OJS utilizando visualizaciones, algoritmos de clustering y reglas de asociación, teniendo en cuenta factores de producción de artículos y citaciones que permitan apreciar cómo trabajan los autores de los diferentes documentos científicos.

3.5.6. Objetivos de la minería de datos

Mediante un conjunto de variables como autores,citaciones y años, identificar como se agrupan los diferentes autores registrados y que figuran como publicados en la base de datos.

3.5.7. Selección de técnicas de minería de datos

Se deben tener en cuenta las técnicas de minería de datos y sus diferentes modos de operación, con el fin de generar el modelo de minería de datos más adecuado que permita un determinado pronóstico de los datos o solo la descripción de cómo se agrupan los elementos estudiados.
A continuación se pueden ver algunas funciones de las técnicas de minería de datos.
Técnicas descriptivas Permiten identificas patrones que explican o resumen los datos como:

- Reglas de asociación
- Clutering

Técnicas predictivas Estiman de variables de interés (a predecir) a partir de valores de otras variables

- regresión
- Clasificación

En este estudio se aplican algunas técnicas de minería de datos, las cuales se han escogido teniendo en cuenta una revisión bibliográfica que permitió determinar cuales se usan en procesos descriptivos ya que no se contempla estudiar el

comportamiento futuro de los datos. A continuación se describen las técnicas de minería de datos que se van a utilizar.
Cluster

Uno de los algoritmos más utilizados para hacer clustering es el k-medias (kmeans), que se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan k elementos aleatoriamente, que representaran el centro o media de cada cluster [6].

El proceso del algoritmo de clúster radica en la división de la información procesada en grupos donde los miembros tienen características similares. Estas similitudes se miden mediante diferentes procesos matemáticos. Esta representación permite obtener una simplificación y fácil comprensión de cómo se comportan los datos.

Reblas de asosición

Permiten expresar patrones de comportamiento entre los datos en función de su aparición conjunta, expresando las combinaciones de valores de los atributos que ocurren más veces. Formalmente son una proposición probabilística sobre la ocurrencia de ciertos estados sobre el conjunto de datos[29].

Algoritmo Apriori

- El algoritmo busca conjuntos de elementos con determinada cobertura mínima.
- Se parte de conjuntos de elementos con un elemento.
- Despues se realiza un proceso incremental hasta que ya no es posible construir conjuntos mas grandes.
- Al final se construye el conjunto de reglas a partir de los conjuntos devueltos.

Capítulo 4

VALIDACIÓN DEL PROCEDIMIENTO

4.1. Comprensión de los datos

En esta sección se muestra cómo se obtuvo la información para el estudio y se analiza para poder determinar unas descripciones de los datos que puedan mostrar detalles de esta información en cuanto a número de registros, cantidad de variables que maneja ese conjunto de datos. Tambien determinar que campos de este conjunto de datos son utiles para el desarrollo del modelo de minería de datos.

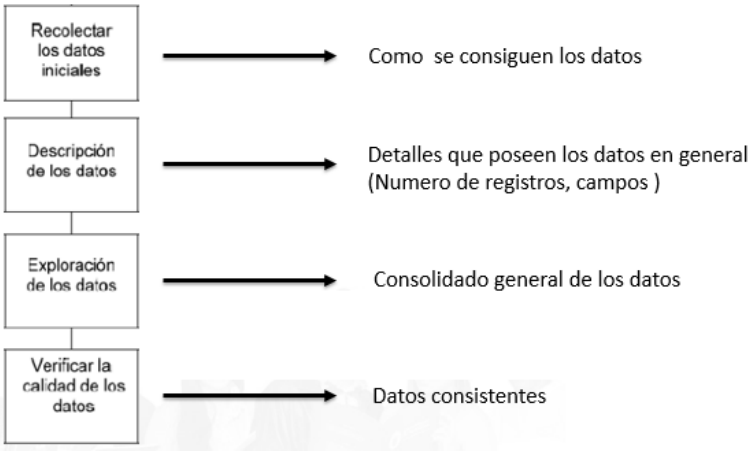


Figura 4.1: Diagrama de análisis de datos

4.1.1. Recolección de la información

Para el desarrollo de este estudio se contó con la ayuda de la vicerrectoría de investigaciones de la universidad de pamplona que mediante una solicitud, permitió el acceso a la base de datos del sistema OJS que esta tiene para el manejo de sus revistas y artículos.

4.1.2. Descripción de la información

En esta sección se revisan características de la información que indiquen volúmenes de datos como número de registros, cantidad de tablas, cantidad de campos, para dimensionar la importancia de los datos que se manejan. Los datos iniciales se encuentran en formato SQL (Lenguaje Estructurado de Consulta) por lo cual esta informacion fue cargada al sistema de bases de datos MYSQL en una base de datos denominada OJS como lo muestra la siguiente imagen(Figura 4.2).

FIGURA 4.2: Estructura del OJS

En la imagen anterior se muestra una pequeña porción de las tablas que componen al sistema OJS, para sabes cuantas tablas lo componen se aplica la siguiente consulta como se ve en la imagen(Figura 4.3).

FIGURA 4.3: Cantidad de tablas del OJS

En lo que tiene que ver a las tablas de la base de datos que son las que poseen las variables que van a permitir construir el conjunto de datos a estudiar, se puede obtener inicialmente la cantidad de registros que estas poseen. La siguiente imagen muestra la cantidad de registros que posee la tabla de autores(Figura 4.4).

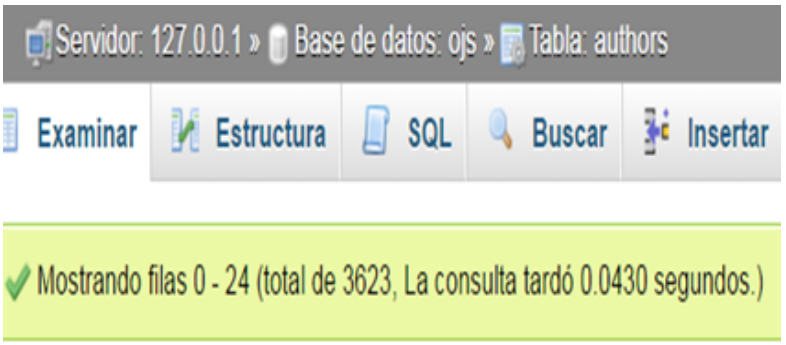


FIGURA 4.4: Cantidad de tablas de la tabla autores

4.2. Exploración de los datos

En esta etapa se utilizó la librería desarrollada para análisis bibliometrico llamada "bibliometrix" [20] que realiza un análisis bibliométrico de un conjunto de datos importado de las bases de datos de *SCOPUS* [30] y Thomson Reuters [31]. Recibiendo un conjunto de parámetros, los cuales pueden ser estudiados dependiendo de la utilidad que se dese aplicar. Inicialmente se emplea la utilidad biblioAnalysis que procesa la matriz con la información que se desea estudiar (Ver anexo 1).

El resultado del proceso de biblioanalisis con la función anterior se asigna a una variable que va a permitir obtener un varios factores de la información incluyendo un conjunto de gráficas que permiten entender estos resultados (Cuadro 4.1).

CUADRO 4.1: Resumen biblioanalisís

Artículos	733
Autores	1276
Apariciones	1761
Unico autor	126
Multiples autores	1150
Artículos por autor	0.574
Autores por artículo	1.74
Co-autores por artículo	2.4
Indice de colaboración	2.22

Se puede determinar valores relacionados con la cantidad de artículos en total, todos los autores involucrados en la muestra, la aparición de varios autores en varios artículos, se observa adicionalmente unos factores como artículos cuya autoría corresponde a un autor o a varios, promedio de autores por artículo, coautoría por artículo e indicador de colaboración. También es posible obtener la producción anual de artículos(Cuadro 4.2):

Cuadro 4.2: Resumen de producción anual

2002	1
2002	1
2004	2
2006	1
2007	1
2008	1
2009	8
2011	20
2012	7
2013	325
2014	104
2015	134
2016	107

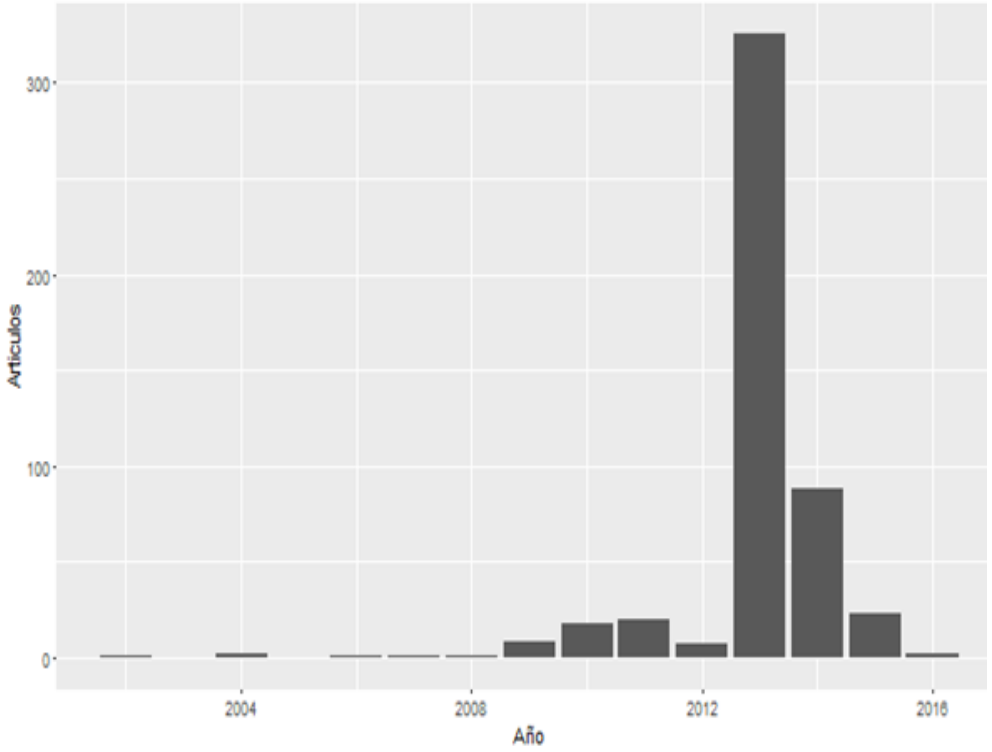


Figura 4.5: Histograma producción anual de artículos

Se cuenta también con un cálculo de una tasa anual de crecimiento porcentual (Annual Percentage Growth Rate) que en este caso fue de: 47.60984

El analisís también muestra resultados en cuanto a autores mas productivos que podemos ver en la siguiente tabla(Cuadro 4.3):

Cuadro 4.3: Resumen de producción anual de artículos

Autores	Artículos	Autores	Artículos fraccionados
A.Parra	16	G.CoteParra	13.00
Y.Navarro	14	E.Capacho	8.08
E.Capacho	13	D.Reyes	7.00
G.CoteParra	13	Y.Navarro	5.58
H.Navia	9	A.Parra	5.25
J.RIVERA	9	J.RIVERA	5.17
M.Rivera	8	H.Navia	4.53
A.Ramón	7	M.OrdoñezSantos	4.33
D.O.	7	M.Garcia	3.70
D.Reyes	7	D.O.	3.50

Se pueden ver las cifras de los autores más productivos en cuanto a generación de artículos(Figura 4.6)

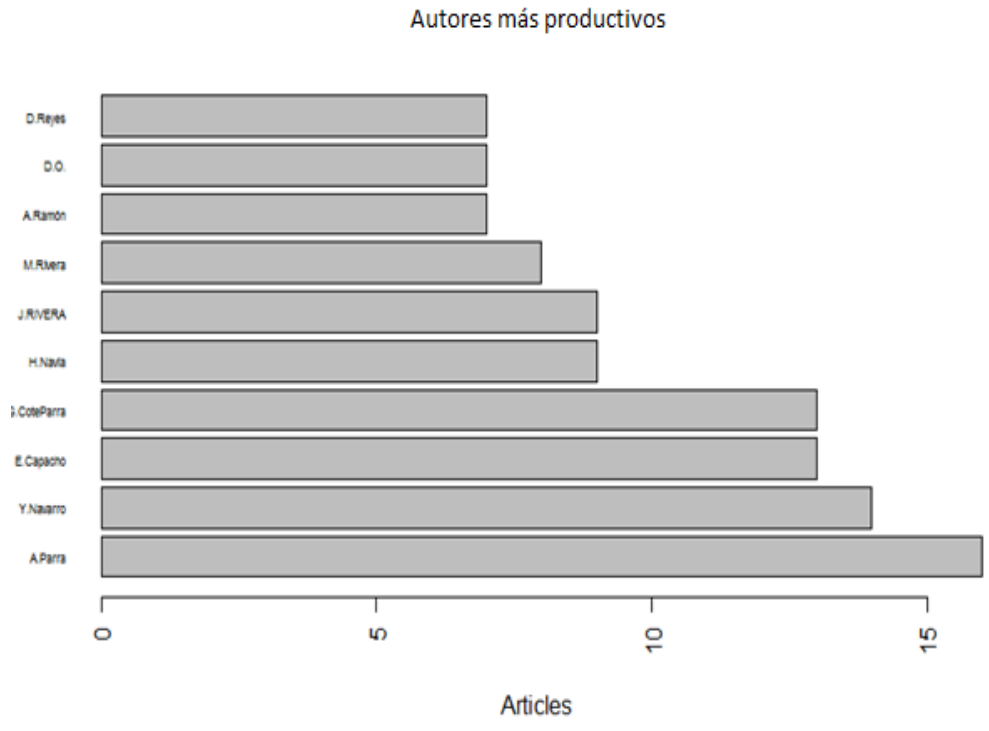


Figura 4.6: Histograma autores más productivos

La cantidad de artículos por las revistas registradas se ve en el cuadro 4.4 y Figura 4.7. En ellas se evidencia la posición de las diferentes revistas según la cantidad de artículos publicados en el transcurso de los años.

Cuadro 4.4: Resumen de producción anual

Sources	Articles
BISTUA	142
FACE	113
ALIMEN	95
CDH	93
RCTA	93
AFDH	58
OWD	55
RA	49
INBIOM	23
COH	8

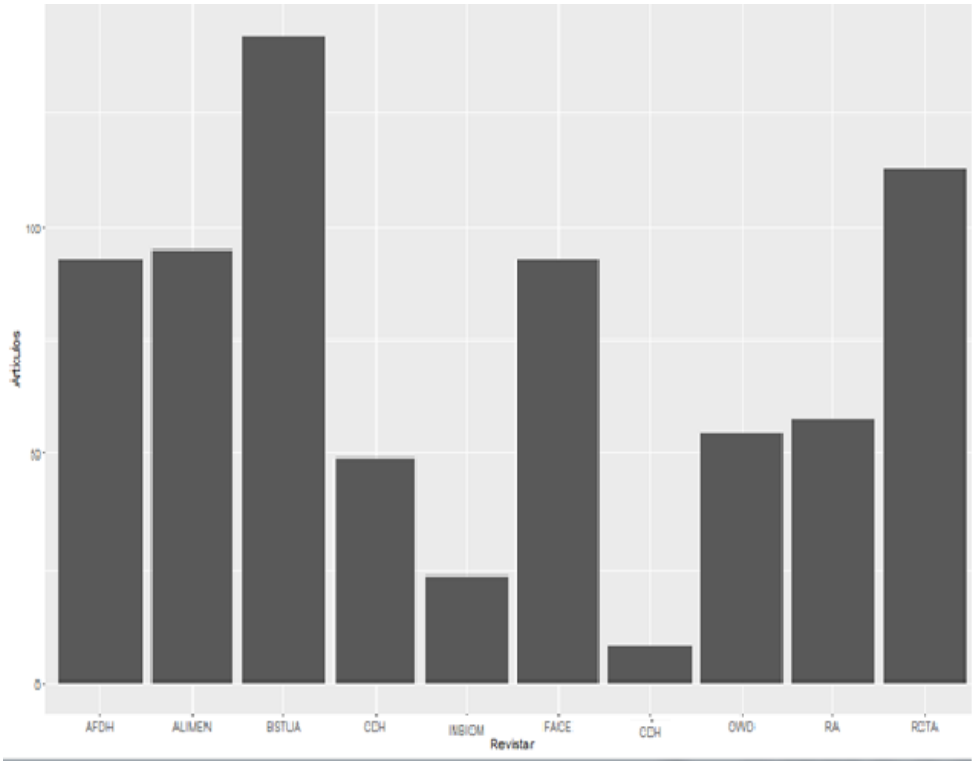


Figura 4.7: Histograma producción de artículos por revista

4.3. Verificación de la calidad de los datos

En esta sección se hace una revisión para identificar y excluir los datos que no se necesiten o que pueden generar diferentes fallos o inconsistencias como lo pueden ser datos de tipo null. En este caso se excuyen los campos suffix y usergroupid que no representan ninguan importancia para el estudio. (ver figura 4.8).

Tabla: authors

Buscar | Insertar | Exportar | Importar | Más

suffix	country	email	url	user_group_id
NULL	CO	yaninetrujillo@unipamplona.edu.co		*NULL*
NULL	CO	a1@a1.com		*NULL*
NULL	CO	a2@a2.com		*NULL*
NULL	BR	a2@a2.com		*NULL*

FIGURA 4.8: Datos inconsistentes

4.4. Preparación de los datos

En esta sección se construirá la estructura de datos que permita contener la información más relevante, con datos completos y si errores para ser posteriormente procesado con las técnicas de minería de datos para generar el modelo. Las etapas que se tiene en cuenta son(Figura 4.9)

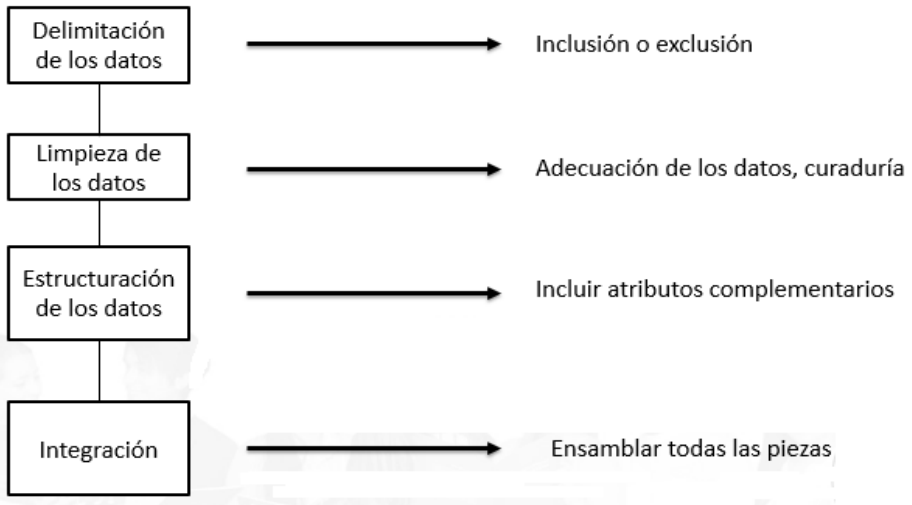


FIGURA 4.9: Preparación de los datos

4.4.1. Selección de los datos

Como fuente principal de información se identificó anteriormente la base de datos que compone al sistema OJS, que es en donde se almacena la información de todas las revistas y demás documentos de producción científica de la universidad de Pamplona. Mediante la revisión de esta base de datos en el entorno del phpmyadmin, se pudo identificar que las tablas que aporta la información más relevante para este estudio corresponden a (Figura 4.10):

- Autores
- Artículos
- Artículos publicados
- Citaciones
- Revistas

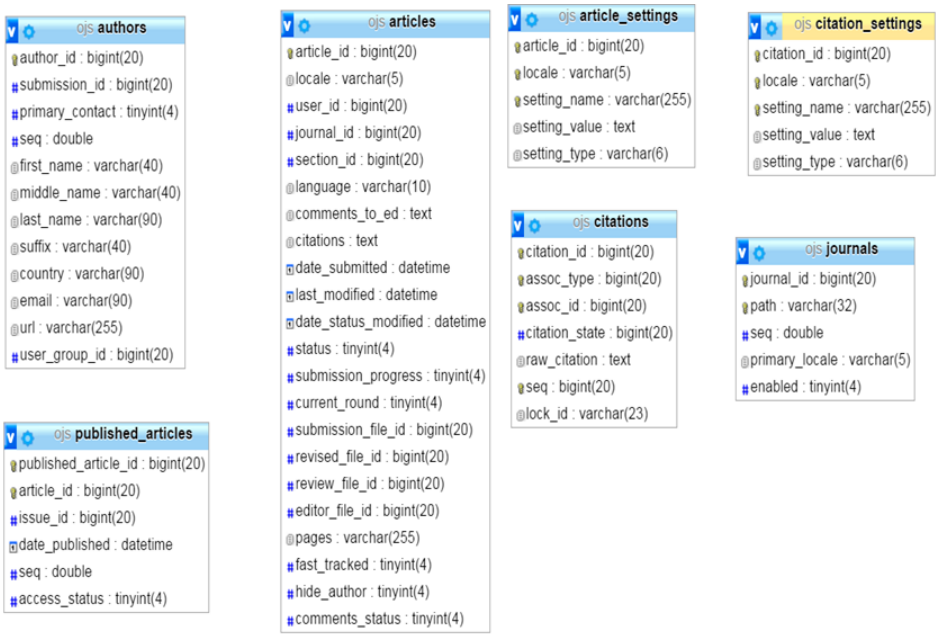


FIGURA 4.10: Base de datos OJS

El conjunto de datos que se va a generar para este estudio se apega a un formato bibliográfico denominado de la empresa ISI, que maneja información acerca de la producción científica de los autores y el cual es base de trabajo para el paquete bibliometrix. En la tabla 4.5 se puede ver los campos que componen el formato ISI.

CUADRO 4.5: Formato ISI [Package bibliometrix version 0.1 Index]

AU	Autores
TI	Título
SO	Revista
DE	Palabras clave
AB	Abstract
TC	Veces citado
PY	Año
DB	Base de datos
CR	Referencias

4.4.2. Limipar los datos

Se procede a seleccionar los campos que van a permitir generar el conjunto de datos de las tablas anteriormente mencionadas en la selección de los datos. En la siguiente figura (figura 4.11), se muestran los campos que se eligieron y su tipo de datos por parte de la tabla autores para añadirlos a el conjunto de datos que se desea establcer.

Nombre	Tipo
author_id	bigint(20)
submission_id	bigint(20)
primary_contact	tinyint(4)
seq	double
first_name	varchar(40)
middle_name	varchar(40)
last_name	varchar(90)
suffix	varchar(40)
country	varchar(90)
email	varchar(90)
url	varchar(255)
user_group_id	bigint(20)

FIGURA 4.11: Tabla autores

Para los datos que tienen que ver con los artículos se realizó el mismo proceso que contempla incluir solo los datos necesarios para generar el conjunto de datos (Figura 4.12).

Nombre	Tipo
article_id	bigint(20)
locale	varchar(5)
user_id	bigint(20)
journal_id	bigint(20)
section_id	bigint(20)
language	varchar(10)
comments_to_ed	text
citations	text
date_submitted	datetime
last_modified	datetime
date_status_modified	datetime
status	tinyint(4)

FIGURA 4.12: Tabla artículos

Otro factor de esta etapa tiene que ver con la delimitación del conjunto de datos, en este caso se tendrán en cuenta los datos que corresponden a los artículos que han sido publicados en las revistas con que cuenta el sistema en su base de datos. Lo cual se puede ver en la siguiente imagen(Figura 4.13).

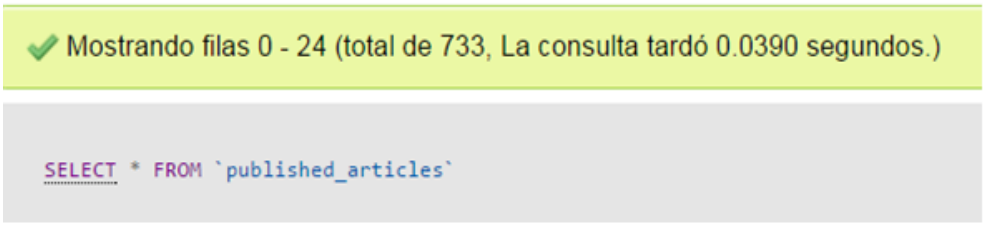


FIGURA 4.13: conteo artículos publicados

Realizando la conexión se puede proceder a ejecutar consultas tal y como se haría en la linea de comandos del intérprete de *MYSQL*(Ver anexo 2). En el proceso de ensamble del conjunto de datos a procesar se implementaron un conjunto de consultas como es el caso de extraer los artículos y sus respectivos autores (Ver anexo 3).

En el anexo anterior se muestra como se extraen los datos correspondientes a los autores y sus respectivas publicaciones teniendo en cuenta criterios como: la revista

a la que está afiliada la publicación, si aparece dentro de los artículos publicados en estas revistas. Finalmente estos datos son guardados en una estructura de datos denominada DATAFRAME que permite realizar operaciones como a cualquier matriz. Terminanda la construcción de la estructura de datos que almacenara la información a estudiar se puede ver que queda de la siguiente forma (Figura 4.14):

AU	TI	SO	DE	AB
A.Diaz;N.Mohallem;R.Millan	ESTUDIO MÖSSBAÜER DE UN NANOCOMPÓSITO PREP...	BISTUA	estudio;mössbaüer;nanocompósito;preparado;partir;...	<p>Se ha estudiado un nanocompósito preparado a ...
A.Parra;Y.Quintero;M.Camargo;J.B	METODO FOTOMETRICO PARA LA DETERMINACIÓN DE...	BISTUA	metodo;fotometrico;para;determinación;concentraci...	<p>Uno de los contaminantes en el ambiente es el O...
N.Fernández;A.Ramírez;F.Solano	PHYSICO-CHEMICAL WATER QUALITY INDICES- A COM...	BISTUA	physico-chemical;water;quality;indices-;comparative...	<p>Water quality assessment can be defined as the ...
E.G.;L.Blanco	APLICACIÓN DEL MODELO DE INTERACCIÓN IÓNICA D...	BISTUA	aplicación;del;modelo;interacción;iónica;pitzer;coefi...	<p>Los coeficientes osmóticos de soluciones de KC...
A.Salazar;J.Rueda	ESTUDIO DEL ACOPLAMIENTO ENERGÉTICO EN UN MA...	BISTUA	estudio;del;acoplamiento;energético;material;fotorre...	<p>A partir de un modelo de interacción de cuatro o...
N.Fernández;A.Ramírez;F.Solano	Dinámica Fisicoquímica y Calidad del Agua en la Micr...	BISTUA	NA	<p>El presente estudio se refiere a la caracterizació...
C.Demicheli;R.Bejarano;R.Sinisterra	PREPARACIÓN DE UN COMPLEJO ANTIMONIO- -CICLOD...	BISTUA	preparación;complejo;antimonio-;-ciclodextrina;para;...	<p>Existe la necesidad de mejorar las limitaciones e...
A.Meneses;E.Hernández	IDENTIFICACION DE EMISIONES DIRECTAS E INDIRECT...	BISTUA	identificacion;emisiones;directas;indirectas;gel;sect...	<p>La planta de tratamiento de aguas residuales Río ...
O.Cáceres;P.Ruiz;G.de Santafé;;A.Marciales.	COMPARACIÓN DEL ESTADO NUTRICIONAL Y EL REND...	BISTUA	comparación;del;estado;nutricional;rendimiento;aca...	<p>En términos generales el estado nutricional se d...
J.G.;M.A.;Y.A;A.G.;A.C.;M.C.	EDAD MATERNA Y/O PATERNA COMO FACTOR DE RIES...	BISTUA	edad;materna;paterna;como;factor;riesgo;genético;ni...	<p>Se estudió, mediante la realización de cariotipos,...
G.Restrepo	LOS ELEMENTOS QUÍMICOS, SU MATEMÁTICA Y RELA...	BISTUA	NA	<p>El sistema periódico de los elementos químicos v...
C.Puentes;J.Castro	TRANSFORMACIÓN QUÍMICA DE LOS 4-N-BENCIL (?-FE...	BISTUA	NA	<p>An easy and simple synthetic route allowed the ...
J.Galviz;F.Ortega;L.Montaño;F.Gamboa	RIQUEZA Y DISTRIBUCION DE LAS ORQUIDEACEAE EN ...	BISTUA	riqueza;distribucion;las;orquideaceae;provincia;pam...	<p>Se estimó la riqueza y distribución de las orquíd...

FIGURA 4.14: Matriz de datos

4.5. Modelado

En esta face se contempla el conjunto de datos que se ha construido con todas las etapas anteriores, con el fin de utilizar una herramienta computacional y asi poder procesar esta información y generar el modelo de mineria de datos.

4.5.1. Selección de técnicas de minería de datos

Como se estableció este estudio como un proceso descriptivo que desea identificar las agrupaciones y colaboraciones de los autores que están registrados en el sistema OJS, se designan técnicas de minería de datos como el agrupamiento y la asociación(Ver cuadro 4.6) con sus respectivos algoritmos y un grupo de visualizaciones que apoyan el entendimiento de los resultados(ver sección 3.5.7).

Cuadro 4.6: Técnicas y algoritmos

Técnica	Algoritmo
Cluster	k-means
Asociación	A priori

4.6. Construcción del modelo y resultados

4.6.1. Aplicación de técnicas de minería de datos

Redes de coautoría y cocitación

En las diversas funcionalidades que posee el paquete bibliometrix existe una que esta orientada al procesamiento y visualización de redes de coautoría y de co-citación que permiten establecer un esquema gráfico de cómo se distribuyen los diferentes autores en trabajo grupal y la forma en que se citan los trabajos de los otros autores. La función biblioNetwork puede crear una colección de redes bibliográficas siguiendo el enfoque propuesto por Batagely y Cerinsek (2013).
Inicialemente la implementación de esta función se realiza para generar la red de colaboracion basada en los datos de estudio se realiza mediante la siguiente sentencia(Ver anexo 4).
En esta setenencia se ejecuta la herramienta igraph que posee rutinas para gráficos sencillos y análisis de redes. Puede Manejar grandes gráficos muy bien y proporciona funciones para generar aleatorio Y gráficos regulares, visualización de gráficos, métodos de centralidad y mucho más [32].
En la siguiente gráfica se observa como los autores conforman grupos debido a que en varios trabajos ellos participan como múltiple autoría lo que genera la una red de coautoría(ver figura 4.15).

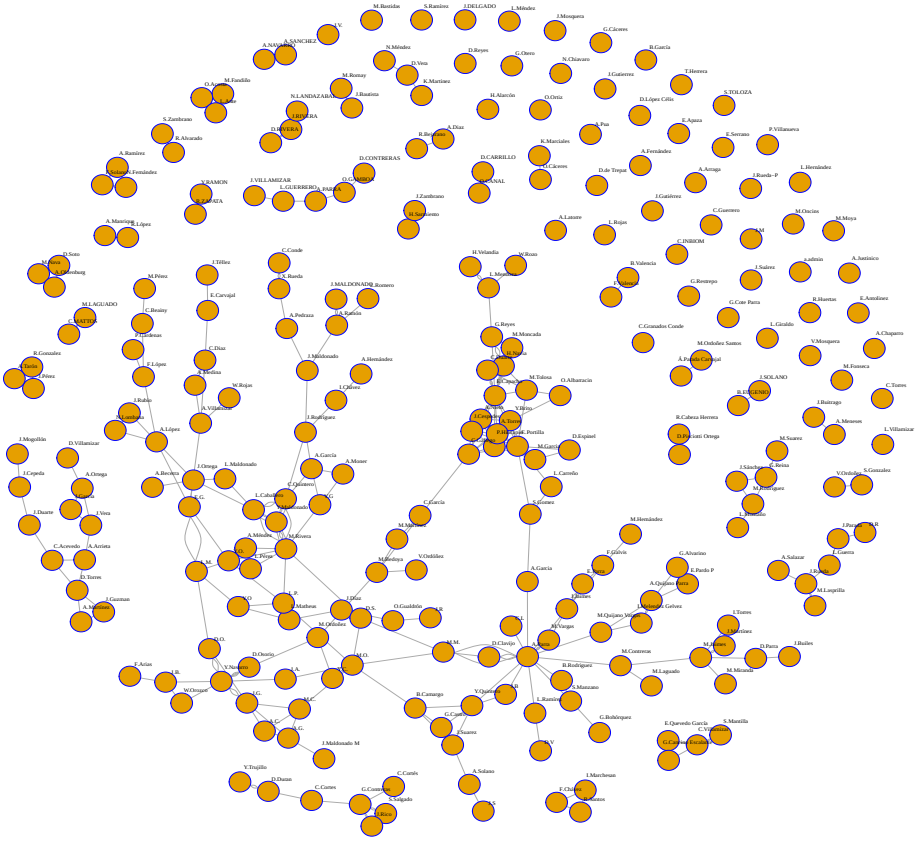


FIGURA 4.15: Red de colaboración

La siguiente funcionalidad que se implemento es la que permite generar la red de cocitación, que indica como los autores incluyen dentro de sus referencias bibliográficas a otros autores que se encuentran dentro este conjunto de datos que se seleccionó. Para este caso se ejecutó una sentencia parecida a la que genera las redes de colaboración salvo que en esta se cambió el parámetro que indica el tipo de red (Ver anexo 5).

En la figura 4.16 que se muestra a continuación permite evidenciar mediante una red de cocitación algunos grupos de autores que se relacionan entre sí, lo que representa las citación de determinados trabajos que realizan los autores para el sustento teórico de sus nuevas producciones.

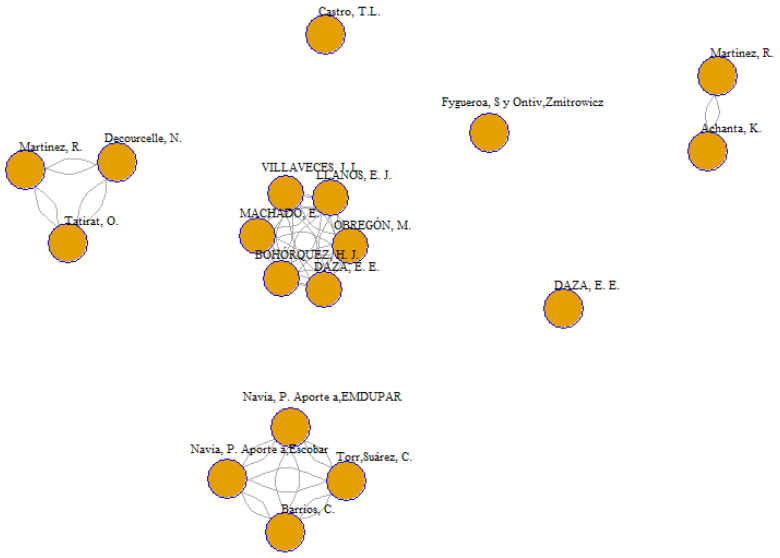


FIGURA 4.16: Red de cocitación

Si se comparan las dos gráficas anteriores de redes de coautoria o colaboración y de cocitación se evidencia fácilmente que el tamaño de la primera es mucho mayor que la segunda, en términos de la cantidad de nodos que representan los autores y los vínculos que unen a estos autores, lo cual nos indica que los trabajos colaborativos entre autores son mayores al actividad de tomar como referencia el trabajo de otro autor.

Aplicación de algoritmos de Agrupamiento

En esta sección se aplicaron algoritmos de agrupamiento (cluster) a los resultados de la ejecución de otro análisis que posee el paquete bibliometrix, el cual permite obtener un factor de dominancia de un conjunto de autores, calculando la posición de dominio de los autores a partir de un objeto de la clase 'bibliometrix'como lo propuso Kumar Y Kumar, 2008 (Ver anexo 6).
La salida de este proceso de puede ver en la siguiente tabla donde este factor permite determinar la relación de los autores entre trabajos que se realizarion como primer autor y trabajos en grupo siendo este el cociente entre estos dos valores. Con esta informacion perteneciante a los autores se pretende ahora identificar como se segmentan o agrupan en términos de este factor de dominancia.Por tal motivo se hace necesario implementar algoritmos de agrupamiento que permitan describir de manera numérica y grafica como se asocian estos indicadores correspondientes a los autores (Figura 4.17).

FIGURA 4.17: Autores por factor de dominancia

Autores	Factor de dominancia	Multi-autoria	Primer autor	Rank por articulo	Posición por DF
A.Ramón	10.000.000	7	7	7	1
X.Rueda	0.8571429	7	6	10	2
C.Acevedo	0.8000000	5	4	17	3
J.Rueda	0.8000000	5	4	19	4
A.Parra	0.7500000	16	12	1	5
M.Garcia	0.7142857	7	5	9	6
Y.Navarro	0.6428571	14	9	2	7
E.G.	0.6000000	5	3	18	8
H.Navia	0.5555556	9	5	4	9
J.RIVERA	0.5555556	9	5	5	10
A.QuijanoParra	0.5000000	6	3	12	11
M.Contreras	0.5000000	6	3	14	12
L.Mendoza	0.4000000	5	2	20	13
D.O.	0.2857143	7	2	8	14
E.Capacho	0.2307692	13	3	3	15
A.García	0.1666667	6	1	11	16
D.Osorio	0.1666667	6	1	13	17
M.OrdoñezSantos	0.1666667	6	1	15	18
N.LANDAZABAL	0.1666667	6	1	16	19
M.Rivera	0.1250000	8	1	6	20

Inicialmente se implementó el paquete cluster que posee todas las funcionalidades para análisis de datos por agrupamiento teniendo en cuenta las variables que se manejen mediante la siguiente sentencia(Ver anexo 7).

El proceso de aplicación de este comando genera unas salidas que corresponden al autor y el clúster al que se le ha asignado según los valores de dominancia. En la siguiente tabla se muestra los autores y el clúster que se les ha asignado.

Para generar una representación gráfica que pueda mostrar la distribución de los autores en sus clúster se debe realizar el llamado de una función clustpot que dibuja un çlusplot"(Package cluster version 2.0.3) bidimensional (gráfico de agrupación) en el dispositivo gráfico actual. La función genérica tiene un método predeterminado y un método de partición que toma esta salida anteriormente generada y establece una representación espacial de los autores en función de su clúster (Ver anexo 7).De forma gráfica se puede apreciar como los autores se agrupan según el clúster al cual fueron asignados.

FIGURA 4.18: Gráfica de cluster por autor

Autor	cluster
A.Ramón	1
M.Garcia	1
A.QuijanoParra	3
A.García	2
X.Rueda	1
Y.Navarro	1
M.Contreras	3
D.Osorio	2
C.Acevedo	3
E.C.	3
L.Mendoza	3
M.OrdoñczSantos	2
J.Rueda	3
H.Navia	1
D.O.	2
N.LANDAZABAL	2
A.Parra	1
J.RIVERA	1
E.Capacho	1
M.Rivera	2

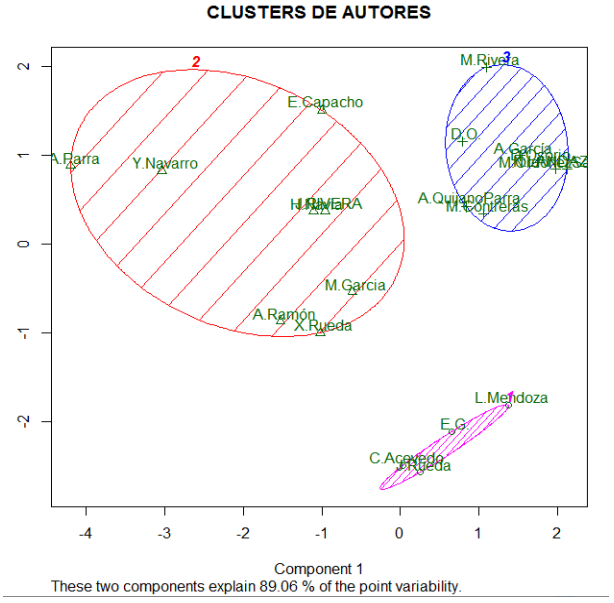


FIGURA 4.19: Gráfica de cluster por autor kmeans

Otra funcionalidad que se tuvo en cuenta para este proceso de aplicación de algoritmos de agrupamiento se denomina factominer(Package FactoMineR version 1.33) [20] con la cual se realiza un análisis similar al anterior(Ver anexo 8) En este proceso el resultado gráfico se obtiene de la siguiente forma(Ver anexo 9). y la gráfica que se ve a continuación permite apreciar ciertas similitudes entre los

resultados del algoritmo de cluster kmean y el realizado por el factominer.

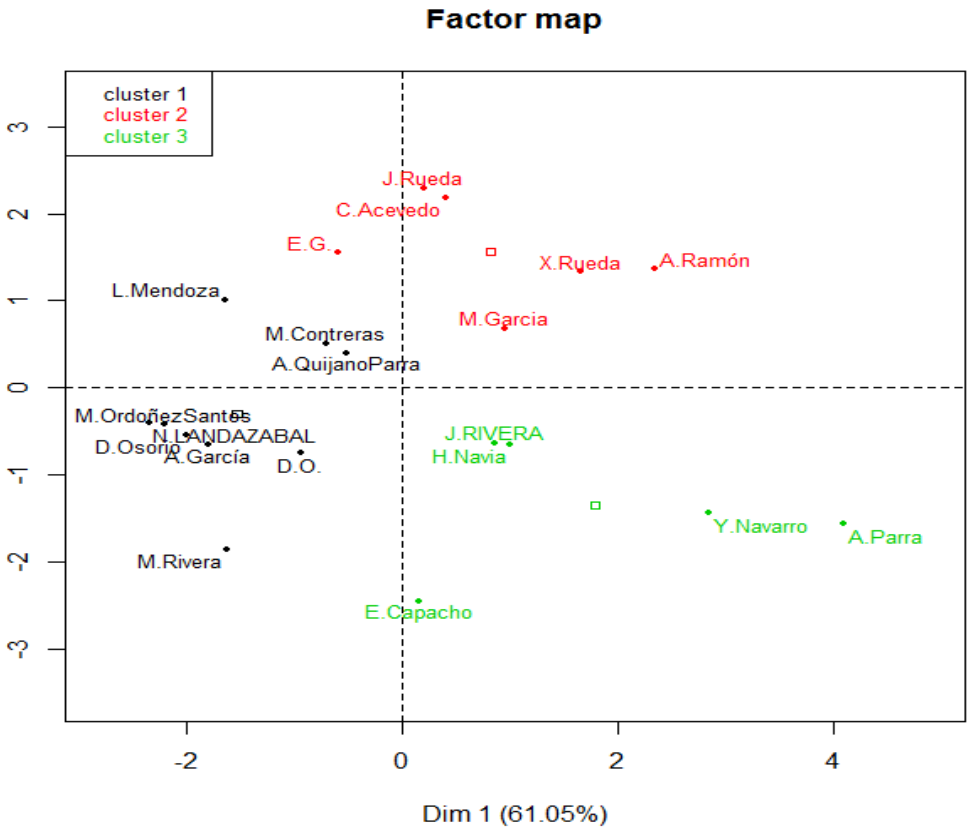


FIGURA 4.20: Gráfica de cluster por autor con Factominer

En la siguiente tabla se muestran las variables asociadas a los tres clusters definidos por el Factominer y su análisis de componentes principales. El valor V.Test muestra la asociación directa o inversa, según su signo. El valor p señala el error estadístico de la asociación de cada variable (todas p¡0.01)

FIGURA 4.21: Tabla de variables asociadas a los tres clusters

Agrupamiento	Variable Asociada	V. Test	Promedio en Categoría	Promedio Total	Valor de Probabilidad
1	Rank.by.DF	15.5555556	3.456804	10.5000000	0.0005466220
	First.Authored	-3.095858	1.6666667	3.9000000	0.0019624447
	Dominance.Factor	-3.317894	0.2752646	0.4991774	0.0009069894
2	Dominance.Factor	3.175054	0.7952381	0.4991774	0.001498087
	Rank.by.DF	-3.216666	4.0000000	10.5000000	0.001296896
3	Multi.Authored	3.729317	12.2	7.65	0.0001919995
	First.Authored	2.565901	6.8	3.90	0.0102908170
	Rank.by.Articles	-3.273268	3.0	10.50	0.0010631149

Técnicas de asociación

Para la implementación de las reglas de asociación las cuales buscan encontrar relaciones entre las variables de un conjunto de datos, para este caso se aplicó este algoritmo a los campos que tiene que ver con los autores y los años en que publicaron sus trabajos. El paquete que permitió aplicar el algoritmo de asociación es llamado Rattle, en el cual es posible cargar un conjunto de datos para posteriormente analizar mediante las diferentes técnicas de minería de datos que esta herramienta posee. Para este caso se escogen las variables que representan los autores y el año en que se realizó la publicación (ver figura 4.22).

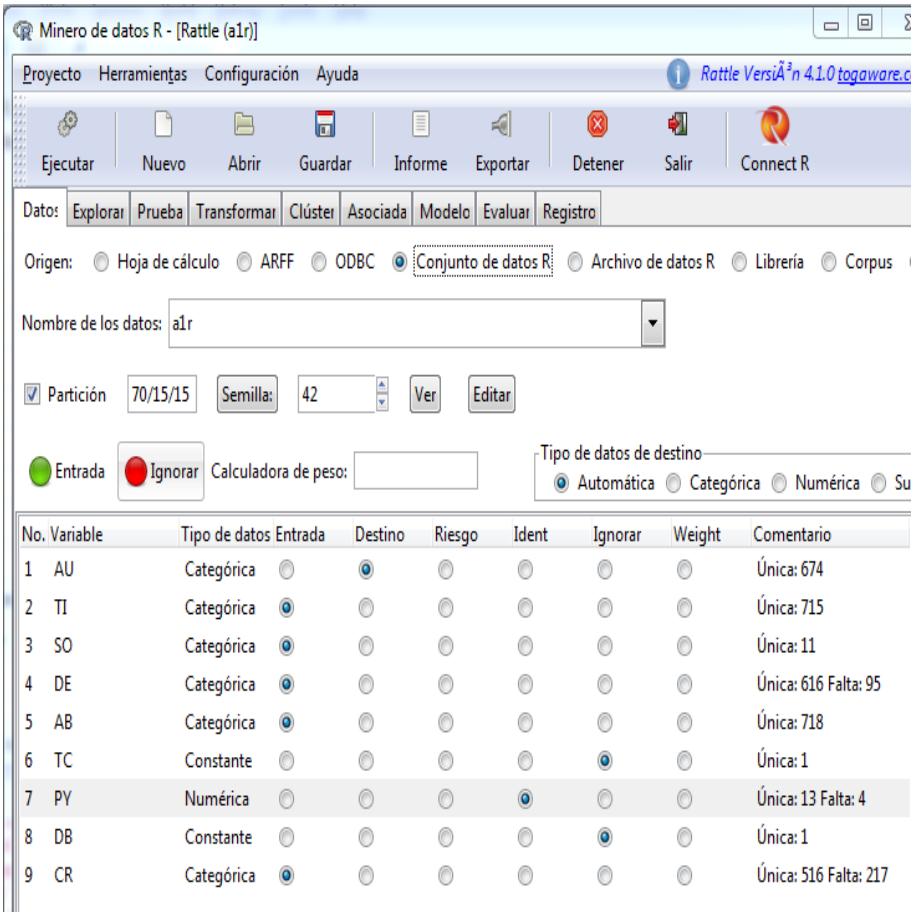
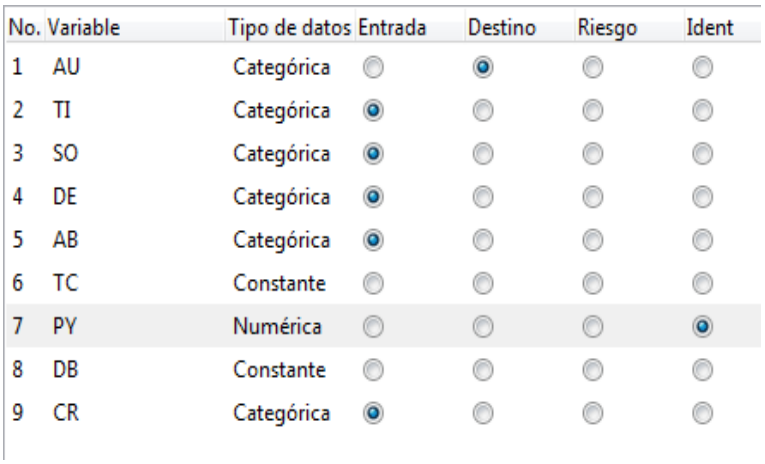


FIGURA 4.22: Configuraciones para técnicas de asociación

Se debe tener en cuenta las opciones de destino e Ident que son importantes para configurar las variables que se van a estudiar (Figura 4.23).



No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident
1	AU	Categórica	○	●	○	○
2	TI	Categórica	●	○	○	○
3	SO	Categórica	●	○	○	○
4	DE	Categórica	●	○	○	○
5	AB	Categórica	●	○	○	○
6	TC	Constante	○	○	○	○
7	PY	Numérica	○	○	○	●
8	DB	Constante	○	○	○	○
9	CR	Categórica	●	○	○	○

FIGURA 4.23: Configuraciones para técnicas de asociación

Con estas configuraciones se debe ir a la pestaña Asociada, que es la que permitirá realzar el proceso de aplicación del algoritmo Aprioro y generar las reglas que muestran las asociaciones entre autores y años de publicaciones. Se debe seleccionar la casilla cesta para finalizar las configuraciones y proceder a generar las reglas de asociación mediante el boton ejecutar(ver figura 4.24).

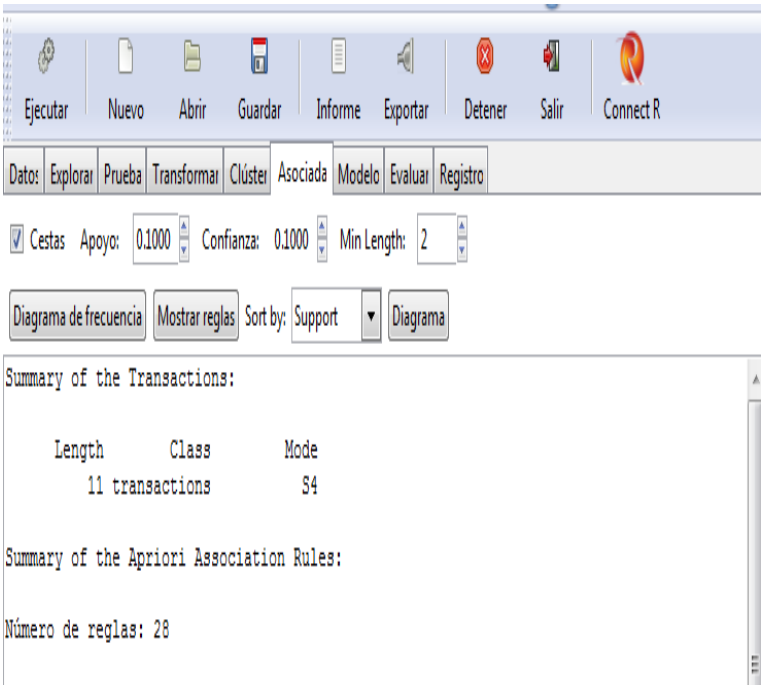


FIGURA 4.24: Configuraciones para técnicas de asociación

Para ver en detalle las reglas y como se asocian los autores según el año de publicación, se selecciona el boton Mostrar reglas que genera en detalle las asociaciones que se generan entre autores(Figura 4.25).

All Rules

	lhs		rhs	support	confidence	lift
[1]	{M.Garcia}	=>	{E.Capacho}	0.1818182	1	5.5
[2]	{E.Capacho}	=>	{M.Garcia}	0.1818182	1	5.5
[3]	{M.Garcia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[4]	{N.Chiavaro}	=>	{M.Garcia}	0.1818182	1	5.5
[5]	{M.Garcia}	=>	{H.Navia}	0.1818182	1	5.5
[6]	{H.Navia}	=>	{M.Garcia}	0.1818182	1	5.5
[7]	{E.Capacho}	=>	{N.Chiavaro}	0.1818182	1	5.5
[8]	{N.Chiavaro}	=>	{E.Capacho}	0.1818182	1	5.5
[9]	{E.Capacho}	=>	{H.Navia}	0.1818182	1	5.5
[10]	{H.Navia}	=>	{E.Capacho}	0.1818182	1	5.5
[11]	{N.Chiavaro}	=>	{H.Navia}	0.1818182	1	5.5
[12]	{H.Navia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[13]	{E.Capacho,M.Garcia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[14]	{M.Garcia,N.Chiavaro}	=>	{E.Capacho}	0.1818182	1	5.5
[15]	{E.Capacho,N.Chiavaro}	=>	{M.Garcia}	0.1818182	1	5.5
[16]	{E.Capacho,M.Garcia}	=>	{H.Navia}	0.1818182	1	5.5
[17]	{H.Navia,M.Garcia}	=>	{E.Capacho}	0.1818182	1	5.5
[18]	{E.Capacho,H.Navia}	=>	{M.Garcia}	0.1818182	1	5.5
[19]	{M.Garcia,N.Chiavaro}	=>	{H.Navia}	0.1818182	1	5.5
[20]	{H.Navia,M.Garcia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[21]	{H.Navia,N.Chiavaro}	=>	{M.Garcia}	0.1818182	1	5.5
[22]	{E.Capacho,N.Chiavaro}	=>	{H.Navia}	0.1818182	1	5.5
[23]	{E.Capacho,H.Navia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[24]	{H.Navia,N.Chiavaro}	=>	{E.Capacho}	0.1818182	1	5.5
[25]	{E.Capacho,M.Garcia,N.Chiavaro}	=>	{H.Navia}	0.1818182	1	5.5
[26]	{E.Capacho,H.Navia,M.Garcia}	=>	{N.Chiavaro}	0.1818182	1	5.5
[27]	{H.Navia,M.Garcia,N.Chiavaro}	=>	{E.Capacho}	0.1818182	1	5.5
[28]	{E.Capacho,H.Navia,N.Chiavaro}	=>	{M.Garcia}	0.1818182	1	5.5

FIGURA 4.25: Configuraciones para técnicas de asociación

Capítulo 5

CONCLUSIONES

Al validar el procedimiento propuesto en este trabajo se puede concluir que:

- Gracias al acceso a la información facilitada por la vicerrectoría de investigación de la universidad de pamplona se pudo realizar este estudio de manera confiable con información real del comportamiento de los diferentes actores inmersos en la producción científica de esta universidad.
- La definición de un procedimiento de minería de datos para el análisis cienciomentro al OJS es totalmente factible ya permite definir un camino seguro y detallado en la extracción de nuevo conocimiento de su base de datos.
- La verificación del procedimiento pudo genera indicadores que corresponden a autores y su producción científica, producción anual, revistas con mas artículos, que dan cuenta del estado en que se encuentra la producción científica.
- Gracias a la revisión bibliográfica se pudo establecer las técnicas de minería de datos mas aptas para el procesamiento de la información.
- Mediante los resultados obtenidos se puede establecer como se comportan y que tipo de relaciones existen entre los elementos que componen la información del OJS
- Con el desarrollo de las etapas planteada se puede establecer exitosamente la validez del procedimiento establecido y su plena funcionalidad.

Capítulo 6

RECOMENDACIONES Y TRABAJO FUTURO

- Consolidar una línea de investigación entre el programa de ingeniera de sistemas y la vicerrectoría de investigación para fomentar el estudio de el sistema OJS en la universidad de pamplona.
- Realizar estudios mas profundos para ver que otras técnicas de minería de datos se pueden aplicar a este sistema y su estructura de datos.
- Detectar y vincular otras comunidades académicas que trabajen con el sistema OJS y que deseen implementar el procedimiento planteado.
- Teniendo en cuenta que el OJS posee funciones de revisión de contenidos, manejo de usuarios y edición de la información, se debe contar con la debida autorización de los propietarios de la información para evitar inconvenientes por la delicadeza de estos contenidos.

Capítulo 7

ANEXOS

7.1. Anexo 1:

- Comandos para generar el Análisis de los datos.

```
library(bibliometrix)

results <- biblioAnalysis(array_ojs)

summary(results)
plot(results, k=10,pause=FALSE)
```

FIGURA 7.1: Implementación analisís bibliométrico

7.2. Anexo 2:

- Comandos para generar la conexón desde R a la base de datos en MYSQL.

```
library(DBI)
library(RMySQL)

con <- dbConnect(MySQL(),
                 user="root", password="[illegible]",
                 dbname="ojs", host="localhost")
on.exit(dbDisconnect(con))
```

FIGURA 7.2: Conexión a la base de datos desde R a mysql

7.3. Anexo 3:

- Consulta SQL para construir el conjunto de datos seleccionado.

```
res<- dbSendQuery(con, "

SELECT a.article_id,GROUP_CONCAT(DISTINCT(concat(substring(aa.first_name,1,1),'.',aa.last_name))SEPARATOR ';')
ase.setting_value as TI, j.path as SO, GROUP_CONCAT(DISTINCT(kl.keyword_text)SEPARATOR ';') as DE,

                GROUP_CONCAT(DISTINCT(concat('[',aa.first_name,']',aus.setting_value,aa.country))SEPARATOR '

                FROM
                articles a LEFT JOIN authors aa ON (aa.submission_id = a.article_id)
                left join article_settings ase ON (ase.article_id=a.article_id and ase.setting_name='title')
                left join journals j ON (a.journal_id=j.journal_id)
                LEFT join `article_search_objects` as so ON a.article_id=so.article_id and so.type=2
                LEFT JOIN article_search_object_keywords as ok ON ok.object_id=so.object_id
                LEFT JOIN article_search_keyword_list as kl ON kl.keyword_id=ok.keyword_id
                left join author_settings as aus on a.article_id=aus.author_id and aus.setting_name='affilia
                INNER join published_articles as pa on pa.article_id=a.article_id GROUP by a.article_id
                ")

array_ojs <- fetch(res)
```

Figura 7.3: Consulta construcción estructura de datos

7.4. Anexo 4:

- Comando para generar la red de colaboración.

```
NetMatrix <- biblioNetwork(array_ojs, analysis = "collaboration",
                           network = "authors", sep = ";")
netDegree <- 2
diag <- Matrix::diag
NetMatrix <- NetMatrix[diag(NetMatrix) >= netDegree,diag(NetMatrix) >= netDegree]
diag(NetMatrix) <- 0
bsk.network <- graph.adjacency(NetMatrix,mode = "undirected")
plot(bsk.network,layout = layout.fruchterman.reingold, vertex.label.dist = 0.2,
     vertex.frame.color = 'blue', vertex.label.color = 'black',
     vertex.label.font = 1, vertex.label = V(bsk.network)$name, vertex.label.cex = 0.5,vertex.size=5)
```

Figura 7.4: Función de redes de colaboración

7.5. Anexo 5:

- Comando para generar la red de cocitación.

```
NetMatrix <- biblioNetwork(array_ojs, analysis = "co-citation",
                           network = "references",sep = ";")
netDegree=2
diag <- Matrix::diag
NetMatrix <- NetMatrix[diag(NetMatrix) >= netDegree,diag(NetMatrix) >= netDegree]
diag(NetMatrix) <- 0
bsk.network <- graph.adjacency(NetMatrix,mode = "undirected")
plot(bsk.network,layout = layout.fruchterman.reingold, vertex.label.dist = 0.5,
     vertex.frame.color = 'blue', vertex.label.color = 'black',
     vertex.label.font = 0, vertex.label = V(bsk.network)$name, vertex.label.cex = 0.7,vertex.size=5)
```

Figura 7.5: Función de redes de cocitación

7.6. Anexo 6:

- Sentencia que permite ordenar a los autores por su factor de dominancia.

```
DF=dominance(results,k=20)
DF
```

FIGURA 7.6: Función de dominancia

7.7. Anexo 7:

- Sentencia que permite aplicar el algoritmo de cluster Kmeans.

```
ModeloKMEANS <- kmeans(domi[-1],3)
domi$grupo <- ModeloKMEANS$cluster

clusplot(domi[-1], ModeloKMEANS$cluster, main='CLUSTERS DE AUTORES',
         color=TRUE, shade=TRUE,
         labels=2, lines=0)
```

FIGURA 7.7: Algoritmo Kmeans

7.8. Anexo 8:

- Sentencia que permite generar los cluster con factominer

```
res<-PCA(DF.PCA , scale.unit=TRUE, ncp=5, graph = FALSE)

res.hcpc<-HCPC(res ,nb.clust=-1,consol=TRUE,min=3,max=10,graph=TRUE)

res.hcpc$data.clust[,ncol(res.hcpc$data.clust),drop=F]
```

FIGURA 7.8: Cluster con factominer

7.9. Anexo 9:

- Sentencia que permite graficar los cluster con factominer

```
plot.PCA(res, axes=c(1, 2), choix="var", new.plot=TRUE, col.var="black",
  col.quanti.sup="blue", label=c("var", "quanti.sup"), lim.cos2.var=0,
  title="")
```

FIGURA 7.9: Cluster con factominer

Bibliografía

[1] Sandra Edith and M Claudia. II Reunión Latinoamericana de Análisis de Redes Sociales Minería de datos , bibliometría y Estrategia metodológica para identificar y Zaida Chinchilla-Rodríguez. 2009.

[2] Salvador Gorbea-Portal. Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y el conocimiento. *Perspectivas em Gestão & Conhecimento*, 3(1):13–27, 2013.

[3] Ricardo Herrera Varela. Bibliomining: minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. pages 1–29, 2006.

[4] Cesar pérez lópez. *Mineria de datos tecnicas y herramientas*. 2008.

[5] José Manuel Molina López. Técnicas Estadísticas de Análisis de Datos. 2006.

[6] José Molina and Jesús García. Técnicas de Minería de Datos basadas en Aprendizaje Automático. *Técnicas de Análisis de Datos*, pages 96 – 266, 2008.

[7] María Moreno, Luis Miguel, Francisco García, and José Martín. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. *Iiis.Org*, page 14, 2008. URL http://www.iiis.org/CDs2010/CD2010CSC/CISCI{_}2010/PapersPdf/CA156FK.pdf.

[8] Dra. María del Carmen Gómez Fuentes. *Bases de datos*. ISBN 8497882695.

[9] José D. Martín Guerrero, Emilio Soria, and Antonio J Serrano. T Ecnicas. *Técnicas De Agrupamiento*, page 11, 2010. URL http://ocw.uv.es/ingenieria-y-arquitectura/2/clustering.pdf.

[10] EcuRed Enciclopedia cubanad. 2016. URL https://www.ecured.cu/Clustering.

[11] EcuRed Enciclopedia cubanad. 2016. URL https://www.ecured.cu/Weka.

[12] José Ángel Gallardo San Salvador. Métodos Jerárquicos de Análisis Multivariante. 1994. URL http://www.ugr.es/{~}gallardo/pdf/cluster-3.pdf.

[13] Marcelo de la Puente. Bibliominería: bibiometría y minería de datos. *Documentos de trabajo*, (14):26, 2010. ISSN 1852-6411. URL http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

[14] J. Araújo and R Arencibia. Contribuciones cortas Informetría, bibliometría y cienciometría: aspectos teórico- prácticos. *Acimed*, pages 1–4, 2002. ISSN 10249435. doi: 10.1016/j.aprim.2011.12.002.

[15] Martinón A. Las matemáticas del siglo xx: una mirada en 101 artículos. 2000.

[16] Valderrama Zurian Gonzalez Alcaide.

[17] Carlos Olmeda-gómez, Antonio Perianes-rodríguez, Mª Antonia Ovalle-perandones, Grupo De, Scimago Universidad, Carlos Iii, and De Madrid Departamento. Madrid (1995-2003). pages 1–16, 2003.

[18] José A Gallardo. CRISP-DM Metodologia para el Desarrollo de Proyectos de Mineria de Datos.

[19] R Development Core Team. Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos. 1:100, 2000.

[20] Corrado Cuccurullo Massimo Aria. Package bibliometrix . 2016.

[21] Small and Sweeny. Las matemáticas del siglo xx: una mirada en 101 artículos. 200.

[22] Wickham Hadley. Packege DBI. 2016. URL http://rstats-db.github.io/DBI.

[23] Hadley Wickham David James, Saikat DebRoy. Packege RMYSQL. 2016. URL https://github.com/rstats-db/rmysql.

[24] Graham Williams. The Rattle Package : Quick Start Guide. pages 2–3, 2011.

[25] John Willinsky, Kevin Stranack, Alec Smecher, James Macgregor, and Atenea Acevedo. Open Journal Systems: Una guía completa para la edición de publicaciones en línea. 2010.

[26] Oracle. 2016. URL https://www.mysql.com/.

[27] Http server proyect. 2016. URL https://httpd.apache.org/download.cgi.

[28] php. 2016. URL http://php.net/.

[29] Segio Luis Perez. Minería de datos (Reglas de asociací on arboles de decisí on). pages 1–26.

[30] elsevier.com. 2016. URL http://www.americalatina.elsevier.com/corporate/es/scopus.php.

[31] thomsonreuters. 2016. URL http://thomsonreuters.com/en.html.

[32] Title Network Analysis and Imports Matrix. Igraph. 2015. doi: 10.1177/001316446902900315.