Instituto Superior Minero metalúrgico de Moa
“Dr. Antonio Núñez Jiménez”

Departamento de Informática

Aplicación web para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por Open Journal System

Trabajo de diploma para optar por el título de Ingeniería en Informática

Autora: Nerina Peña Olivero

Tutor: Ing. Miguel Barrera Fernández

DECLARACIÓN DE AUTORÍA

Declaro que soy el único autor de este trabajo y autorizo al Departamento de Informática del Instituto Superior Minero metalúrgico de Moa para que hagan el uso que estimen pertinente con este trabajo.

Para que así conste firmamos la presente a los _____ días del mes de _____ del _____.

_____
Nerina Peña Olivero

_____
Ing. Miguel Barrera Fernández

Agradecimientos

Muchas personas me han apoyado en este largo recorrido, muchos me han ayudado a levantarme cuando me he caído y me han animado a seguir adelante pese a las dificultades.

Quiero agradecer a mi tutor por la excelente labor desarrollada en la dirección de esta tesis.

A mis profesores por todos estos años de paciencia y apoyo que sin duda me han servido para crecer profesionalmente, y que sin su ayuda no hubiese sido capaz de realizar.

A mis amigos y compañeros, que me apoyaron y me dieron su amistad durante este largo período.

A mi familia por su constante interés, apoyo moral y amor.

A J.Luis por su apoyo incondicional y por tanta paciencia durante este largo período.

A todos solo puedo decir infinitamente GRACIAS.

Dedicatoria

A Pipo, donde quiera que estés estoy segura que estás disfrutando este momento tanto como yo. Gracias por ser el promotor de esta idea y por tantos años de sacrificio y amor. Ya la deuda está cumplida!

Resumen

La identificación de posibles evaluadores en una revista científica podría decirse que es uno de sus principales procesos de gestión. El Open Journal System por defecto tiene implementado un sistema de búsqueda, pero se pudo constatar que para este propósito contiene limitaciones: la búsqueda está concebida principalmente en la obtención de documentos por una necesidad de información dada y no en la identificación de posibles revisores o expertos para una temática determinada y la no existencia de cierto mecanismo u opción que permita realizar un agrupamiento de usuarios del sistema. La presente investigación se realizó con el objetivo de desarrollar una aplicación web que permita la conformación y agrupamiento de perfiles de usuario para la identificación de posibles árbitros en revistas gestionadas por el Open Journal System de manera que ayude a los consejos editoriales a procesar el gran volumen de información que contienen sus bases de datos.

En la presente investigación, se describe el proceso de desarrollo de la aplicación web, para tal propósito se utilizó la metodología para proyectos de minería de datos: CRISP-DM, así como el procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por el Open Journal System.

A partir de la implementación de la aplicación se muestran resultados experimentales en cuanto a: la evaluación del grado de ajuste de una clasificación a un conjunto de datos y la calidad de la clasificación, obteniéndose resultados satisfactorios.

Abstract

Identifying potential evaluators in a scientific journal is arguably one of its key management processes. The Open Journal System default has implemented a search, but it was found that for this purpose contains limitations: Search is designed primarily in obtaining documents by a need for information given and not in identifying possible reviewers or experts for a given subject and the absence of some mechanism or option that allows for a grouping of system users.

This research was conducted with the aim of developing a web application that allows the creation and grouping of user profiles to identify potential arbitrators managed by the Open Journal System magazines so that helps the editorial boards to process the large volume of information contained in their databases.

In this investigation, the development process of the web application is described, for this purpose the methodology for mining projects data was used: CRISP-DM and the procedure for the formation and grouping of user profiles in scientific journals managed by the Open Journal System.
Since the implementation of the application experimental results are shown in terms of: evaluating the degree of adjustment of a classification to a set of data and the quality of the classification, obtaining satisfactory results.

Índice

Índice de tablas y gráficas

Índice de figuras

Introducción

La información hoy en día es un tesoro muy valioso tanto para empresas, organizaciones o para simples usuarios, ya que para todos lo más importante es obtener una información oportuna y de buena calidad. A pesar de esto, el constante avance tecnológico de la Informática y las Telecomunicaciones, el desarrollo de sistemas gestores de bases de datos más poderosos y la acumulación rápida de datos, han llevado a un aumento de la información, donde Internet constituye una de las principales fuentes de extracción de información. Este veloz crecimiento de la información trae consigo que se requiera la selección de un mayor número de expertos que certifiquen la validez de la información y la necesidad de procesar un gran volumen de esta.

Las revistas científicas encontraron en Internet un camino para llegar a un mayor número de lectores. Actualmente una gran cantidad de revistas científicas se gestionan sobre sistemas informáticos, que permiten el envío en línea de los artículos y la selección de los evaluadores que avalarán su calidad para que podamos apropiarnos del conocimiento contenido en dicha información.

Gran parte de este conocimiento existe en forma de lenguaje natural, y pesar que la especie humana posee habilidades extremadamentes sofisticadas para detectar patrones y descubrir tendencias, es evidente que no podemos realizar la tarea de analizar el gran volumen de datos que se almacena electrónicamente, por ejemplo, en las bases de datos de revistas científicas.

Para la gestión de revistas científicas entre los sistemas informáticos más utilizados se encuentran el Open Journal Systems (OJS), el Sistema Electrónico de Gestión Editorial (SEGE) y el Quark Publishing System 7 (QPS 7) (Barrera-Fernández, 2011).

Existe una tendencia a la utilización del OJS para la gestión y publicación de las revistas en formato digital. Lo que trae como consecuencia que se recolecten un gran conjunto de datos, principalmente de carácter textual que pueden ser procesados computacionalmente para ayudar a los consejos editoriales en algunos de los procesos de gestión que se llevan a cabo en una revista. Para este propósito la minería de textos es especialmente apropiada

La minería de textos pretende identificar relaciones y modelos en la información no estructurada, así como proveer de una visión selectiva y perfeccionada de la información contenida en documentos escritos y sacar consecuencias para la acción, detectar patrones no triviales e incluso, información sobre el conocimiento almacenado en las mismas (Tan, 1999).

Para lograr sus propósitos, la minería de textos necesita combinar varias técnicas, de ahí que sea un campo multidisciplinario que incluye el análisis de textos, la extracción de información, el agrupamiento, la construcción de resúmenes, la clasificación, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos (Tan, 1999) (Dixon, 1997).

La identificación de posibles evaluadores en una revista científica podría decirse que es uno de sus principales procesos de gestión. El OJS por defecto tiene implementado un sistema de búsqueda, por medio del cual de cierta

forma se contribuye a la identificación de posibles evaluadores. Pero se pudo constatar que para este propósito contiene las siguientes limitaciones:

- La búsqueda en este sistema está concebida principalmente en la obtención de documentos por una necesidad de información dada y no en la identificación de posibles revisores o expertos para una temática determinada.
- No existencia de cierto mecanismo u opción que permita realizar un agrupamiento de usuarios del sistema.

Por las cuestiones anteriormente tratadas, por la cantidad de información almacenada en las bases de datos de las revistas gestionadas con OJS y la poca existencia de mecanismos u opciones en el sistema para procesar estos datos, se dificulta la identificación de los autores de artículos de la propia plataforma que podrían servir cómo posibles evaluadores de un artículo determinado, lo que genera en ocasiones que el proceso de selección de evaluadores sea un poco lento, que se desconozcan algunos de los posibles candidatos y que de conocerse no se sepa con claridad la estructura jerárquica de estos respecto al tema del artículo científico a evaluar.

En función de esta situación se hace necesario crear una aplicación informática que permita conocer la similitud existente entre los perfiles de usuarios y la clasificación de estos, generados de la información contenida en una revista gestionada con OJS, con la finalidad de favorecer el proceso de identificación de posibles árbitros de artículos científicos. Para tal propósito se utilizará el procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por el Open Journal System, el cual es una propuesta del Ing. Miguel Barrera Fernández.

Los elementos anteriores dan lugar, a la formulación del siguiente **problema científico**:

¿Cómo favorecer el proceso de identificación de investigadores similares que publican en revistas gestionadas con el OJS para la ayuda en el proceso de selección de árbitros?

El **objeto de estudio** de esta investigación radica en:

La minería de textos en el Open Journal System.

El **Campo de acción:**
Informatización del proceso de Clasificación y agrupamiento jerárquico de perfiles de usuarios en revistas gestionadas con el Open Journal System.

Para dar solución al problema científico definido se ha formulado el **objetivo general** siguiente:

Desarrollar una aplicación Web que permita la conformación y agrupamiento de perfiles de usuario para la identificación de posibles árbitros en revistas gestionadas con el OJS.

Para dar cumplimiento al objetivo y resolver la situación problémica planteada, proponemos las siguientes **tareas de investigación**:

- Establecer las bases teóricas que permitan dar solución al problema científico planteado, así como el estudio de los antecedentes del sistema.
- Seleccionar la metodología, tecnologías y herramientas para guiar el proceso de desarrollo del proyecto.

- Aplicar la metodología escogida, así como el procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas con el OJS.

Idea a defender: La utilización de la aplicación web para la selección de posibles evaluadores de artículos científicos en revistas gestionadas con el OJS facilitará la agilización de este proceso.

Los **Métodos Teóricos y Métodos Empíricos** para la investigación científica que se utilizaron se describen a continuación:

Como **Métodos teóricos** se utilizaron:

Análisis y Síntesis: este método se utiliza para desglosar el problema en partes o subproblemas, para de esta forma comprobar el funcionamiento de los mismos, luego integrarlos para corroborar las relaciones entre estas partes, y su integración con un todo llegando así a una mejor solución, también para arribar a conclusiones generales de la investigación.

Histórico – Lógico: para la búsqueda de antecedentes del software y las herramientas utilizadas.

Como **Métodos Empíricos** se utilizaron:

Análisis de documentos: para elaborar los fundamentos teóricos que se relacionan con el campo de acción.

Entrevistas: para determinar los requerimientos funcionales del sistema Informático que se quiere construir. Se llevó a cabo un diálogo con personas expertas en la materia.

El trabajo está estructurado en 3 capítulos.

Capítulo 1. Fundamentación Teórica: Contiene la fundamentación teórica del tema, se aborda el lenguaje de programación y las tecnologías que se utilizan en el desarrollo de la aplicación. También se exploran soluciones existentes similares al campo de acción para tener una guía de las posibles automatizaciones que se pueden realizar.

Capítulo 2. Descripción de la solución propuesta: Este capítulo estará enfocado a la solución del problema en cuestión, en el mismo se realizará una descripción de la propuesta aplicando la metodología para proyectos de minería de datos CRISP-DM. En su última parte el capítulo estará enfocado a la implementación del sistema informático en cuestión, de manera que quede cubierto el objetivo de la investigación así como la Fase de Implementación de la metodología escogida.

Capítulo 3.Estudio de Factibilidad: En este capítulo se realiza un estudio para determinar la factibilidad del proyecto, basado en la metodología Coste-Beneficio. Además de un estudio de los esfuerzos requeridos para la realización del sistema propuesto.

Capítulo I. Fundamentación Teórica.

Introducción.

En el presente capítulo se describe de forma general los aspectos relacionados con el objeto de estudio y el campo de acción en que se trabaja. Este capítulo constituye la base teórica para la comprensión del trabajo que se desarrolla y sus principales aspectos.

1.1 Minería de datos

El origen de la minería de datos se relaciona con dos factores. Por una parte, la disponibilidad de grandes cantidades de datos almacenados electrónicamente; y por otra parte, la necesidad de transformar toda esta información en conocimiento útil para la toma de decisiones en diferentes escenarios de aplicación. Así pues, la minería de datos es el resultado "natural" de la evolución de los sistemas de información (Jeria,2007).

La minería de datos, es la etapa central del proceso de descubrimiento de conocimiento en bases de datos. En ella se realizan varias tareas que permiten identificar distintos tipos de patrones en un conjunto de datos. En general, estas tareas son de dos tipos: descriptivas y predictivas. Las tareas descriptivas caracterizan las propiedades generales de los datos y construyen descripciones compactas de estos. Por su parte, las tareas predictivas hacen inferencias sobre los datos conocidos con el objetivo de predecir el comportamiento de datos nuevos.

1.1.2 Minería de texto

La minería de datos se enfoca en el análisis de grandes bases de datos. Debido a ello, sus métodos consideran solamente información estructurada, principalmente numérica y booleana, y descuidan otros tipos de información. Como consecuencia de esta situación, muchos logros de la minería de datos parecen tareas muy difíciles de realizar con datos no-estructurados o semiestructurados. De ahí, el surgimiento de la minería de texto como una respuesta a la incapacidad de los métodos de minería de datos para analizar información textual.

La minería de texto se define, parafraseando la minería de datos, como el proceso de descubrimiento de patrones interesantes –y posiblemente nuevos conocimientos– en un conjunto de textos (Jeria, 2007). La idea es que estos patrones no deben existir explícitamente en ningún texto de la colección, y deben surgir de relacionar el contenido de varios de ellos (Jeria, 2007).

La minería de texto es también un proceso multidisciplinario que conjuga métodos provenientes de distintas áreas. Por ejemplo, en la etapa de preprocesamiento se emplean algunos métodos provenientes principalmente de la recuperación de información, mientras que en la etapa de descubrimiento se usan varios métodos de la minería de datos. Estos últimos son en su mayoría de tipo estadístico, aunque también algunos incorporan técnicas provenientes del aprendizaje automático.

1.1.3 Definición de Perfiles de Usuario de las TIC

Para Samper (2005) perfil es una palabra que procede de la expresión latina pro filare, que significa diseñar los contornos. Un perfil será un modelo de un

objeto, una representación compacta que describe sus características más importantes, que puede ser creado en la memoria de un ordenador y puede utilizarse como representante del objeto en las tareas computacionales. Las aplicaciones más conocidas que crean y gestionan perfiles incluyen la personalización, la gestión de conocimiento y el análisis de datos.

Se reconoce también la procedencia de perfil, derivada de la psicología, dentro de esta disciplina es entendido como el conjunto de medidas diferentes de una persona o grupo, cada una de las cuales se expresa en la misma unidad de medición. Esto es, que ciertas características de un individuo son medidas mediante pruebas que arrojan puntuaciones diferentes, estas puntuaciones constituyen su perfil, el cual es utilizado con fines diagnósticos (Corti, 2000).

Atendiendo el anterior planteamiento se puede entender el perfil del usuario como el conjunto de rasgos distintivos que lo caracterizan.
Para Samper (2005) existen distintos tipos de perfiles, desde el perfil psicológico del comportamiento de un individuo, hasta el perfil del funcionamiento de un programa de ordenador. En principio, se puede hacer un perfil de todo, y por consiguiente, las características representadas en el perfil dependerán de la naturaleza del objeto modelado.

1.1.4 Métodos para la creación de perfiles

Según Samper (2005) pueden considerarse tres métodos principales para crear perfiles: el método explícito o manual; el método colaborativo o de composición

a partir de otros perfiles, y el método implícito, que utiliza técnicas específicas para extraer las características automáticamente. Este autor afirma que:

- Método explícito los datos serán introducidos directamente por el usuario, escribiéndolos en su perfil de usuario o respondiendo a formularios.

- Método colaborativo se podrá crear y modificar un perfil de usuario a partir de su interacción colaborativa con otros perfiles con los que se relaciona, recurriendo a conocimiento específico del dominio y heurísticas inteligentes.

- Método implícito, los perfiles de usuario se crearán y se modificarán automáticamente, recurriendo en la mayoría de los casos a técnicas de Inteligencia Artificial.

Como se puede observar en la figura 1.1 el perfil se construye a partir de las características que identifican y caracterizan a un usuario de otro y de los factores de influencia que lo circundan (Naranjo y Álvarez, 2003).

Cada usuario tiene sus propios intereses y necesidades, de acuerdo con su desarrollo cognoscitivo, del ambiente en que se desenvuelve y de su experiencia de vida, lo cual los hacen únicos, de los perfiles de usuarios pueden derivarse innumerables estudios, que permitan determinar el nivel de interacción entre ellos, la experticia en dependencia de los campos recogidos en su perfil, la compatibilidad a nivel de similitud o distancia entre ellos, conglomerados de usuarios respondiendo a los parámetros definidos en su perfil, etc.

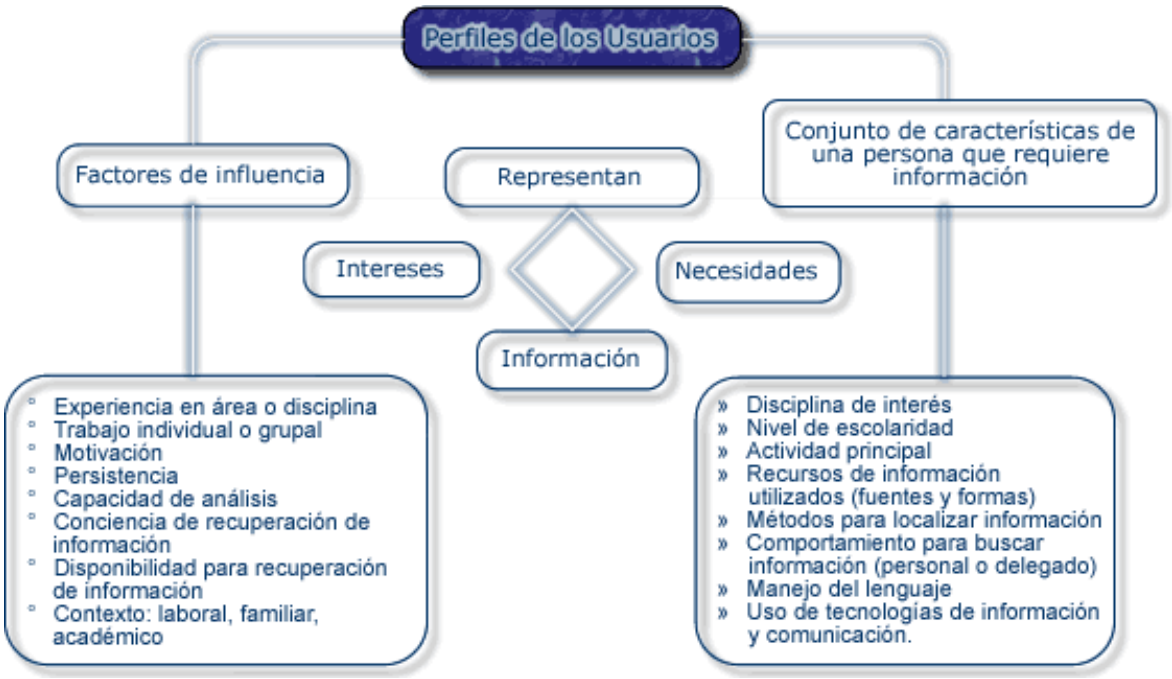


Figura 1.1. Perfiles de Usuario. Fuente: (Naranjo y Álvarez, 2003).

En este acápite no se pretende conceptualizar las TIC y su desarrollo en las organizaciones, sino plasmar una coyuntura cultural acerca del empleo de estas tecnologías. Como se ha podido observar se ha ido mencionando en cada proceso de datos, información, conocimiento e inteligencia y el empleo o integración de las TIC en actividades que justifican su uso en cada proceso como herramienta de apoyo.

En la presente investigación, emplearemos para la generación de los perfiles de usuarios los datos obtenidos mediante el método explícito, que es el que se

utiliza para que el perfil sea construido a partir de la información suministrada por el usuario en el momento en el que este interactúa con el sistema OJS.

1.1.5 Clasificación y Agrupamiento de Perfiles de Usuario

Podríamos decir que en el trabajo con de perfiles de usuarios, un problema de clasificación surge cuando se quiere decidir si un perfil de usuario pertenece a una categoría preestablecida de perfiles de usuarios.

Para la clasificación de perfiles de usuario se utiliza la técnica de Análisis de Clusters (o Análisis de conglomerados), es una técnica de Análisis Exploratorio de Datos para resolver problemas de clasificación. Su objetivo consiste en ordenar objetos (personas, cosas, animales, plantas, variables, etc,…) en grupos (conglomerados o clusters) de forma que el grado de asociación/similitud entre miembros del mismo cluster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters. Cada cluster se describe como la clase a la que sus miembros pertenecen.

Un clúster se define como un conglomerado de objetos que comparten muchas características entre sí, o dicho de otra manera, aquel en que los perfiles de los objetos en un mismo grupo sean muy similares entre sí, pero son muy disimilares a aquellas en los objetos que pertenecen a otros clústeres (Karlson, 2008; Romesburg, 2004; Mooi y Sardtedt, 2011). Un análisis de clústeres se define como la partición de las observaciones en grupos de manera que las disimilitudes por parejas (medida de cuán diferentes son dos elementos) entre

los elementos asignados a un clúster sean menores con respecto a elementos pertenecientes a otros clústeres (Hastie et al., 2009).

Los seres humanos poseen una capacidad innata para formar clústeres, estableciendo categorías para todos los elementos que le rodean y luego ubicando dentro de cada una de esas categorías cada nuevo elemento que se visualice (clasificación). La técnica de clústering es una de las técnicas más utilizadas para análisis y exploración de datos. Esta técnica tiene aplicaciones en estadística, ciencias de la computación, biología, ciencias sociales y psicología (Von-Luxburg, 2007).

El análisis de clúster puede ser clasificado de acuerdo al resultado que el mismo genera. Una primera clasificación crea, por un lado, una jerarquía de clústeres (clústering jerárquico) y por otro lado una partición en clústeres (partitioning clustering), (Arabie y Hubert, 1996). En el primer caso, los clústeres son formados por anidación o por des-anidado de los elementos. En el caso de des-anidación, se parte de un clúster global que contiene todas las observaciones en un sólo clúster y en función del algoritmo de desagrupamiento que se emplea, se van formando subclústeres hasta llegar a un número de clústeres igual al número de elementos (cada elemento constituye un clúster o "singleton").

En el caso de anidación sucede exactamente lo contrario, es decir, desde los "singleton" se llega a un clúster global que agrupa todos los elementos. El segundo caso (partición de clústeres), es una partición simple del conjunto de

datos en subconjuntos disjuntos, de tal manera que cada elemento se encuentre en uno u otro subconjunto.

Otra clasificación del análisis de clústeres genera como resultado clústeres exclusivos, o clústeres en los cuales cada clúster tiene su subconjunto exclusivo de elementos, que no se repiten en otro(s) clúster, clústeres no disjuntos, que corresponde al caso en que un mismo elemento(s) puede existir en diferentes clústeres al mismo tiempo; clúster borroso (fuzzy cluster) en el que los elementos no pertenecen per se a un determinado clúster, sino que su pertenencia al conjunto de clústeres está asociado a un peso, siendo 1 el peso que indica que el elemento pertenece completamente a un clúster dado y 0 el peso que indica que el elemento no tiene ninguna relación de pertenencia con un clúster dado (Abony y Balázs, 2007). La última clasificación de análisis de clúster de la que se hará mención puede generar clústeres completos o clústeres parciales. En el caso de los clústeres completos, todos los elementos se agrupan dentro de algún subgrupo, en tanto en el caso de la partición parcial, ciertos elementos no pueden ser ubicados en ninguno de los subconjuntos resultantes. Estos elementos corresponden generalmente al ruido del conjunto de datos.

También existe una clasificación de los clúster generados en función de las metas que persiga el análisis de clúster. Así, estos pueden ser bien separados, cuando se basan en prototipo, basados en densidad, basados en grafos y en propiedades compartidas. Los clústeres bien separados siguen una conceptualización idealista del término clúster, es decir, los clústeres son formados por elementos con un grado de similitud muy fuerte entre sí, y se

encuentran lejos de otros elementos. En ciertos casos se emplean umbrales mínimos para definir si una similitud es suficiente para establecer el agrupamiento o no. Este tipo de clústeres también son conocidos como clústeres naturales. Los clústeres basados en prototipos, o también conocidos como clústeres basados en el centro (centered-based cluster) corresponden a clústeres que se forman siguiendo un prototipo como lo puede ser un centroide o un medoide, a partir del cual los elementos se agrupan (Ding et al., 2008). Los clústeres basados en grafos se forman a partir de nodos que se encuentran conectados entre sí a través de links (conexiones) o aristas, siguiendo la tipología propia de un grafo de acuerdo a la teoría de grafos, que será abordada en la siguiente sección. En este caso, el agrupamiento se centra en la conexión entre los elementos. Así, los elementos (o nodos) con un determinado grado de similaridad se encuentran conectados y los elementos (o nodos) con un nivel de disimilaridad se encuentran desconectados. En el caso de los clústeres basados en densidad, el agrupamiento se da por regiones de mayores densidades de elementos, siendo la separaciones entre dos clústeres zonas de baja densidad de elementos (Ding et al., 2008). Los clústeres basados en propiedades compartidas van un paso más allá de todos los tipos de clústeres mencionados previamente, estos incluyen clústeres que puedan estar enlazados entre sí mediante ciertos elementos en común.

En este caso centraremos el estudio en cuanto al agrupamiento de los usuarios por medio del clúster jerárquico ya que por medio de este no se requiere hacer inferencias sobre el número de clusters y se permite apreciar claramente las relaciones de agrupación entre los datos e incluso entre grupos de ellos. Observando las sucesivas subdivisiones podemos hacernos una idea sobre los

criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc.

El clústering jerárquico es una herramienta de análisis de datos que se basa en la construcción de clústeres de los mismos bajo un orden de jerarquía. Se agrupa dentro de los métodos de aprendizaje automático no supervisado, y tienen como objetivo global hacer una exploración de datos. La idea, tras este, es construir árboles binarios que sucesivamente se fusionen en grupos en dependencia de su similaridad (Han et al., 2006). A partir del estudio del árbol que se genere, se puede extraer información útil que ayude a comprender la estructura de los datos. Se diferencia de otros métodos de formación de clústeres en el hecho de que no se requiere introducir a priori un número de clústeres en el que se hará la partición (Manning et al., 2008), por el contrario, tal y como se explicará a continuación, el número de clústeres óptimo en el que se deben partir los datos puede ser estimado a través del árbol de jerarquía resultante. En el resto de los métodos de partición de datos en clústeres se asigna una "tarea" inicial (un centroide por ejemplo) a los clústeres, con respecto a la cual estos se terminan de conformar.

Existen dos tipos de procesos de clústering jerárquico, según el entorno desde el que se inicie éste. Si el proceso se inicia en un único clúster que contiene todos los casos agrupados como un sólo conjunto, se dice que se trata de un proceso de clústering divisivo (de arriba a abajo). En él, el clúster general se va subdividiendo hasta llegar a un nivel en donde cada una de los casos conforma un clúster (o singleton). El segundo procedimiento, conocido como clustrering de aglomeración en nido (*aglomerative nesting clustering*)

sigue la ruta contraria; en él, a partir de los singleton, los casos se van agrupando, hasta llegar a un nivel en donde se forma un único clúster (de abajo hacia arriba) (Cimiano et al., 2004). De estos dos métodos, el método aglomerativo suele ser empleado más comúnmente debido a que tiene menor grado de complejidad. El clústering jerárquico divisivo tiene la desventaja de que cuando el número de datos en los que se hace la partición es muy grande, es computacionalmente costoso examinar todas las posibles particiones. Por esta razón, se suele recurrir a métodos heurísticos, que luego puede conducir a imprecisiones en los resultados (Han et al., 2012).

En el proceso de aglomeración de casos en clústeres, en un primer momento, se producen los clústeres en función de la comparación entre las variables en cada caso individual y a continuación, de uno u otro elemento de los clústeres conformados (una vez que los clústeres dejen de ser singleton). El elemento característico de cada clúster que se emplee para hacer la comparación depende del método de aglomeración que se seleccione. Los métodos más comúnmente empleados son: agrupación por promedio, agrupación completa, agrupación individual, y finalmente, agrupación basada en un centroide (Everitt et al., 2011). A pesar de ello, existen otros métodos que se pueden emplear, tal como el conocido método de la mínima varianza (de Ward), el método de la mediana, el método de máxima probabilidad de igual varianza (o EML), el método de McQuitty y el método flexible-beta.

1.1.6 Métodos para identificar expertos

Existen varios métodos reconocidos para la identificación de expertos. Cuando se utiliza información de tipo subjetivo, hablaremos de métodos cualitativos; éstos suelen estar basados en opiniones y son los que durante mucho tiempo han surtido de expertos a numerosas actividades de investigación y desarrollo (I+D).

Para evitar las limitaciones de estos métodos surgen otros en los que se utiliza información objetiva basada en datos que faciliten la selección de los expertos más adecuados de entre una muestra de los disponibles en una determinada materia; éstos son los métodos cuantitativos. Entre los métodos cuantitativos y cualitativos estarían los métodos semi-cuantitativos donde el juicio subjetivo se intenta cuantificar mediante reglas o definiciones (Cabrera, 2015).

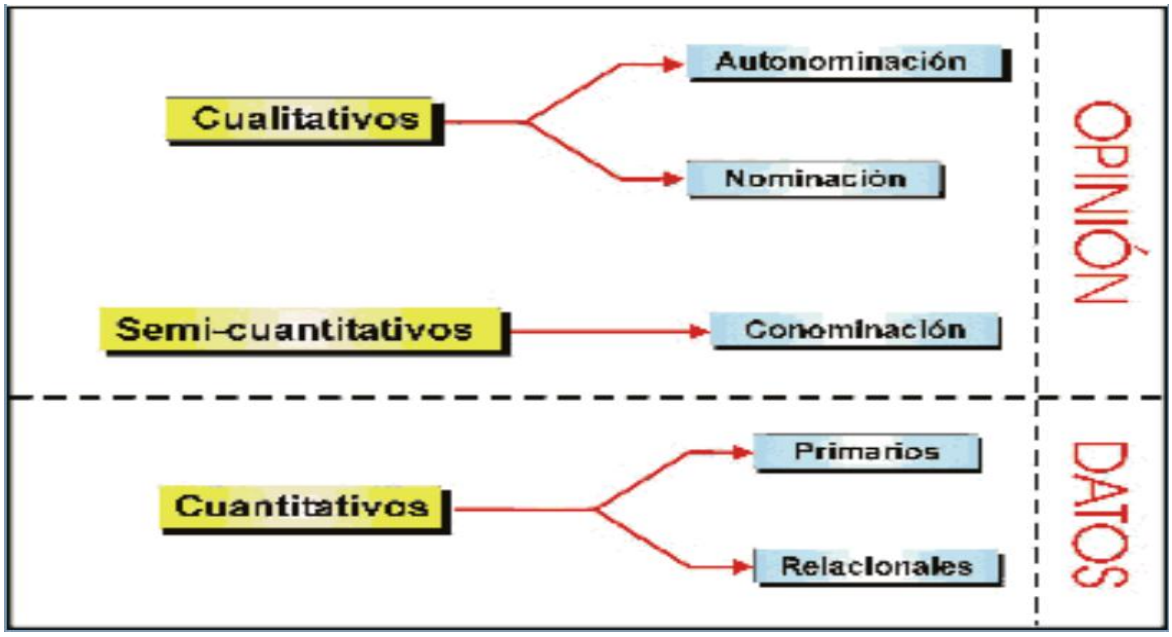


Figura 1.2. Métodos de identificación de expertos. Fuente: (Cabrera, 2015).

En el presente trabajo para la identificación de posibles expertos se utilizarán los métodos cuantitativos que surgieron para evitar las limitaciones de los métodos cualitativos que utilizan solamente la información subjetiva, puesto que se utilizará información objetiva basada en datos, que facilitará la identificación de posibles evaluadores, de entre una muestra de los disponibles en una determinada materia.

Antecedentes

Se cuenta con el procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por el Open Journal System, es una propuesta del Ing. Miguel Barrera Fernández.

El procedimiento consta de cuatro etapas, basadas en la minería de textos: conformación y transformación, representación del espacio vectorial, selecciones de rasgos, clasificación y agrupamiento. Una vez conformados los perfiles de usuarios a partir de información textual, se le realizan operaciones de transformación, para luego llevarlos a una representación de espacio vectorial. Se aplican técnicas de selección de rasgos y de lo obtenido se realiza una clasificación y agrupamiento que ayudará a la identificación de posibles evaluadores de artículos. (Barrera-Fernández, 2015)

Con la aplicación del procedimiento se muestran resultados experimentales en cuanto a: la evaluación del grado de ajuste de una clasificación a un conjunto de datos para doce combinaciones posibles y la calidad de la clasificación por medio de la medida f, obteniéndose resultados satisfactorios.

Destacar que las experiencias y opiniones del autor del trabajo mencionado anteriormente son utilizadas en el desarrollo de esta investigación, así como la utilización del procedimiento propuesto para el desarrollo de la aplicación informática.

1.3 Lenguajes, tecnologías y herramientas utilizadas

1.3.1 Lenguaje de Programación

Un lenguaje de programación es un lenguaje que puede ser utilizado para controlar el comportamiento de una máquina, particularmente una computadora. Consiste en un conjunto de reglas sintácticas y semánticas que definen su estructura y el significado de sus elementos, respectivamente(Wikipedia, 2012).

1.3.1.2 R

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico. Se distribuye bajo la licencia

GNUGPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux. (R Foundation). Entre las principales características de este lenguaje tenemos:

- Proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas.
- Permite cargar dinámicamente bibliotecas desarrollas en C, C++ o Fortran y hereda de S su orientación a objetos.
- Puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.
- Su gran capacidad gráfica, permite generar gráficos con alta calidad.
- Posee su propio formato para la documentación basado en LaTeX.
- Puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como GNU Octave y su equivalente comercial, MATLAB.
- Posee la interfaz, RWeka para interactuar con Weka que permite leer y escribir ficheros en el formato arff y enriquecer R con los algoritmos de minería de datos de dicha plataforma.
- Tiene disponible el paquete tm muy útil para la minería de textos.

1.3.2 Entornos de desarrollo integrados (IDE)

Un entorno de desarrollo integrado, llamado también IDE (sigla en inglés de *Integrated Development Environment*), es un programa informático compuesto por un conjunto de herramientas de programación. Puede dedicarse en

exclusiva a un solo lenguaje de programación o bien poder utilizarse para varios.

Un IDE es un entorno de programación que ha sido empaquetado como un programa de aplicación, es decir, consiste en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica (GUI). (Wikipedia, 2012)

1.3.2.1 R Studio

Es un entorno de desarrollo integrado (IDE) para R (lenguaje de programación). Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. Está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux). RStudio tiene la misión de proporcionar el entorno informático estadístico R. Permite un análisis y desarrollo para que cualquiera pueda analizar los datos con R (Wikipedia, 2012).

1.3.3 Sistema Gestor de Base de Datos

Consiste en un conjunto de programas, procedimientos y lenguajes que nos proporcionan las herramientas necesarias para trabajar con una base de datos. Incorporar una serie de funciones que nos permita definir los registros, sus campos, sus relaciones, insertar, suprimir, modificar y consultar los datos.

1.3.3.1 MySQL

Es un sistema de gestión de bases de datos relacional, multihilo y multiusuario. Por un lado se ofrece bajo la GNU GPL para cualquier uso compatible con esta licencia, pero para aquellas empresas que quieran incorporarlo en productos

privativos deben comprar a la empresa una licencia específica que les permita este uso. Está desarrollado en su mayor parte en ANSI C. (Wikipedia, 2012)

Una base de datos es una colección estructurada de tablas que contienen datos. Esta puede ser desde una simple lista de compras a una galería de pinturas o el vasto volumen de información en una red corporativa. Para agregar, acceder a y procesar datos guardados en un computador, usted necesita un administrador como MySQL Server.
En el caso de la presente investigación los datos del caso de estudio fueron obtenidos de un servidor de bases de datos en particular: MySQL Server.

1.4 Metodologías de trabajo

Antes de comenzar a desarrollar cualquier proyecto de minería de, se debe seleccionar una metodología de trabajo. Esta metodología se debe seguir paso a paso, con el objetivo de comprender cada una de sus fases, una metodología que explique cuando se debe hacer cada actividad y su razón.
Para este proyecto en particular el objetivo principal de la correcta aplicación de una metodología es alcanzar el objetivo propuesto al inicio de esta investigación, así como para la ayuda a las personas encargadas del proceso de toma de decisiones en el Consejo Editorial de una revista científica gestionada con OJS.
Tomando en cuenta la gran cantidad de información que se maneja dentro de las revistas científicas y el progreso tecnológico de la última década, se han creado diferentes metodologías para realizar un análisis utilizando minería de datos, estableciendo ciertos parámetros que se deben cumplir, dependiendo de la información que se tenga y de lo que concretamente se desea, aunque los resultados sean obtenidos en un plazo no muy corto ya que una metodología

consta de ciertas fases sucesivas que hay que seguir respetando el orden jerárquico.

Ciertas empresas han desarrollado metodologías para que el usuario pueda seguir utilizando al máximo la información. SAS[1] por ejemplo, puso a disposición de los usuarios la metodología SEMMA por sus siglas en inglés (*Sample, Explore, Modify,Model, Assess*). La empresa Microsoft también tiene su propia guía para proyectos de minería de datos, llamada por su mismo nombre. IBM creó la metodología CRISP- DM por sus siglas en inglés (*Cross-Industry Standard Processfor Data Mining*).

Las metodologías SEMMA, Microsoft y CRISP-DM comparten la misma esencia, estructurando el proyecto de *Data Mining* en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de minería de datos en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM y Microsoft, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto, donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM y Microsoft comienzan realizando un análisis del problema empresarial, para su transformación en un problema técnico.

Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto, pudiendo ser

[1] SAS: http://www.sas.com/
SAS es uno de los líderes en Business Analytics y servicios de software, y uno de los mayores proveedores independientes de Inteligencia de Negocio del mercado.

integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Otra diferencia significativa entre la metodología SEMMA, Microsoft y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Analizando la propuesta metodológica de Microsoft se puede ver que está íntimamente vinculada a la aplicación de las herramientas de su propia compañía (Microsoft) especialmente en lo que respecta a la integración de servicios, vista de origen de datos y diseñador de minería de datos. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto, siendo su distribución libre y gratuita.

A continuación se muestra en la tabla 1.1 la comparación entre las metodologías:

Metodologías	CRISP-DM	SEMMA	Microsoft
Estructura	Fases y niveles	Fases	Fases
Niveles	Parte de lo general a lo específico	No tiene	No tiene
Fases	Análisis del problema Análisis de datos Preparación de Datos Modelado Evaluación Explotación	Muestreo Exploración Manipulación Modelado Valoración	Definir el problema Preparar los datos Explorar los datos Generar modelos Explorar y validar los modelos Implementar y actualizar los modelos
Herramientas	Genéricas	SAS	Microsoft
Procesos	Iterativo e interactivo entre fases	Iterativo e interactivo entre fases	Iterativo e interactivo entre fases
Documentación	Modelo de referencia Guía de usuario	No se especifica	No se especifica
Objetivos	Se centra en los objetivos empresariales del proyecto	Se centra en las características técnicas del desarrollo del proceso	Se centra en los objetivos empresariales del proyecto

Tabla 1.1. Cuadro comparativo de metodologías

1.4.1 Elección de la Metodología

Se escogió CRISP-DM como la metodología más apropiada para el desarrollo del trabajo de tesis por considerarla más completa que SEMMA, principalmente porque posee una fase de desarrollo dedicada íntegramente al entendimiento del negocio, y por su flexibilidad, al permitir trabajar con cualquier herramienta de explotación de datos.

A continuación se ampliaran las definiciones hechas sobre la metodología CRISP-DM:
La Metodología CRISP-DM, aunque se desarrolló para llevar adelante grandes proyectos, es suficientemente amplia y flexible para aplicarla a proyectos de cualquier tamaño. En la figura 1.3, se esquematiza el ciclo de vida de un proyecto desarrollado con la metodología.

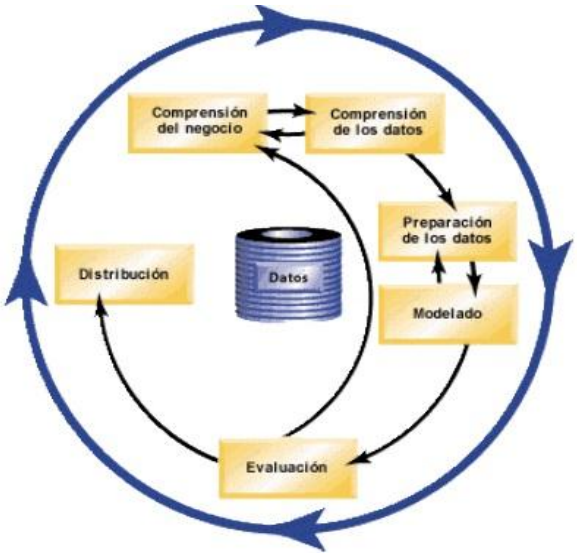


Figura 1.3. Modelo de proceso CRISP–DM ([CRISP-DM, 2000])

El ciclo de vida de la metodología consiste en seis fases, cuya sucesión no es rígida, y se puede mover entre ellas siempre que se requiera. Las flechas indican la dependencia más importante y frecuente entre las fases. El círculo exterior simboliza la naturaleza cíclica de los proyectos de minería de datos.

La metodología se presenta en términos de un proceso jerárquico. Consiste en un juego de tareas descriptas en niveles de abstracción (de lo general a lo específico): la fase, la tarea genérica o subfase, la tarea especializada y el caso del proceso.

El contexto de CRISP-DM se maneja entre lo genérico y el nivel especializado, dentro del cual se distinguen cuatro dimensiones diferentes:

- ✓ *Dominio de la aplicación:* Especifica el área en que el proyecto de minería de datos tiene lugar.
- ✓ *Tipo de problema:* Describe la clase y objetivos del proyecto.
- ✓ *Aspecto técnico:* Cubre procesos específicos de la minería de datos, describe diferentes desafíos que normalmente ocurren.
- ✓ *Herramienta técnica:* Especifica que se aplica durante el proceso de minería de datos.

1.4.2 Fases.

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 (figura 1.4), Esta fase inicial se focaliza en el entendimiento de los objetivos y requerimientos desde una perspectiva de negocios. Este conocimiento se convierte en una definición de problema de minería de datos y en un plan preliminar diseñado para llevar a cabo los objetivos.

El primer objetivo es comprender a fondo, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo el cliente tiene muchos objetivos y restricciones que compiten, los cuales deben ser correctamente equilibrados. El objetivo del analista es destapar factores importantes en el principio del proyecto esto puede influir en el resultado final. Una consecuencia probable de descuidar este paso debe ser a expensas de hacer un gran esfuerzo de producir las respuestas correctas a las preguntas incorrectas.

Finalmente esta fase debe terminar con la descripción del plan intencionado para alcanzar los objetivos de minería de datos y así alcanzar los objetivos de

negocio. El plan debería especificar los pasos para ser realizados durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas.

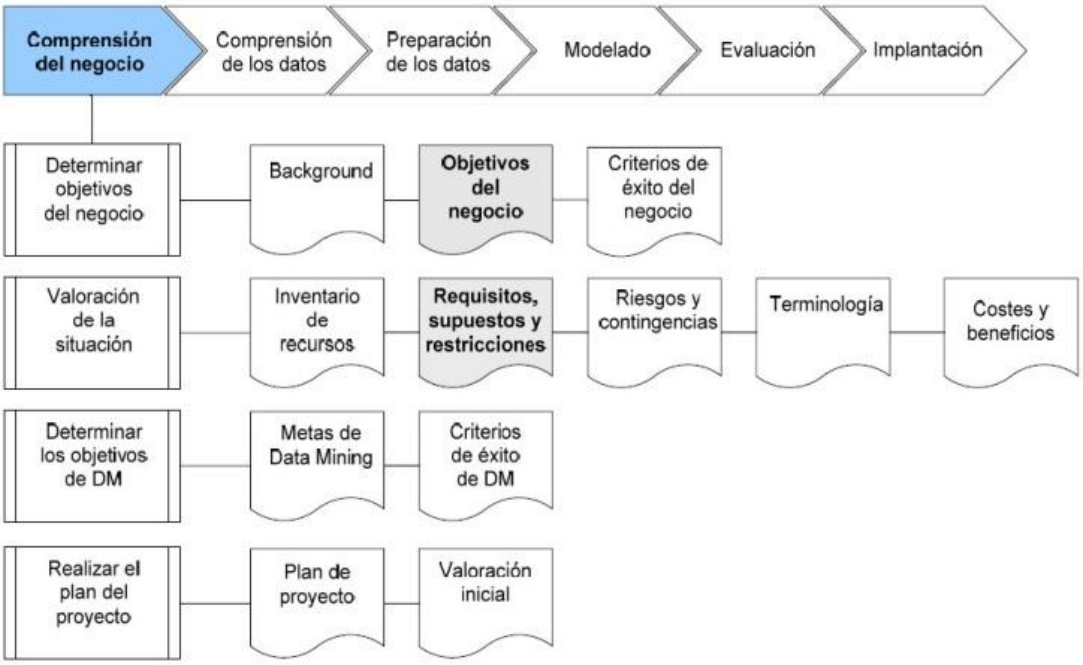


Figura 1.4. Fase de comprensión del negocio ([CRISP-DM, 2000]).

Fase de comprensión de los datos

La fase de comprensión de los datos (figura 1.5), comienza con una colección inicial y continúa con actividades tendientes a familiarizarse con ellos, para identificar los problemas de calidad, descubrir primeras vistas internas o detectar interesantes subconjuntos para formular hipótesis sobre la información oculta. Esta colección inicial incluye carga de datos, si es necesario, para la comprensión de los datos, significando un esfuerzo que posiblemente conduce a los pasos iniciales de la preparación de datos.

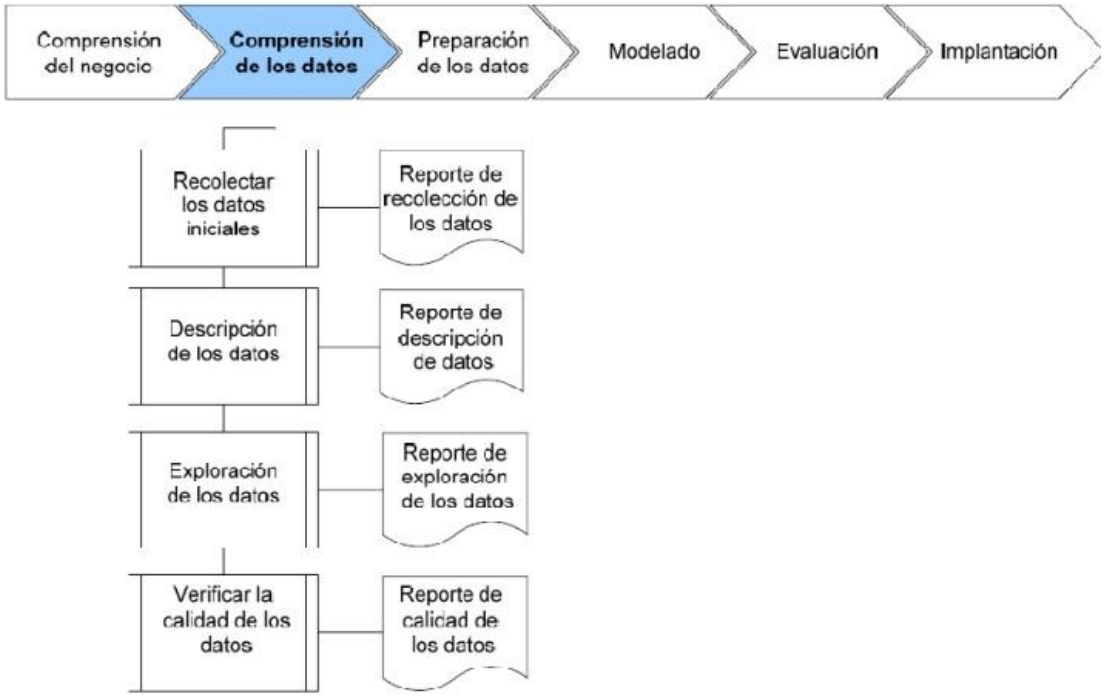


Figura 1.5. Fase de comprensión de los datos ([CRISP-DM, 2000]).

Fase de preparación de los datos

La fase de preparación de datos (figura 1.6) cubre todas las actividades para construir el conjunto de datos finales a partir de los datos iniciales en bruto. Este conjunto de datos llamado datos de aprendizaje conforman las tablas de escenarios que se utilizaran para entrenar los modelos de minería de datos.

Las tareas de preparación de datos se suelen desarrollar múltiples veces y no tienen un orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos como también la transformación y limpieza de los datos por herramientas de modelamiento.

En algunas ocasiones se desarrollan operaciones de preparación de datos, tales como la producción de atributos derivados, el ingreso de nuevos registros o la transformación de valores para atributos existentes.

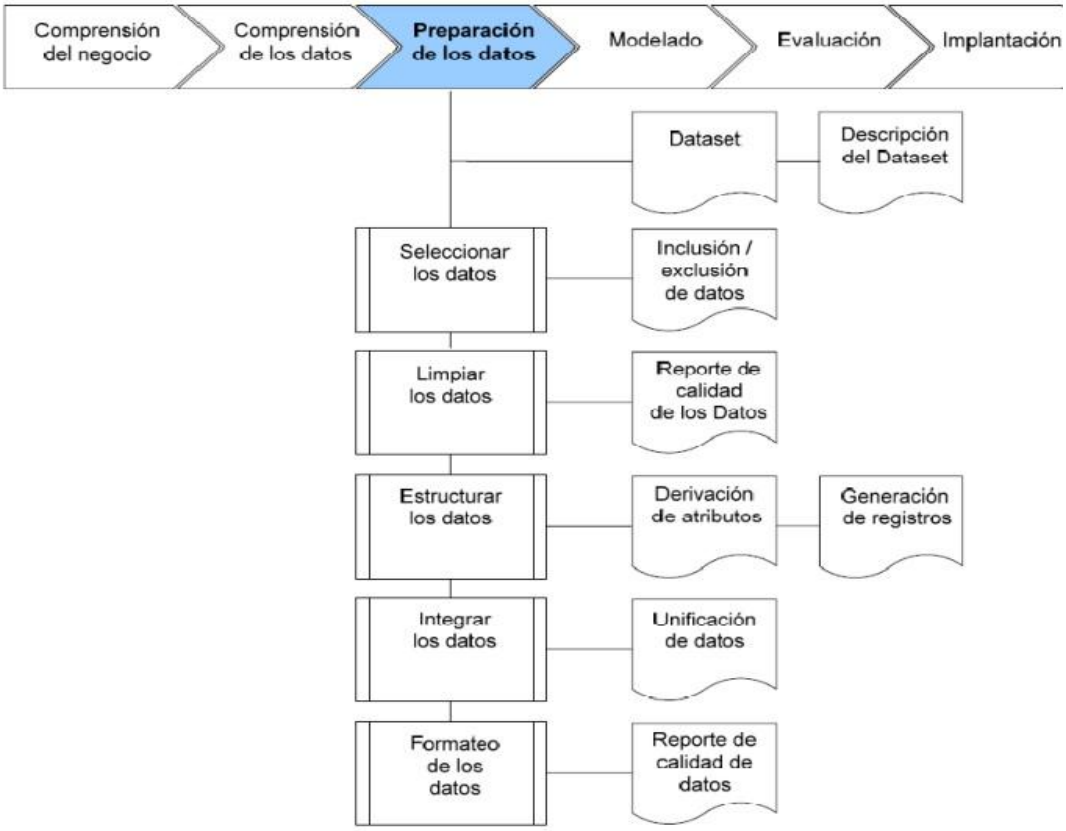


Figura 1.6. Fase de preparación de los datos ([CRISP-DM, 2000]).

Fase de modelado

En esta fase de CRISP-DM (figura 1.7), se seleccionan las técnicas de modelado más apropiadas para el proyecto de *Data Mining* específico. Las técnicas a utilizar en esta fase se eligen en función delos siguientes criterios:

- Ser apropiada al problema.

- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Típicamente hay algunas técnicas para los mismos tipos de problemas de minería de datos, aunque algunas técnicas tienen requerimientos específicos en la forma de los datos, y muchas veces es necesario volver a la fase de preparación de datos.

Como primer paso, se selecciona la técnica de modelado real que se utilizará. Aunque la selección de una herramienta fue realizada durante la fase de comprensión del negocio, esta tarea se refiere a la técnica de modelado específico. Si se aplican múltiples técnicas, cada tarea se realizará separadamente para cada técnica.

El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios exitosos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre nuevas técnicas más eficientes; él se pone en contacto con analistas del negocio para hablar de los resultados de la minería de datos en el contexto de negocio.

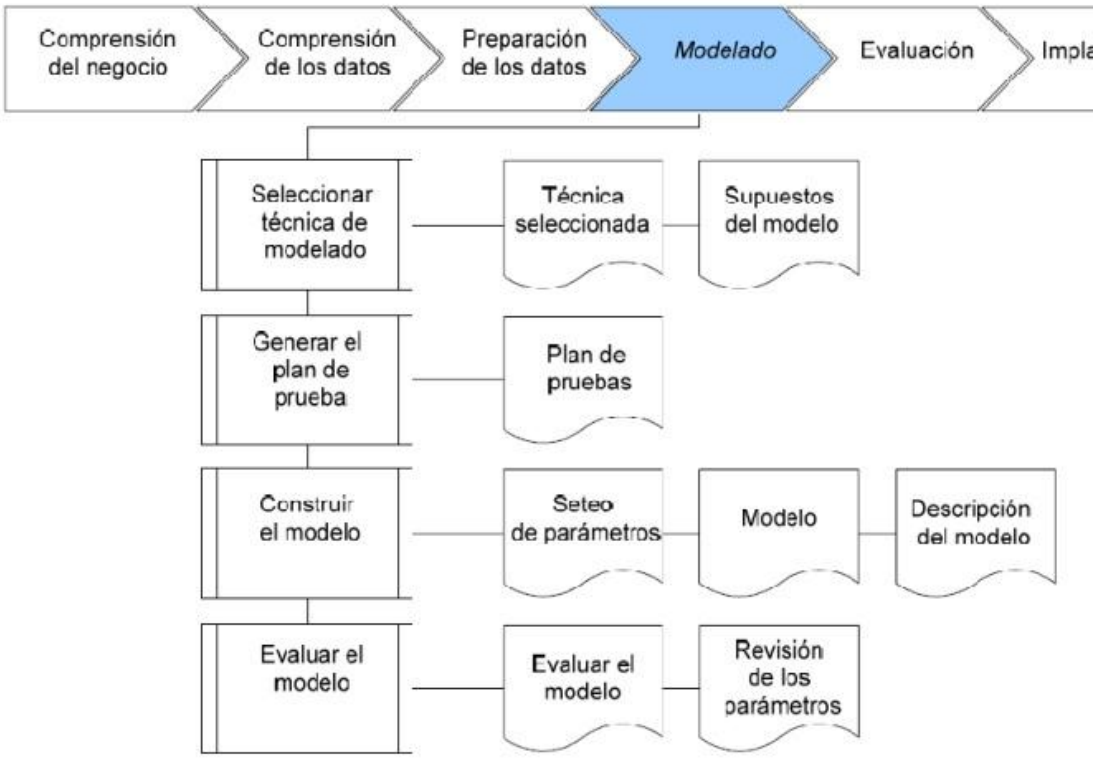


Figura 1.7. Fase de modelado ([CRISP-DM, 2000]).

Fase de evaluación

En esta fase (figura 1.7), ya se ha construido el modelo (o modelos) que parecen tener alta calidad desde una perspectiva de análisis de datos. Antes de proceder a la implementación final del modelo, se debe evaluar de forma más exhaustiva el modelo y revisar los pasos ejecutados para construirlo y asegurarse que interprete de forma adecuada los objetivos del negocio. Un objetivo clave es determinar si hay algún aspecto importante del negocio que no haya sido suficientemente considerado. Al final de esta fase, se debería haber alcanzado alguna decisión en el uso de los resultados de minería de

datos. Los pasos de la evaluación trata factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

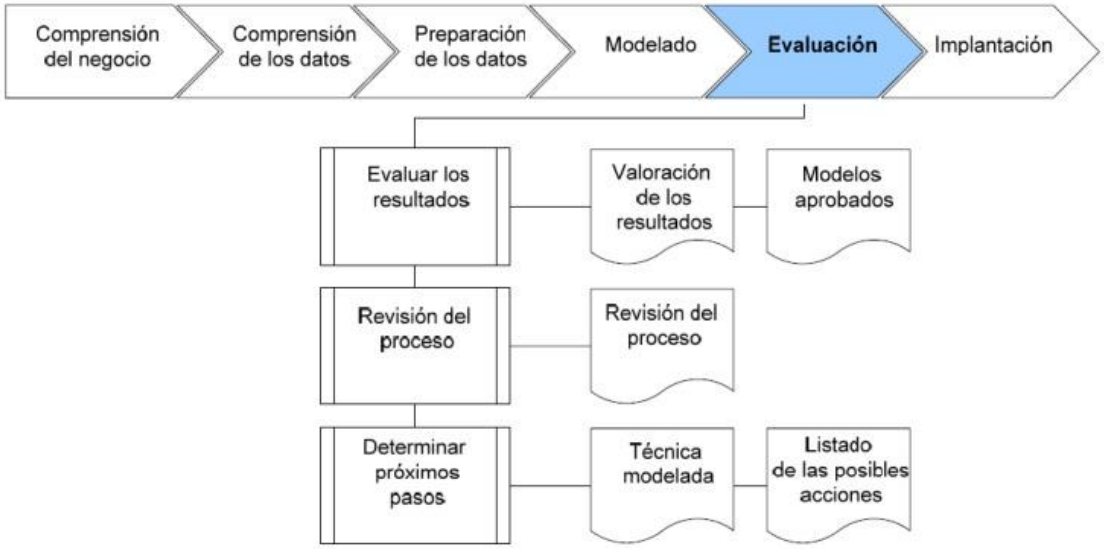


Figura 1.8. Fase de evaluación ([CRISP-DM, 2000]).

Fase de implementación

En esta fase (figura 1.9),la creación del modelo no es, por lo general, el final del proyecto. Aún si el propósito del modelo es incrementar el conocimiento de los datos, el conocimiento obtenido necesitara ser reorganizado y presentado de una manera que el cliente pueda utilizarlo. A menudo, la aplicación involucra modelos “vivos” dentro de los procesos de toma de decisiones de la

organización, por ejemplo, en la personalización de páginas web en tiempo real.

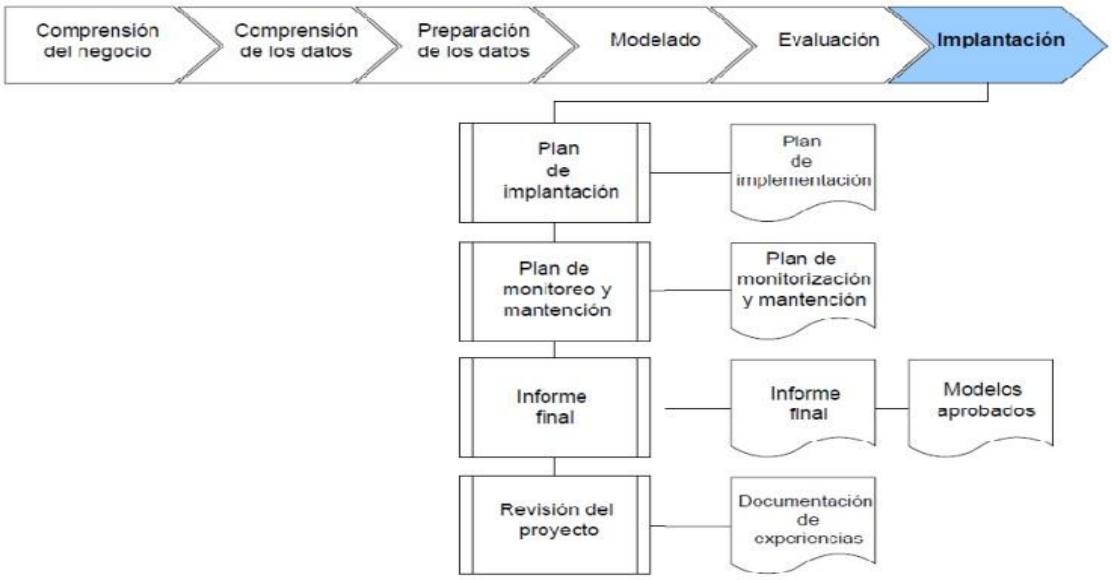


Figura 1.9. Fase de implementación ([CRISP-DM, 2000]).

1.5 Conclusiones del Capítulo.

Después del estudio realizado de los diferentes aspectos tratados en este capítulo quedan sentadas las bases para el desarrollo de la Aplicación web para la creación y agrupamiento de perfiles de usuario en revistas científicas gestionadas con Open Journal System.

- Metodología: CRISP-DM.
- Lenguajes de Programación: R
- Herramientas de Desarrollo: R Studio.

Capítulo 2. Descripción de la Solución Propuesta.

2.1 Introducción.

En el presente capítulo se hace una descripción detallada de la solución que se propone para el problema planteado, para lograr tal propósito se aplican las primeras cinco fases de la metodología CRISP-DM, antes mencionada, en virtud de lograr cumplimentar los objetivos propuestos al inicio de esta investigación.

2.2 Comprensión del negocio

A continuación se describirá detalladamente la aplicación de la primera fase de la metodología CRISP-DM, así como cada una de sus tareas, esta fase es Comprensión del negocio.

2.2.1 Definición del problema

Por la cantidad de información almacenada en las bases de datos de las revistas gestionadas con OJS (información textual) y la poca existencia de mecanismos u opciones en el sistema para procesar estos datos, se dificulta la identificación de los autores de artículos de la propia plataforma que podrían servir cómo posibles evaluadores de un artículo determinado, lo que genera en ocasiones que el proceso de selección de evaluadores sea un poco lento, que se desconozcan algunos de los posibles candidatos y que de conocerse no se sepa con claridad la estructura jerárquica de estos respecto al tema del artículo científico a evaluar.

2.2.1.1 Objetivos del negocio

Una solución propuesta es crear una aplicación informática que permita conocer la similitud existente entre los perfiles de usuarios y la clasificación de estos, generados de la información contenida en una revista gestionada con OJS, con la finalidad de favorecer el proceso de identificación de posibles árbitros de artículos científicos.

2.2.1.2 Criterios de éxito del negocio.

- Crear los perfiles de usuario y clasificarlos según un tema dado.
- Ordenar de forma jerárquica los posibles evaluadores respecto a un tema.
- Mejorar el proceso de selección de expertos teniendo en cuenta la calidad y transparencia de la selección, y así como el tiempo para realizar este análisis.

2.2.2 Valoración de la situación

2.2.2.1 Inventario de Recursos

Valoración del hardware

El servidor donde debe estar instalado el sistema propuesto, debe cumplir con los siguientes requerimientos mínimos:

- Procesador Pentium 1.5 Ghz.
- 1 GB de Memoria RAM
- Disco Duro de 40 GB.

Evaluando el hardware existente y tomando en cuenta la configuración mínima necesaria, no se requirió realizar inversión inicial para la adquisición de nuevos equipos, ni tampoco para mejorar o actualizar los equipos existentes.

Identificar orígenes de datos y almacenes de conocimientos

¿Qué tipos de datos están disponibles para el análisis?

Se cuenta con la Base de datos de la Revista Científica Minería y Geología, la que está gestionada por Open Journal System, los datos son de carácter textual, como por ejemplo: resúmenes y palabras claves de artículos, nombre completo de los autores así como su grado científico, correo electrónico, etc.

Identificar recursos personales

¿Se dispone del personal necesario para completar el proyecto?

Tratándose de un Proyecto de Trabajo de Diploma, se cuenta con 1 persona que realizará todos los roles de un proyecto. Se espera que el tiempo con que se dispone para realizar la solución propuesta el personal sea suficiente. Para la realización de entrevistas y encuestas de apoyo al proceso de implementación de la solución propuesta, se cuenta con el consejo editorial de la Revista Minería y Geología.

2.2.2.2 Requisitos, supuestos y restricciones

Determinación de requisitos

El requisito fundamental es el objetivo del negocio, mencionado con anterioridad, que es la creación de una aplicación informática que permita clasificar y agrupar los perfiles de usuario de los posibles evaluadores de un

artículo determinado a la vez que se conoce la similitud existente entre ellos, con la finalidad de favorecer el proceso de identificación de posibles árbitros de artículos científicos.

Determinación de supuestos

- No existen factores económicos que impidan la realización del proyecto, evaluando el hardware existente y tomando en cuenta la configuración mínima necesaria, no se requirió realizar inversión inicial para la adquisición de nuevos equipos, ni tampoco para mejorar o actualizar los equipos existentes, tampoco proceden los gastos de representación.
- Los datos con los que se cuenta para el análisis poseen la calidad requerida para su uso, aunque es necesario tomar en consideración que el trabajo con datos textuales es uno de los problemas que tiene la minería de textos, con los que hay que trabajar y mitigar los riesgos que esto pueda ocasionar al proyecto.
- Con la implementación de la solución propuesta, no solo se pretende comprender el modelo creado, sino visualizar los resultados.

Comprobación de restricciones

No se cuentan con restricciones para la implementación de la solución, se disponen de todos los permisos para acceder a los datos para su posterior análisis, no se tienen restricciones legales para el uso de los datos y no se presentan restricciones financieras.

2.2.2.3 Riesgos y Contingencias

El constante acceso a la base de datos, la unión de tablas, la extracción y limpieza de datos con el fin de crear y probar los modelos supone una pérdida

de la información. Para mitigar esto se hace preciso trabajar sobre copias de la base de datos para conservar legible la base de datos original del sistema.
Los factores ambientales tales como: la familiaridad con el modelo de proyecto utilizado, la experiencia en la aplicación, la capacidad del equipo para asimilar los cambios, la motivación, la estabilidad de los requerimientos y la dificultad en el lenguaje de programación también pueden considerarse riesgos para el proyecto. Además del tiempo que se dedicara a la implementación de la solución propuesta, no existen otros riesgos inmediatos en el proyecto. Sin embargo, el tiempo es siempre importante, por lo que el proyecto inicial se programa para 6 meses.

2.2.2.4 Terminologías

La terminología o glosario de términos de este proyecto está disponible en el Anexo 1 (Terminología de negocio) y Anexo 2 (Terminología de minería de datos).

2.2.2.5. Análisis de Coste-Beneficio

Ver Capítulo 3.

2.2.3 Determinación de los objetivos de Minería de Datos.

2.2.3.1 Metas de Minería de Datos

- Creación de un perfil para cada autor de artículo científico que publique en la revista.
- Clasificación y agrupamiento de perfiles de usuario
- Visualización en orden jerárquico de los autores que pudieran servir de árbitros al artículo que se analiza.

2.2.3.2 Criterios de éxito de Minería de Datos

Para lograr las metas de minería de datos, se aplicará el Procedimiento para la Conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por Open Journal System, mencionado anteriormente, el cual consta de las siguientes etapas:

1. *Conformación y transformación.*

Se pueden considerar según (Arco-García et al., 2007) dos grandes tipos de operaciones con los corpus textuales: *operaciones de conformación* y *operaciones de transformación.* El primer tipo incluye las operaciones que tienen el objetivo de conformar el propio corpus mediante la adición de textos, el ordenamiento de estos en el corpus, y su delimitación y segmentación. Este tipo de operación se lleva a cabo mediante la desambiguación de los nombres de los autores de artículos científicos y algunas operaciones que incluyen el ordenamiento y delimitación de los datos donde finalmente quedan conformados los perfiles de usuario. El segundo tipo de operaciones, las de transformación, se realizan sobre un corpus ya conformado.

2. *Representación del espacio vectorial.*

Los perfiles de usuario se generarán y se almacenarán en una nueva vista de la base de datos del propio sistema OJS. Desde el punto de vista matemático, la base de datos es una tabla que contiene una matriz en la que cada fila representa a un usuario y cada columna indica la presencia, o no, de un determinado término en su perfil correspondiente

3. *Selección de rasgos*

La selección de rasgos usada para representar un dominio tiene un efecto profundo en la calidad del modelo producido. Los rasgos bien seleccionados pueden mejorar la exactitud de las técnicas de minería de textos sustancialmente y reducir la cantidad de datos necesarios para obtener el nivel de funcionamiento deseado (Forman, 2003).

Las técnicas de selección de rasgos toman como entrada un conjunto de rasgos y producen como salida un subconjunto de esos rasgos, los cuales son relevantes para el problema que se quiera resolver (Lanquillon, 2001). Obviamente, realizar una búsqueda exhaustiva es intratable desde el número de rasgos que es usualmente muy grande en el dominio de textos. Por tal motivo, la selección de rasgos puede ser guiada por heurísticas.

4. *Clasificación y agrupamiento*.

Podríamos decir que en el trabajo con perfiles de usuario textuales un problema de clasificación surge cuando se quiere decidir si un perfil de usuario pertenece a una categoría preestablecida de perfiles de usuarios. Si la representación de estos perfiles de usuarios textuales se realiza por medio del Modelo de Espacio Vectorial (MSV) entonces podríamos calcular la similitud existente entre una categoría determinada y los perfiles de usuario. Para esto

tendríamos que utilizar una de las medidas de distancias para el cálculo de la similitud entre vectores.

Salton, establece un modelo matemático para la recuperación de información basado en el cálculo del coeficiente de similitud entre vectores (Salton, 1971, 1989; Salton y McGill, 1983; Salton et al., 1975). Este modelo de cierta forma responde a las necesidades del presente estudio, ya que para obtener el grado de relevancia de los usuarios según su perfil con respecto a una categoría determinada, es posible establecer la similaridad entre los vectores de los usuarios respecto al vector categoría, o sea cada vector lo constituirá un usuario y será posible determinar la similitud de cada usuario con respecto a una categoría. El sistema tomará un valor real que será tanto mayor cuanto más similares sean los usuarios respecto a una categoría.

2.2.4 Realización del Plan de Proyecto

2.2.4.1 Plan de Proyecto

Fase	**Tiempo**	**Recursos**	**Riesgos**
Comprensión del negocio	2 semana	1 persona (Rol de analista)	Incapacidad del personal que laborará en el proyecto para lograr un entendimiento con el cliente, tanto para la delimitación de las metas y objetivos como para la aceptación de la solución que se propone.

Comprensión de los datos	3 semanas	1 persona (Rol de analista)	Problemas en los datos, problemas tecnológicos tanto de software y hardware, así como la preparación del personal.
Preparación de los datos	5 semanas	1 persona (Asesor de minería de datos)	Problemas en los datos, problemas tecnológicos tanto de software y hardware, así como la preparación del personal.
Modelado	4 semanas	1 persona (Asesor de minería de datos)	Problemas en los datos, problemas tecnológicos tanto de software y hardware, así como la preparación del personal. Incapacidad para encontrar un modelo adecuado
Evaluación	2 semana	1 persona (Rol de analista)	Incapacidad para la evaluación crítica de los resultados obtenidos.
Implementación	8 semana	1 persona (Asesor de minería de datos)	Incapacidad para implementar los resultados. Dificultad del lenguaje de Programación.

2.3 Comprensión de los datos

Se describirá detalladamente la aplicación de la segunda fase de la metodología CRISP-DM, así como cada una de sus tareas, esta fase es Comprensión de los datos.

2.3.1 Recolección de datos iniciales

2.3.1.1 Reporte de recolección de datos.

Datos existentes: Se cuenta con la Base de datos de la Revista científica Minería y Geología, la cual consta de 115 tablas, las cuales recogen de manera general datos del autor (nombre completo, e-mail, grado científico, institución a la que pertenecen, etc.), datos de los artículos (título, resúmenes, palabras claves, área o disciplina de estudio, comentarios, referencias bibliográficas, galerías, citas textuales, fecha de publicación, fecha de la última modificación, etc.). De este gran conjunto de datos, para este estudio, solo van a ser relevantes las tablas: *authors, authors_settings* y *article_settings*, a partir de las cuales se generarán otras.

Existen datos suficientes para obtener conclusiones generales o realizar predicciones, se disponen de atributos suficientes para utilizar los métodos de modelado. Hasta este momento no se están utilizando datos con otros orígenes.

2.3.2 Descripción de los datos

2.3.2.1 Reporte de descripción de datos

Cantidad de datos: Como se mencionaba con anterioridad, se cuenta con la Base de datos de la Revista científica Minería y Geología, la cual consta de 115

tablas. Para este estudio se utilizarán solo 3 tablas (*authors, authors_settings* y *article_settings*), los datos están representados de forma textual (lenguaje natural). En la siguiente figura (Figura 2.1) se muestra el proceso de obtención de datos.

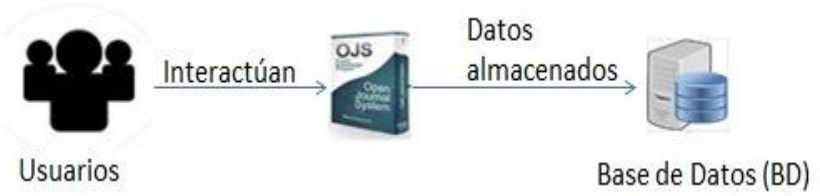


Figura 2.1. Obtención de datos. (Creación propia)

Calidad de los datos: Los datos seleccionadas para la implementación de la aplicación para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por Open Journal System cuentan con la calidad requerida para su análisis, además de proporcionar información relevante para el dominio del problema.

2.3.3 Exploración de datos

Reporte de exploración de datos.

Aunque de manera general los datos que se han tomado para su análisis y para la posterior implementación de la solución propuesta poseen la calidad requerida, se pueden observar errores en las primeras exploraciones, por ejemplo la aparición de etiquetas HTML, siendo esto no relevante pues puede ser solucionado aplicando técnicas de transformación como la eliminación de etiquetas HTML

Los datos seleccionados pueden adaptarse tanto a los objetivos de negocios que fueron delimitados en la primera fase de la aplicación de la metodología como a los objetivos de minería de datos que tributan a la correcta implementación de la solución propuesta para el problema de selección de expertos para evaluar artículos científicos de revistas gestionadas con OJS.

Los atributos *abstract*(referido a resumen de artículo) y *keywords*(referido a palabras claves de artículos) de la tabla *article_settings* son relevantes para la creación de los perfiles de usuario, así como el atributo *last_name*, estos serán la base para la aplicación del procedimiento para la Conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por Open Journal System, propuesto por Barrera Fernández, 2015.

2.3.4. Verificación de la calidad de datos

2.3.4.1 Reporte de calidad de datos.

Datos perdidos: Los datos relevantes para el análisis, como *last_name* y *abstract* no contienen valores vacíos o codificados como sin respuesta ($null$, ? o 999). En cambio algunos valores del atributo *keywords* presentan datos perdidos, lo que pudiera ocasionar ruido. Por tratarse de información textual no se aplicaran técnicas de predicción para completar estos valores.

Los errores de datos*:* suelen ser errores tipográficos cometidos al introducir los datos. La mayoría de los orígenes de datos se generan automáticamente, por lo que no es un problema grave, en el caso de los atributos relevantes para

este estudio se presenta este problema en algunos nombres de autores de artículos, muchos de los cuales pueden ser solucionados aplicando transformaciones, los que no puedan ser solucionados por problemas ortográficos pueden constituir ruidos y causar problemas en futuras fusiones o transformaciones, otro problema es la posibilidad de representar el nombre de los autores de diferentes formas en los metadatos bibliográficos presentes en los repositorios digitales. Este se puede manifestar de dos formas diferentes, (1) pueden aparecer nombres de autores iguales, pero que no se refieren al mismo autor y (2) aparecen nombres diferentes, pero que se refieren al mismo autor. (Alonso-Sierra, Hidalgo-Delgado y Leiva-Mederos, 2014)

Para poder crear por cada autor un identificador único (id), lograr la generación de perfiles de usuario reales, de forma tal que no exista ambigüedad en los nombres de los autores de artículos, hay que realizar una homogeneidad en los datos de la tabla *authors* de la base de datos del sistema OJS. Eliminando incongruencias entre nombres de un mismo autor que deberían ser iguales, esto ocurre a veces por un mal procesamiento de la información, principalmente por errores de escritura. No es objetivo de esta investigación realizar la desambiguación de nombres.

Los errores de mediciones: incluyen datos que se introducen correctamente, pero se basan en un esquema de mediciones incorrecto. No se presentan para atributos relevantes.

Las incoherencias de codificación: suelen incluir unidades no estándar de medidas o valores incoherentes, como el uso de M y masculino para expresar el género. No se presentan para atributos relevantes.

Los metadatos erróneos: incluyen errores entre el significado aparente de un campo incluido en un nombre o definición de campo. No se presentan para atributos relevantes.

2.4 Preparación de los datos

A continuación se describirá detalladamente la aplicación de la tercera fase de la metodología CRISP-DM, así como cada una de sus tareas. En esta fase de preparación de los datos, se cubrieron todas las actividades necesarias para construir el conjunto de datos final o datos de aprendizaje. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

La primera tarea ejecutada, fue finalizar el proceso de selección de los datos y la conformación de las tablas de escenarios. Para ello se considera junto a los datos o atributos investigados en la etapa anterior, un conjunto de datos que permiten identificar claramente los autores de artículos científicos. Esto será de vital importancia para la conformación de los perfiles de estos usuarios del sistema OJS.

2.4.1 Selección de los datos

2.4.1.1 Inclusión / Exclusión de datos

Los datos considerados relevantes para cumplir los objetivos de minería de datos son los siguientes:

- *last_name*(apellidos del autor)
- *abstract*(resumen de artículo)
- *keywords* (palabras claves del artículo)

El origen de estos datos se encuentran en dos tablas que serán fundamentales a la hora de construir las tablas de escenarios, estas son las tablas que contienen los datos generales del autor (tabla: *authors*) y la tabla que almacena todos los datos de los artículos científicos publicados en la revista (tabla: *article_settings*).

2.4.2 Limpieza de datos

2.4.2.1 Reporte de calidad de datos

Luego de realizar la exploración de datos en esta etapa se pudo constatar que se encontraron las mismas deficiencias en los datos que en la fase anterior, por tanto el Reporte de calidad de datos se ajusta al de la fase anterior.
Ver epígrafe 2.3.4.1.

article_id	locale	setting_name	setting_value
3	es_ES	cleanTitle	Particularidades geológicas del complejo ofiolitic...
3	es_ES	abstract	<p>En este artículo se presentan los resultados ob...
3	es_ES	sponsor	

Figura 2.2 Ejemplo de ruido en los datos

La tarea de limpieza de datos se hará una vez que se complete la integración de datos y de conjunto con la tarea de formateo de los datos, esto se realizará con el objetivo de modificar solamente los datos que son relevantes para este estudio.

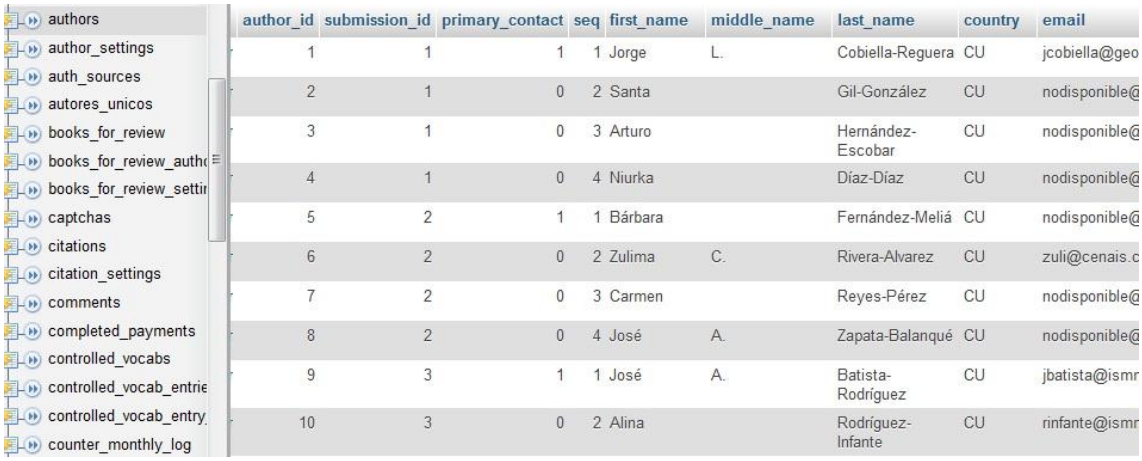
2.4.3 Estructuración de los datos

2.4.3.1 Derivación y generación de atributos

No es necesaria la creación de nuevas columnas o filas en las tablas ya existentes, la tarea más conveniente para esta investigación es la integración de datos, que se detallará en el próximo epígrafe.

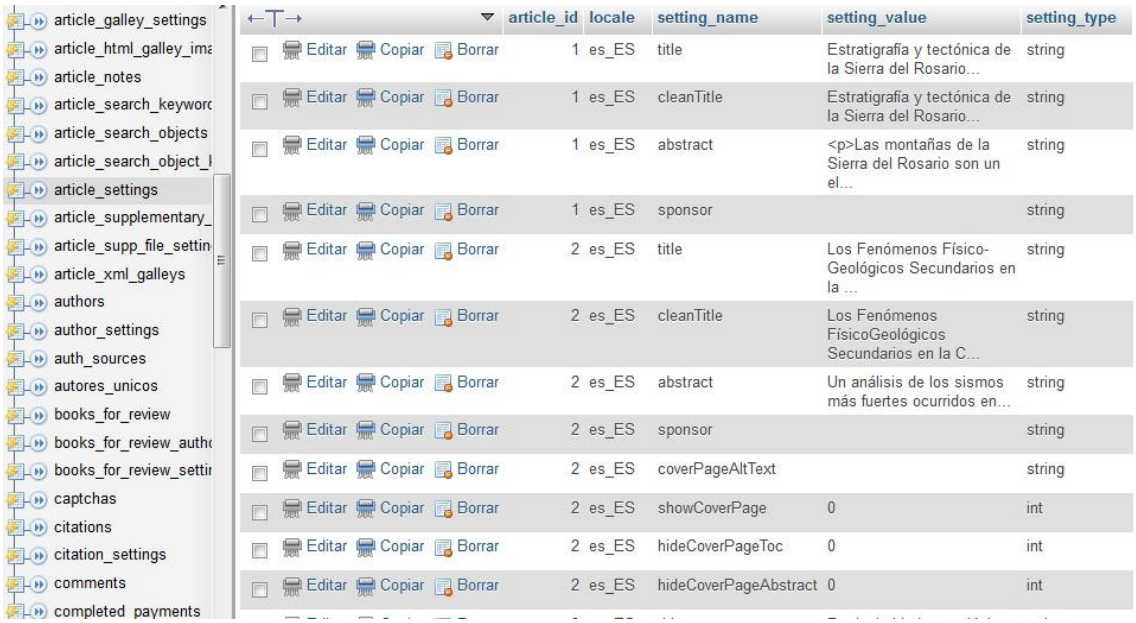
2.4.4 Integración de datos

El origen de los datos relevantes para esta investigación se encuentra en tablas diferentes, como ya se ha ido mencionando con anterioridad, estos conjuntos de datos contienen el mismo identificador único por lo que pueden ser fusionados para organizarlos en un único origen. En las figuras 2.3 y 2.4 se muestran las tablas *authors* y *article_settings*, las que contienen los datos relevantes para la confección de las tablas escenario, que contendrá la unificación de los datos, estas son la materia prima para crear los modelos de minería de datos



author_id	submission_id	primary_contact	seq	first_name	middle_name	last_name	country	email
1	1	1	1	Jorge	L.	Cobiella-Reguera	CU	jcobiella@geo
2	1	0	2	Santa		Gil-González	CU	nodisponible@
3	1	0	3	Arturo		Hernández-Escobar	CU	nodisponible@
4	1	0	4	Niurka		Díaz-Díaz	CU	nodisponible@
5	2	1	1	Bárbara		Fernández-Meliá	CU	nodisponible@
6	2	0	2	Zulima	C.	Rivera-Alvarez	CU	zuli@cenais.c
7	2	0	3	Carmen		Reyes-Pérez	CU	nodisponible@
8	2	0	4	José	A.	Zapata-Balanqué	CU	nodisponible@
9	3	1	1	José	A.	Batista-Rodríguez	CU	jbatista@ismr
10	3	0	2	Alina		Rodríguez-Infante	CU	rinfante@ismr

Figura 2.3. Fragmento de la Tabla *Authors*



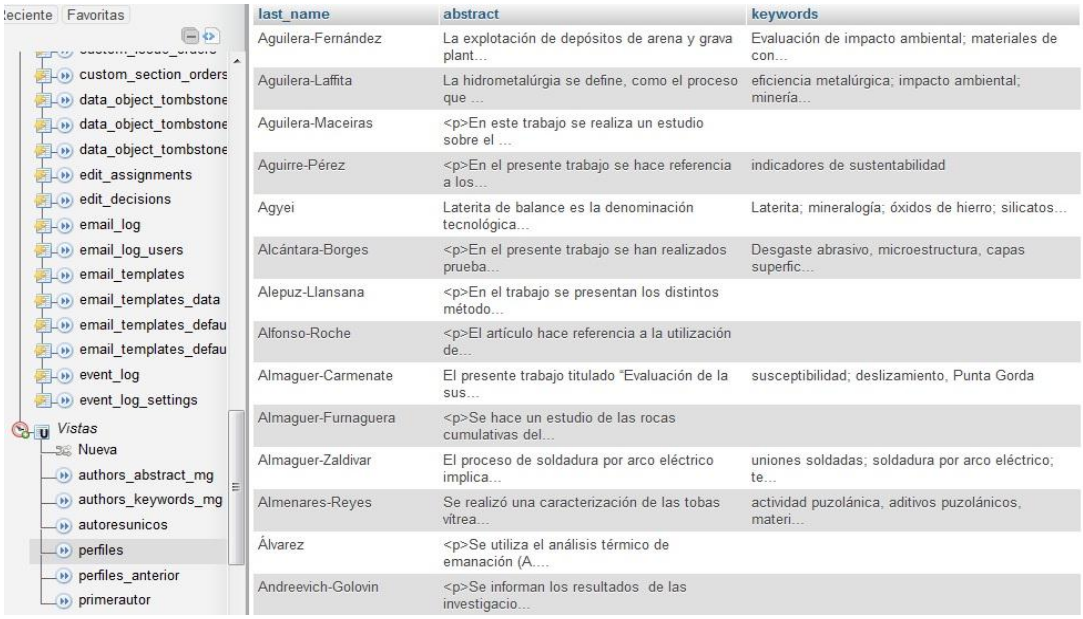
article_galley_settings
article_html_galley_ima
article_notes
article_search_keyworc
article_search_objects
article_search_object_l
article_settings
article_supplementary_
article_supp_file_settin
article_xml_galleys
authors
author_settings
auth_sources
autores_unicos
books_for_review
books_for_review_autho
books_for_review_settir
captchas
citations
citation_settings
comments
completed_payments

	article_id	locale	setting_name	setting_value	setting_type
Editar Copiar Borrar	1	es_ES	title	Estratigrafía y tectónica de la Sierra del Rosario...	string
Editar Copiar Borrar	1	es_ES	cleanTitle	Estratigrafía y tectónica de la Sierra del Rosario...	string
Editar Copiar Borrar	1	es_ES	abstract	<p>Las montañas de la Sierra del Rosario son un el...	string
Editar Copiar Borrar	1	es_ES	sponsor		string
Editar Copiar Borrar	2	es_ES	title	Los Fenómenos Físico-Geológicos Secundarios en la ...	string
Editar Copiar Borrar	2	es_ES	cleanTitle	Los Fenómenos FísicoGeológicos Secundarios en la C...	string
Editar Copiar Borrar	2	es_ES	abstract	Un análisis de los sismos más fuertes ocurridos en...	string
Editar Copiar Borrar	2	es_ES	sponsor		string
Editar Copiar Borrar	2	es_ES	coverPageAltText		string
Editar Copiar Borrar	2	es_ES	showCoverPage	0	int
Editar Copiar Borrar	2	es_ES	hideCoverPageToc	0	int
Editar Copiar Borrar	2	es_ES	hideCoverPageAbstract	0	int

Figura 2.4. Fragmento de la Tabla *article_settings*

2.4.4.1 Unificación de los datos

La tabla perfiles es una tabla generada por una vista que tiene como objetivo integrar todos los autores de artículos, los resúmenes y palabras claves de sus artículos. Como se ve en la figura 2.5, ya se encuentran en un solo origen estos datos (*last_name, abstract* y *keywords*) los que servirán para realizar la creación de los futuros perfiles de usuarios, para la posterior clasificación y agrupamiento de dichos perfiles con el fin de contribuir a la selección de expertos de la revista científica.



last_name	abstract	keywords
Aguilera-Fernández	La explotación de depósitos de arena y grava plant…	Evaluación de impacto ambiental; materiales de con…
Aguilera-Laffita	La hidrometalúrgia se define, como el proceso que …	eficiencia metalúrgica; impacto ambiental; minería…
Aguilera-Maceiras	<p>En este trabajo se realiza un estudio sobre el …	
Aguirre-Pérez	<p>En el presente trabajo se hace referencia a los…	indicadores de sustentabilidad
Agyei	Laterita de balance es la denominación tecnológica…	Laterita; mineralogía; óxidos de hierro; silicatos…
Alcántara-Borges	<p>En el presente trabajo se han realizados prueba…	Desgaste abrasivo, microestructura, capas superfic…
Alepuz-Llansana	<p>En el trabajo se presentan los distintos método…	
Alfonso-Roche	<p>El artículo hace referencia a la utilización de…	
Almaguer-Carmenate	El presente trabajo titulado "Evaluación de la sus…	susceptibilidad; deslizamiento, Punta Gorda
Almaguer-Furnaguera	<p>Se hace un estudio de las rocas cumulativas del…	
Almaguer-Zaldivar	El proceso de soldadura por arco eléctrico implica…	uniones soldadas; soldadura por arco eléctrico; te…
Almenares-Reyes	Se realizó una caracterización de las tobas vítrea…	actividad puzolánica, aditivos puzolánicos, materi…
Álvarez	<p>Se utiliza el análisis térmico de emanación (A.…	
Andreevich-Golovin	<p>Se informan los resultados de las investigacio…	

Figura 2.5. Unificación de datos

2.4.5 Formateo de los datos

Debido a que las técnicas de agrupamiento requieren de instancias únicas en el conjunto de datos a ser minados, la vista minable obtenida de la unificación de los datos fue transformada. La principal razón para esta transformación se debe a que se tenía varios registros con el mismo identificador, los que hacían referencia al mismo autor, la transformación en los datos consistió en establecer un único identificador por autor de artículos científicos, lo que trajo consigo el cumplimiento del primer objetivo de minería de datos: creación de un perfil para cada autor de artículo científico que publique en la revista.

2.4.5.1 Normalización de los datos

A partir de la vista minable obtenida de la unificación de datos se realizaron nuevas transformaciones como:

- Normalización de términos (sustitución de mayúsculas por minúsculas, sustitución de caracteres acentuados por sus homólogos no acentuados).
- *Stemming* (reducir cada forma lingüística a su raíz, *stem* o lema correspondiente),
- Eliminación de *stopwords* (una lista de palabras vacías es utilizada para eliminar términos comunes que no aporta información relevante).
- Eliminación de espacios dobles entre términos, caracteres numéricos, signos de puntuación, etiquetas HTML y términos con menos de 3 caracteres.

2.4.5.2 Reporte de calidad de datos

Ya obtenidos los perfiles de usuario y realizado las tareas de limpieza y formateo de datos, están sentadas las bases para pasar a la Fase de Modelado y no existen problemas de formateo que afecten el tiempo de modelado.

2.5 Modelado

Como se mencionó en epígrafes anteriores para el desarrollo de la solución propuesta al problema de identificación de expertos que podrían servir como

posibles evaluadores y que avalarán los artículos que llegan a las revistas científicas gestionadas con OJS se utilizó el procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas por el Open Journal System, propuesto por Barrera Fernández, 2015.

Para la fase de modelado se utilizarán los supuestos y conclusiones del autor de dicho procedimiento a partir de su segunda etapa (Ver Figura 2.6), se describirán las tareas pertinentes a esta fase como son: selección de técnicas de modelado, creación y descripción del modelo, etc. La primera etapa del mencionado procedimiento quedó cumplida en la fase de preparación de los datos (Ver epígrafes 2.4.4 hasta 2.4.5.1).

Figura 2.6. Etapas del procedimiento.

2.5.1 Representación espacio-vectorial de los perfiles de usuarios

Podemos considerar una base de Perfiles de Usuarios *(U)*, compuesta por usuarios u_i , donde han sido ingresados un conjunto de términos *(T),* formado por *n* términos t_j, en la que cada usuario u_i contiene un número de términos, como resultado de los campos suscritos en el perfil. De esta forma, es posible representar a cada usuario como un vector perteneciente a un espacio *n*-

dimensional, siendo *n* el número de términos ingresados en el perfil que forman el conjunto *T:*

$U_i = (t_{i1}, t_{i2}, t_{i3}; \ldots \ldots \ldots; t_{in})$

Donde cada uno de los elementos t_{ij} de este vector puede representar la presencia, ausencia o relevancia del término t_j en el usuario u_i en su perfil.

El proceso de construcción de los vectores-usuarios en las tablas relacionadas con los perfiles de usuarios generará automáticamente la representación de los usuarios extrayendo los contenidos de información de los perfiles. Por lo que se creará una asociación automática de la representación de cada usuario en función de los contenidos de información de este, o sea, determinar los pesos de cada término extraído de su perfil en el vector usuario u_i. Su función sería:

$$F: U \times T \rightarrow [0,1]$$

La representación de cada vector-usuario tendrá *n* componentes, de los cuales los que estén referenciados en el perfil tendrán un valor diferente de 0, mientras que los que no estén referenciados tendrán un valor nulo o 0.

La frecuencia de aparición de un término en un perfil de cierta forma determina su importancia en él, sugiriendo que dichas frecuencias pueden ser utilizadas para resumir el área de conocimiento en que se mueve el usuario y por ende los principales intereses en cuanto a investigación.

Siguiendo lo que describe el modelo de espacio vectorial y dando continuidad a los métodos usados para almacenar los términos recogidos en el perfil de cada usuario, se continúa con el proceso de selección, a ello le sigue determinar la importancia o peso de cada término en el vector-usuario. El cálculo de la importancia o peso de cada término se conoce como ponderación del término.

Gerald Salton utiliza este concepto de peso en su modelo de recuperación basado en el espacio vectorial. En dicho modelo, se forma una matriz término/documento que representa la base de datos. Cada vector de la matriz representa un documento; cada elemento del vector tendrá valor 0 (cero) si dicho documento no contiene el término; o el valor del peso del término si lo contiene (López-Herrera, 2006; Pérez et al., 2010; Salton, 1971, 1989; Salton y McGill, 1983; Salton et al., 1975; Samper, 2005).

Un primer enfoque se basa en contar las ocurrencias de cada término en un documento, medida que se denomina frecuencia del término *i*-ésimo en el documento *j*-ésimo, y se nota como $tf_{i,j}$. Una segunda medida de la importancia del término es la conocida como frecuencia documental inversa de un término en la colección, conocida normalmente por sus siglas en inglés: *idf* (inverse document frequency), como reflejan (Baeza-Yates y Ribeiro-Neto, 1999; López-Herrera, 2006) y que responde a la siguiente expresión:

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{n_i}\right) \qquad (5)$$

Donde N es el número de documentos de la colección, y n_i el número de documentos donde se menciona al término *i*-ésimo, si asociamos al caso de la

presente investigación a N con U como el número de usuarios de la base de datos de perfiles de usuarios, y n_i como el número de usuarios que contienen en su perfil el término *i,* entonces es posible determinar la importancia o peso de cada término en el perfil de cada uno de los usuarios.

Finalmente se tendría una matriz de *vectores-usuarios* por términos como se muestra a continuación en la figura 2.7.

$$\begin{array}{c|c} & \begin{matrix} t_1 & t_2 & t_3 & \dots & t_n \end{matrix} \\ \hline \begin{matrix} User_1 \\ User_2 \\ User_3 \\ \vdots \\ User_n \end{matrix} & \begin{pmatrix} 1 & 2 & 1 & & n \\ 1 & 1 & 1 & \cdots & n \\ 0 & 2 & 1 & & n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & n & n & \cdots & n \end{pmatrix} \end{array}$$

Figura 2.7. Matriz de perfiles de usuarios

Luego del cálculo de la importancia o peso por medio de la ecuación (1) tendríamos una matriz de peso relacionado con los términos obtenidos en cada uno de los perfiles de usuarios como se muestra a continuación en la figura 2.8:

$$\begin{array}{c|c} & \begin{matrix} t_1 & t_2 & t_3 & \dots & t_n \end{matrix} \\ \hline \begin{matrix} User_1 \\ User_2 \\ User_3 \\ \vdots \\ User_n \end{matrix} & \begin{pmatrix} w_{11} & w_{12} & w_{13} & & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ w_{31} & w_{32} & w_{33} & & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nn} \end{pmatrix} \end{array}$$

Figura 2.8. Matriz del peso (W) de los términos en los perfiles de usuarios

De esta manera queda establecida en una tabla de la base de datos del sistema la matriz de los términos correspondiente a cada uno de los usuarios partiendo de su perfil, que representa las áreas de conocimiento donde incursionan los usuarios.

2.5.2Selección de rasgos

Una pregunta que es común en un problema de clasificación es si todos los rasgos descriptivos serán útiles a la hora de conformar las reglas de clasificación. Intentando responder a esta pregunta, aparece el problema de la selección de rasgos.

La selección de rasgos usada para representar un dominio tiene un efecto profundo en la calidad del modelo producido. Los rasgos bien seleccionados pueden mejorar la exactitud de las técnicas de minería de textos sustancialmente y reducir la cantidad de datos necesarios para obtener el nivel de funcionamiento deseado (Forman, 2003).

Las técnicas de selección de rasgos toman como entrada un conjunto de rasgos y producen como salida un subconjunto de esos rasgos, los cuales son relevantes para el problema que se quiera resolver (Lanquillon, 2001). Obviamente, realizar una búsqueda exhaustiva es intratable desde el número de rasgos que es usualmente muy grande en el dominio de textos. Por tal motivo, la selección de rasgos puede ser guiada por heurísticas.

Se propone la aplicación de algunos de los criterios de selección de rasgos en dominios textuales, entre ellos:

- Eliminar las palabras gramaticales(Sam et al., 2000)
- Eliminar todos los términos cuyas frecuencias están por encima de un umbral superior o por debajo de un umbral inferior especificado. Estos términos tienen poco poder discriminante.
- Eliminar todos los términos cuya frecuencia de documentos es menor que un umbral predeterminado. Esto es basado en la suposición de que

términos que ocurren solamente en muy pocos documentos improbablemente llevan información general de la clase específica y algunas veces tienden a ser ruidosos. Además, usar términos de ocurrencias infrecuentes no es estadísticamente confiable.

- Implementar medidas que cuantifiquen la calidad de los términos, considerando aquellos términos que sobrepasen un umbral determinado.
- Entropía de los términos, como grado de información que transmiten (Arco, 2007).

Será necesario tener mucho cuidado con la aplicación de algunos de estos criterios a la hora de elegir los umbrales correctos que determinen un buen conjunto de palabras de frecuencia media ya que la eliminación de todas las palabras muy frecuentes pude producir pérdida en la exhaustividad, mientras que la eliminación de las palabras poco frecuentes puede ocasionar pérdidas en la precisión.

2.5.3 Clasificación de perfiles de usuarios

Podríamos decir que en el trabajo con corpus de perfiles de usuarios textuales un problema de clasificación surge cuando se quiere decidir si un perfil de usuario pertenece a una categoría preestablecida de perfiles de usuarios. Si la representación de estos perfiles de usuarios textuales se realiza por medio del Modelo de Espacio Vectorial (MSV) entonces podríamos calcular la similitud existente entre una categoría determinada y los perfiles de usuarios. Para esto tendríamos que utilizar una de las medidas de distancias para el cálculo de la similitud entre vectores.

Salton, establece un modelo matemático para la recuperación de información basado en el cálculo del coeficiente de similitud entre vectores (Salton, 1971,

1989; Salton y McGill, 1983; Salton et al., 1975). Este modelo de cierta forma responde a las necesidades del presente estudio, ya que para obtener el grado de relevancia de los usuarios según su perfil con respecto a una categoría determinada, es posible establecer la similaridad entre los vectores de los usuarios respecto al vector categoría, o sea cada vector lo constituirá un usuario y será posible determinar la similitud de cada usuario con respecto a una categoría. El sistema tomará un valor real que será tanto mayor cuanto más similares sean los usuarios respecto a una categoría.

Muchas medidas de similitud entre documentos pueden ser utilizadas, las que han reportado los mejores resultados en dominios textuales son: similitud de Dice, Jaccard y Coseno (Fakes et al., 1992). Entre ellas, la distancia euclideana ha sido la más utilizada para comparar vectores de frecuencias de documentos para un vocabulario de *n* términos (Korfhage, 1977). Por lo que se propone su utilización para la clasificación de los usuarios. La relación coseno medirá el coseno del ángulo entre documentos (*perfiles de usuarios*) y consultas (*categorías*), ya que éstos se representarán como vectores en un espacio multidimensional de dimensión *t*. Así, podemos expresar la medida de similitud entre un documento d_i y una consulta q_k, siendo *n* el número de términos, como:

$$sim(d_i, q_k) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{q_k}}{|\overrightarrow{d_i}| . |\overrightarrow{q_k}|} = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A^2{}_j \cdot \sum_{j=1}^{n} B^2{}_j}} \quad (5)$$

Un ejemplo de cálculo de la similitud, tomado de [Raymond, 2005], puede observarse en la figura 2.9 donde aparecen representados dos documentos d_1, d_2 y una consulta *q* respecto a los ejes t_1, t_2y t_3.

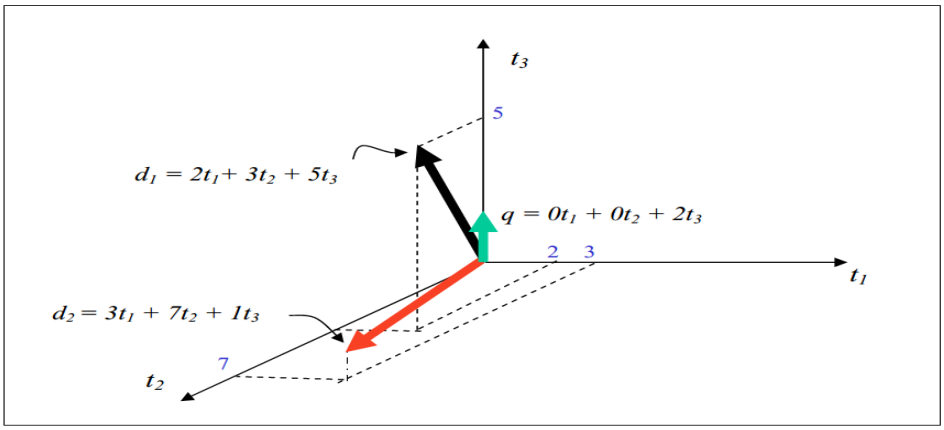


Figura 2.9. Representación gráfica de una consulta *q* junto a dos documentos d_1, d_2 utilizando el modelo vectorial. Fuente: [Raymond, 2005].

El cálculo de la similitud entre los documentos d_1, d_2 y la consulta *q* del ejemplo, se efectuará como sigue:

$$sim(d_1,q) = \frac{2 \cdot 5}{\sqrt{(4+9+25) \cdot (0+0+4)}} = 0.81$$

$$sim(d_2,q) = \frac{2 \cdot 1}{\sqrt{(9+49+1) \cdot (0+0+4)}} = 0.13$$

Teniendo en cuenta que d_1= *(2, 3, 5),* d_2 *= (3, 7, 1) y q = (0, 0, 2).*

De los resultados se deduce que el documento d_1 es bastante más similar a la consulta *q* que el documento d_2, o lo que es lo mismo, que el ángulo θ_1 entre el vector que representa a d_1 y el vector que representa a *q* es menor que el ángulo θ_2 entre el vector que representa a d_2 y el vector que representa a *q*, tal y como puede verse en la **figura 2.10**.

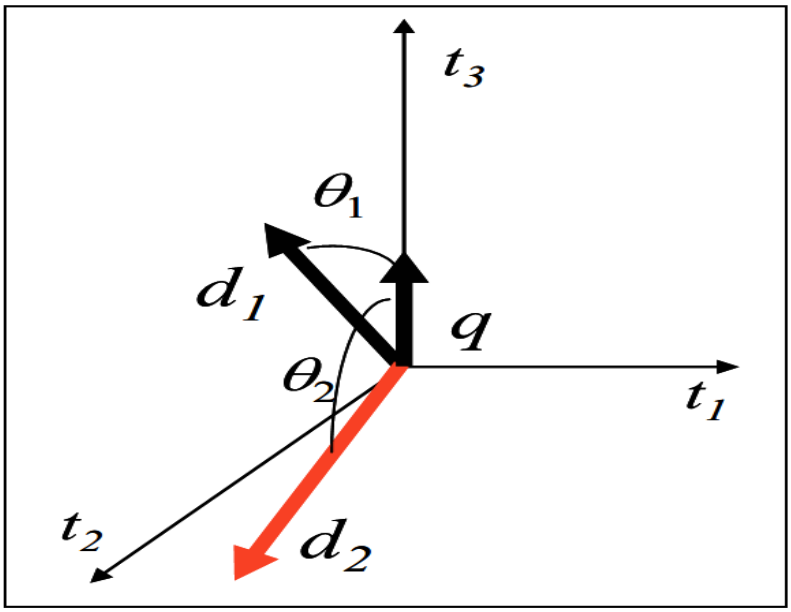


Figura 2.10. Representación gráfica de los ángulos θ_1 y θ_2 entre los vectores de los documentos d_1 y d_2 y la consulta q, para el ejemplo de cálculo de similitud en el modelo vectorial descrito. Fuente: (Raymond, 2005).

Al contar con una medida de similitud como la del coseno entre cada documento (perfil de usuario) y una consulta dada, será posible considerar un umbral en la recuperación de los documentos (perfiles de usuarios), de forma que se consideren relevantes aquellos cuyo valor en la fórmula (5) sea, por ejemplo, mayor o igual a 0.6. De este modo podemos considerar búsquedas no exactas. Los documentos (perfiles de usuarios) pueden entonces presentarse al usuario en un orden decreciente de similitud entre ellos.

2.5.4 Agrupamiento de perfiles de usuarios

La elección de una métrica apropiada influenciará la forma de los grupos, ya que algunos pueden estar cerca unos de otros de acuerdo a una distancia y más lejos de acuerdo a otra.

2.5.4.1 Similitud de perfiles de usuarios

Para el caso de la presente investigación se hará uso de la misma función para el cálculo de la similitud, que la utilizada en la clasificación de los usuarios. La función de similitud del coseno:

Función del coseno:

$$F\ cos(A,B) = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A^2{}_j \cdot \sum_{j=1}^{n} B^2{}_j}}$$

Donde A_j y B_j son, respectivamente, los pesos asociados al término t_j en la representación de los usuarios *A* y *B*.

Una matriz de similitud puede quedar representada simétricamente, donde cada elemento δ_{ij}de *M* representa la similaridad entre el estímulo *i* y el estímulo *j* como se muestra a continuación:

$$M = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & & \delta_{1n} \\ \delta_{21} & \delta_{22} & \delta_{23} & \cdots & \delta_{2n} \\ \delta_{31} & \delta_{32} & \delta_{33} & & \delta_{3n} \\ & \vdots & & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \delta_{n3} & \cdots & \delta_{nn} \end{pmatrix}$$

Figura 2.11. Matriz de similitud de los usuarios

De esta manera queda determinada la matriz de similitud de los usuarios que contiene el sistema, de forma tal que pueden ser identificados los niveles de

compatibilidad entre estos usuarios partiendo de su perfil. Todo esto también brinda la posibilidad de establecer conglomerados de usuarios.

2.5.4.2 Agrupamiento jerárquico para identificar conglomerados de usuarios

Las estrategias jerárquicas (aglomerativas o divisivas) construyen una jerarquía de agrupamientos, representada tradicionalmente por un árbol llamado dendograma (Pascual-González, 2010). En el caso de las técnicas aglomerativas, el dendograma parte generalmente de grupos unitarios, hasta que algún criterio de parada se ejecute, o hasta conseguir el grupo formado por todos los puntos, mientras que las divisivas comienzan generalmente con todos los puntos en un clúster y van dividiendo en cada nivel dos grupos de acuerdo a algún criterio prefijado.

2.5.4.3 Agrupamiento

Suponiendo que la mejor alternativa a partir de la aplicación del Coeficiente de correlación cofenética es el uso de la distancia euclidiana y el método de distancia media (linkage average) para el agrupamiento y su visualización por medio de un dendrograma, procederemos a realizar el cluster jerárquico, el cual según Pascual-González (2010) funciona de la siguiente manera:

1. Empezar con N clústeres (el número inicial de elementos) y una matriz $N \times N$ simétrica de distancias.
2. Dentro de la matriz de distancias, buscar aquella entre los clústeres U y V que sea la menor entre todas, d_{UV}.
3. Juntar los clústeres U *y* V en uno solo. Actualizar la matriz de distancias:

I. Borrando las filas y columnas de los clúster *U y V*.

II. Formando la fila y columna de las distancias del nuevo clúster (*UV*) y el resto de los clústeres.

4. Repetir los pasos (2) y (3) un total de (*N−1*) veces, o sea si todos los puntos están en un mismo clúster, terminar; sino, volver a los pasos (2) y (3).

Para la representación de los usuarios del sistema a través del análisis de clúster jerárquico según los pasos anteriores:

Partiendo de una matriz de distancia o similitud en el caso de la presente investigación se determina la distancia entre sus elementos por medio de algunas de las métricas de distancias disponibles, obteniéndose una matriz simétrica como se describe a continuación:

Sean (*1, 5, 8.5, 7.2, 4.5, 7.8, 6.7, 3.6, 2.2, 2.0)* distancias calculadas y (u_1, u_2, u_3, u_4, u_5) usuarios en el sistema.

	U_1	U_2	U_3	U_4	U_5
U_1	0				
U_2	1	0			
U_3	5	4.5	0		
U_4	8.5	7.8	3.6	0	
U_5	7.2	6.7	2.2	2.0	0

Como punto de partida es considerado cada elemento de la matriz un clúster, por tanto se busca el menor valor, entonces se conforma el primer clúster, donde quedaría identificado como u_{21}, conformándose una nueva matriz con la unión del clúster compuesto por u_2 y u_1 y las distancias de u_{21} au_3, u_4 yu_5.

	U_1	U_2		*distancia*
U_3	5	4.5	(5+4.5)/2	4.75
U_4	8.5	7.8	(8.5+7.8)/2	8.15
U_5	7.2	6.7	(7.2+6.7)/2	6.95

A partir de lo anterior se construye la nueva matriz quedando de la siguiente forma:

	U_{21}	U_3	U_4	U_5
U_{21}	0			
U_3	4.75	0		
U_4	8.15	3.6	0	
U_5	6.95	2.2	2.0	0

Nuevamente elegimos el menor valor de distancia que es entre U_4 y U_5, fusionándolos en un cluster que denominaremos U_{54}calculamos la distancia entre U_{54}, U_{21} y U_3. Entre U_{54} y U_{21} buscamos las distancias entre todos los pares de puntos y calculamos la media.

	U_{54}		
U_{21}		U_4	U_5
	U_1	8.5	7.2
	U_2	7.8	6.7

Siendo la media de los cuatros valores 7.55 quedando la matriz representada de la siguiente forma:

	U_{21}	U_{54}	U_3
U_{21}	0		
U_{54}	7.55	0	
U_3	4.75	2.9	0

Procedemos nuevamente a identificar el valor más pequeño siendo 2.9 y unimos a U_3 con U_{54} como U_{543} quedando finalmente la matriz de distancia representada de la siguiente forma:

	U_{21}	U_{543}
U_{21}	0	
U_{543}	6.62	0

Se comprueba la condición de *N-1* elementos, o sea solo queda representada la distancia 6.62 entre el clúster u_{543} y u_{21}.

Para construir el dendrograma (figura 2.6) que representa a los usuarios del sistema, se resume que:

- Para la distancia 6.62 se tiene ($u_{543} – u_{21}$).
- Para la distancia 2.9 se tiene ($u_{54} – u_3$).
- Para la distancia 2.0 se tiene ($u_4 – u_5$).
- Para la distancia 1.0 se tiene ($u_1 – u_2$).

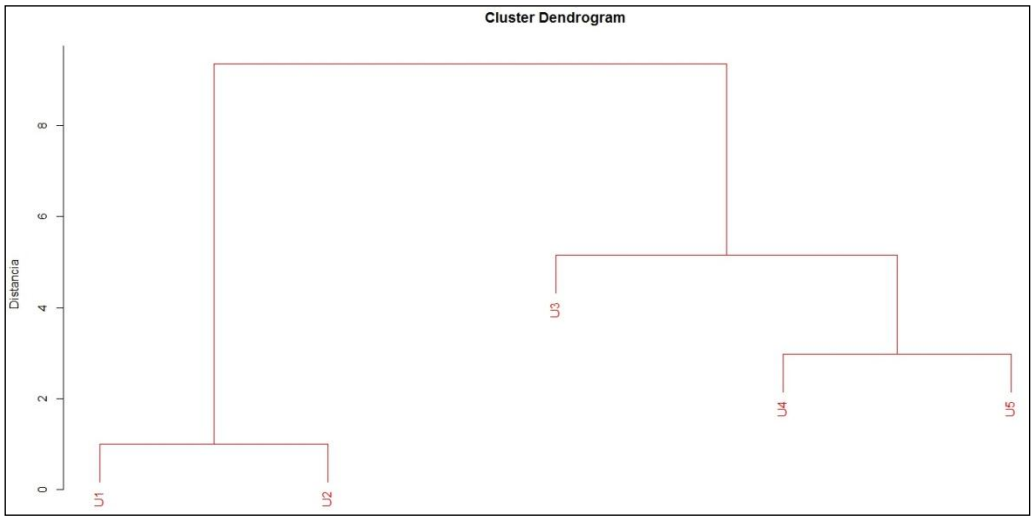


Figura 2.12. Representación de los usuarios del sistema a partir del análisis de clúster jerárquico.

2.5.5 Evaluación del modelo

Para la evaluación inicial del modelo se tomó una muestra real de 15 perfiles de usuario a los que se les aplicaron las etapas antes mencionadas del procedimiento para la conformación y agrupamiento de perfiles de usuario en revistas científicas gestionadas con el Open Journal System. El resultado obtenido de esta evaluación arrojó el árbol de agrupamiento o dendograma para la muestra escogida

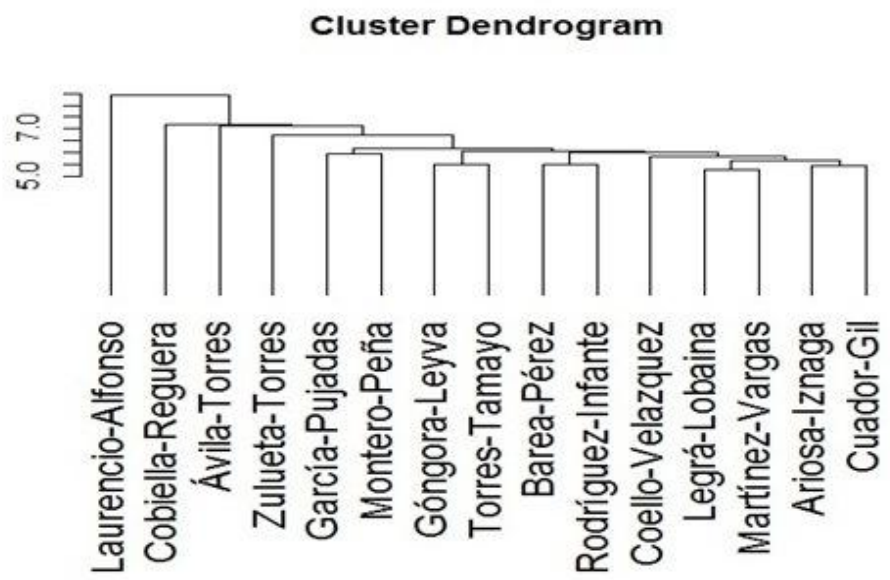


Figura 2.15 Dendograma obtenido

2.6 Evaluación

En este punto, ya se ha completado la mayor parte del proyecto de minería de datos. También se determinó, en la fase de modelado, que los modelos son técnicamente correctos y efectivos en función de los criterios de rendimiento de minería de datos que se definieron previamente. Sin embargo, antes de continuar, deben evaluarse los resultados de los esfuerzos utilizando los criterios de rendimiento del negocio establecidos en el inicio del proyecto. Es la clave para asegurar que la organización puede utilizar los resultados que se han obtenido.

Partiendo del análisis de los objetivos del negocio y cómo influyen los modelos en ellos, se determinó que se cumplen los criterios de éxito propuestos, se logra la creación de los perfiles de usuario y su clasificación según un tema dado, se ordena de forma jerárquica los posibles evaluadores, lo que trae

consigo que se aprecien mejoras en el proceso de selección de expertos, lo que contribuirá en gran medida a la toma de decisiones del consejo editorial de una revista científica gestionada con OJS.

A raíz de la precisión y relevancia de los resultados de modelado, se determinó que para el cumplimiento total de los objetivos del negocio, que se definió como: la creación de una aplicación informática que permita conocer la similitud existente entre los perfiles de usuarios y la clasificación de estos, generados de la información contenida en una revista gestionada con OJS, con la finalidad de favorecer el proceso de identificación de posibles árbitros de artículos científicos; las tareas de evaluación se van a retomar en el siguiente capítulo que detallará el proceso de desarrollo de dicha aplicación, dado que es aconsejable probar el modelo en un problema real, o sea, es conveniente evaluar cómo interactúa la aplicación y la implementación de los modelos, en virtud de comparar los resultados obtenidos y comprobar su calidad. El tiempo y las restricciones lo permiten.

2.7 Implementación

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, o ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso de negocio de la organización. La metodología CRISP-DM no impone una manera a seguir en esta etapa, o sea que la implementación se realiza de acuerdo a las necesidades de la organización en la que se trabaja y así como los objetivos que se propusieran al inicio del proyecto de minería de datos. En el caso particular de esta investigación, se ha decidido en esta última etapa del

proyecto, desarrollar una aplicación web para la clasificación y agrupamiento de los perfiles de usuarios de las revistas científicas gestionadas con OJS, además de la validación funcional de dicha aplicación. En este epígrafe se mostrarán algunos fragmentos de código de las principales funcionalidades del sistema y las pruebas funcionales a cada una de ellas.

Como se mencionó con anterioridad la aplicación se desarrolló en el lenguaje de programación R, haciendo uso de la muy popular librería *shiny* para crear aplicaciones web. Esta divide la aplicación en dos partes o archivos, el *ui.R* que se encarga de la interfaz de usuario y la parte *server.R* cuyo trabajo es el procesamiento de los datos.

Implementación de la interfaz de usuario

```
library(shiny)
shinyUI(fluidPage(
  titlePanel("Sis_CAP "),
   sidebarLayout(
    sidebarPanel(
      fileInput('file1', 'Leer desde otra fuente de datos',
                accept=c('text/csv',
                         'text/comma-separated-values,text/plain',
                         '.csv')),
      tags$hr(),
      checkboxInput('header', 'Cabecera', TRUE),
      radioButtons('sep', 'Separador',
                   c(Coma=',',
                     Punto_y_coma=';',
                     Tabulador='\t'),
                   ','),
      radioButtons('quote', 'Comillas',
                   c(None='',
                     'Doble_Comilla'='"',
                     'Comilla_Simple'="'"),
                   '"'),
      br(),
      actionButton("perfiles", "Ver Perfiles de Usuario"),
      br(),
      br(),
      textInput("palabras", "Buscar Investigadores similares"),
      helpText("Es necesario que introduzca las palabras claves del articulo que se analiza. "),
      submitButton("Buscar")
      ),
    mainPanel(
      tabsetPanel(type = "tabs",
        tabPanel("Lectura del archivo", tableOutput("contents")),
        tabPanel("Agrupamiento", plotOutput("dendograma")),
        tabPanel("Investigadores similares", tableOutput("resultb")),
        tabPanel("Agrupamiento de investigadores", plotOutput("investig")),
        tabPanel("Perfiles", tableOutput("perfiles"))
```

Implementación de lectura de archivo externo

```
shinyServer(function(input, output) {

  #LEER DE OTRO ARCHIVO

  output$contents <- renderTable({
    inFile <- input$file1

    if (is.null(inFile))
      return(NULL)

    read.csv(inFile$datapath, header=input$header, sep=input$sep,
             quote=input$quote)

  })
```

Implementación de la normalización de datos

```
 #3- Operaciones sobre el dataset
txtdataset <- subset(txtrepo, datapath!="\\N") #Eliminamos columnas vacÃas

 #Creamos el corpus
 mycorpus <- Corpus(VectorSource(txtdataset$setting_value), readerControl = list(reader = readPlain,

 #4- Operaciones sobre el corpus

 #EliminaciÃ³n de nÃºmeros
 mycorpus <- tm_map(mycorpus, content_transformer(removeNumbers))

 #EliminaciÃ³n de signos de puntuaciÃ³n
 mycorpus <- tm_map(mycorpus, content_transformer(removePunctuation))

 #Convierte a minÃºsculas todas las palabras
 mycorpus <- tm_map(mycorpus, content_transformer(tolower))

 #Elimina los stopwords del idioma espaÃ±ol Ej. de, la, el, en
 mycorpus <- tm_map(mycorpus, content_transformer(removeWords), stopwords("spanish"))

 #Elimina las acentuaciones
 mycorpus <- tm_map(mycorpus, content_transformer(stemDocument), language="spanish")

 #Aplicamos stemming (reducir las palabras a su raÃz, como de "clasificar" a "clasific").
 mycorpus <- tm_map(mycorpus, stemDocument, language="spanish")

 #Elimina los espacios en blanco sobrantes entre cada palabra o sea espacios dobles
 mycorpus <- tm_map(mycorpus, content_transformer(stripWhitespace))
```

Implementación de la representación del espacio vectorial y selección de rasgos

```
#5- RepresentaciÃ³n VSM y selecciÃ³n de rasgos

#Creamos lo que se conoce como una matriz de tÃ©rminos del documento.
#asignandole pÃ©so a los tÃ©rminos por el mÃ©todo por defecto(binary)
mycorpus.vsm.dtm<-DocumentTermMatrix(mycorpus, control=list(wordLengths=c(3,Inf), weighting=weightBin))

#Asignandole peso a los tÃ©rminos por el mÃ©todo(weightTfIdf)
mycorpus.vsm.dtm.tdidf<-DocumentTermMatrix(mycorpus, control=list(wordLengths=c(3,Inf), weighting=weightTfIdf

mycorpus.vsm.dtm<-DocumentTermMatrix(mycorpus, control=list(wordLengths=c(3,Inf), weighting=weightTfIdf))

#Covertimos el formato de los datos del corpus a una matriz
mycorpus.vsm.dtm.data <- as.matrix(mycorpus.vsm.dtm)
```

Implementación del cálculo de matriz de similaridad y análisis de conglomerados jerárquicos

```
#Calculamos la matriz de similitud por el mÃ©todo del coseno
similitudCosine<- as.matrix(simil(mycorpus.vsm.dtm.data, method = "cosine"))

#Calculamos la matriz de distancia por el mÃ©todo del coseno
distanciaCosine<- as.matrix(dist(mycorpus.vsm.dtm.data, method = "cosine"))

#Calculamos la matriz de distancia o disimilitud Euclidean
distanciaEuclidean<- as.matrix(dist(mycorpus.vsm.dtm.data, method = "Euclidean"))

#7- AnÃ¡lisis de conglomerados
a<-plot(hclust(dist(distanciaEuclidean), method="average"),labels=row.names(distanciaEuclidean),ylab
```

2.7.1 Validación funcional.

Durante todo el ciclo de elaboración del software es preciso velar, controlar y garantizar su correcta calidad, haciendo posible el cumplimiento de los requerimientos que precisamente satisfacen las necesidades del cliente. Este aspecto debe estar presente de forma paralela desde la concepción del producto hasta la fase de producción del mismo. Para verificar lo antes mencionado se recurre a la realización de las pruebas de software.

2.7.1.1 Pruebas de software.

Las pruebas de software es un concepto que a menudo es conocido como verificación y validación. Integra las técnicas de diseño de casos de prueba en una serie de pasos bien planificados que dan como resultado una correcta construcción del software. Entre algunas de las técnicas que se llevan a cabo para el proceso de prueba se encuentran las técnicas de caja negra y de caja blanca. (Fernández E., 2011)

2.7.1.2 Pruebas de caja negra.

Prueba de caja negra es aquel elemento que se estudia desde el punto de vista de las entradas que recibe y las salidas o respuestas que produce, sin tener en cuenta su funcionamiento interno. En otras palabras, de una caja negra solamente interesará su forma de interactuar con el medio que le rodea (en ocasiones, otros elementos que también podrían ser cajas negras) entendiendo qué es lo que hace, pero sin dar importancia a cómo lo hace. Por tanto, de una caja negra deben estar muy bien definidas sus entradas y salidas, es decir, su interfaz; en cambio, no se precisa definir ni conocer los detalles internos de su funcionamiento.

Cuando se habla de un software, la prueba de caja negra se refiere a las pruebas que se llevan a cabo sobre la interfaz del mismo. Los métodos de prueba de la caja negra se centran en los requisitos funcionales del mismo e intentan encontrar errores de las siguientes categorías: (Fernández E. , 2011).

1. Funciones incorrectas o ausentes.
2. Errores de interfaz.
3. Errores en estructuras de datos o en acceso a bases de datos externas
4. Errores de rendimiento.
5. Errores de inicialización y terminación.

2.7.1.3 Pruebas de la aplicación web

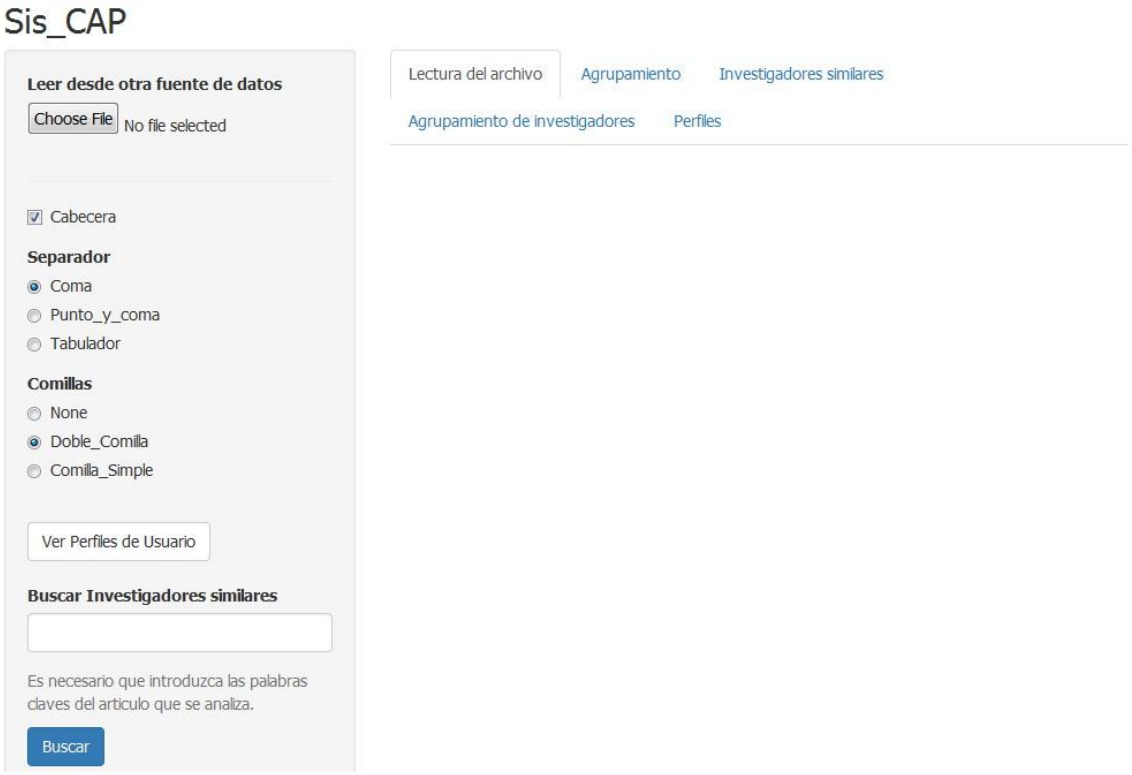


Figura 2.16 Interfaz principal

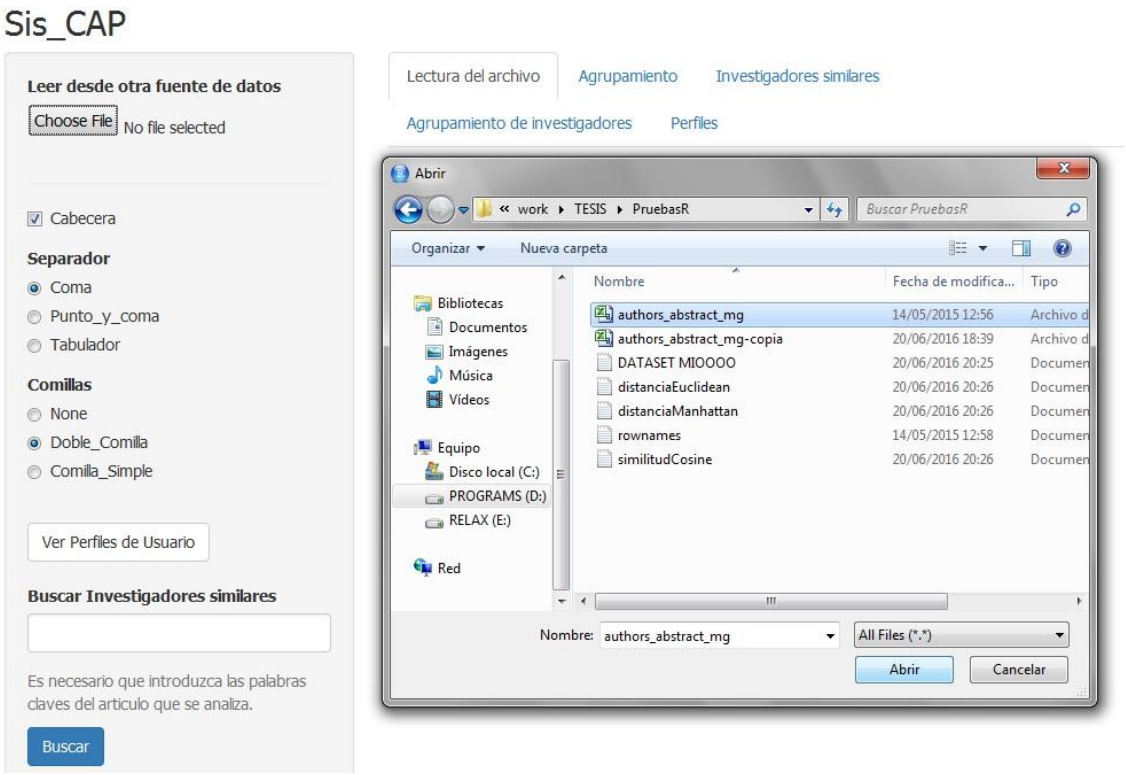


Figura 2.17 Cargar datos de archivo externo

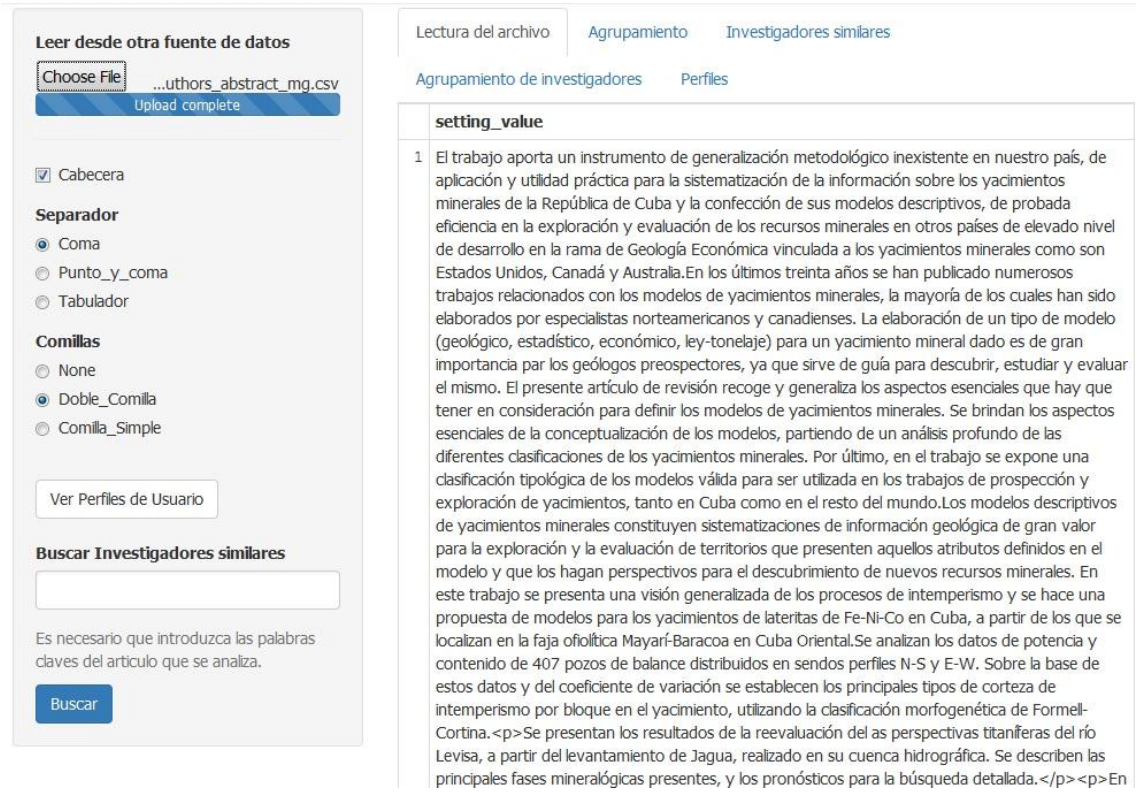


Figura 2.18 Carga de archivo externo realizada con éxito

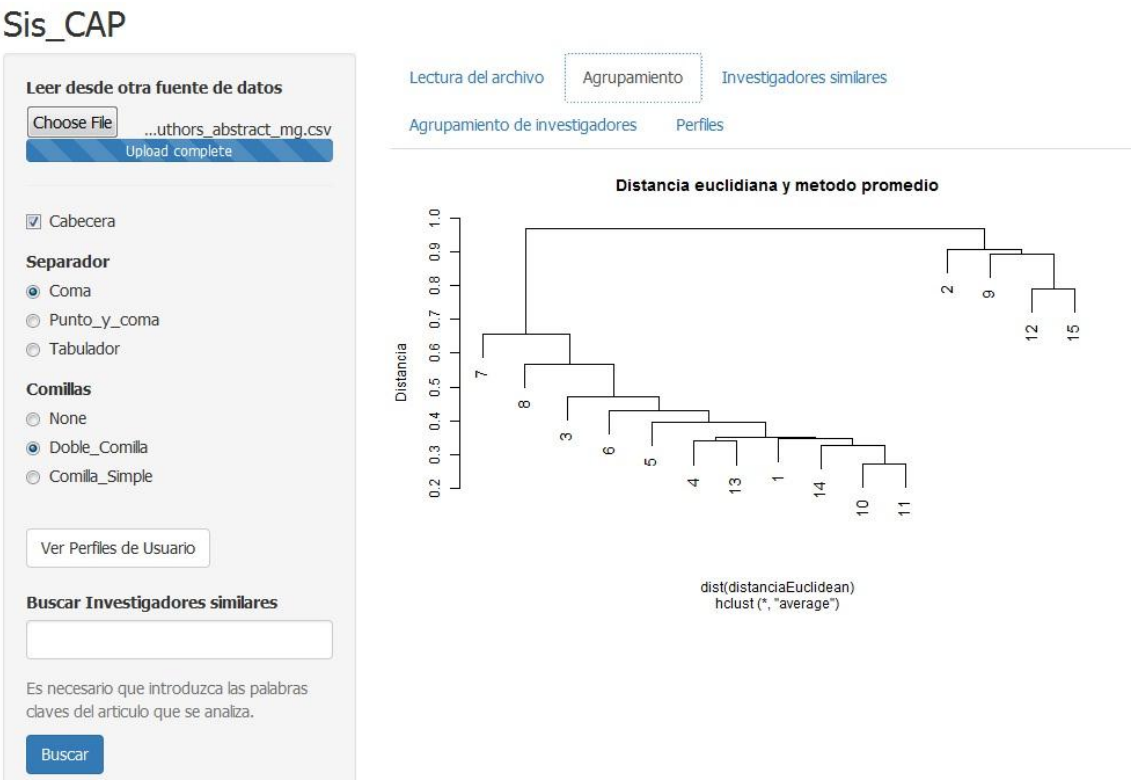


Figura 2.19 Agrupamiento de la lectura del archivo externo realizado con éxito

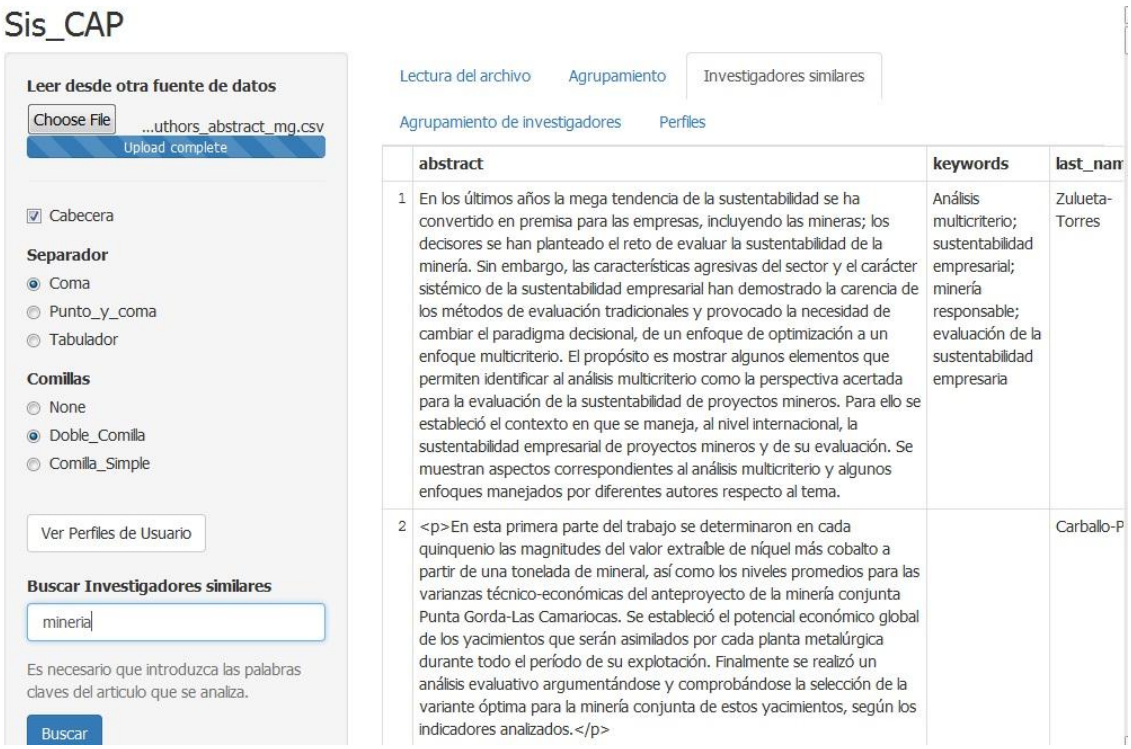


Figura 2.20 Clasificación de PU según tema realizado con éxito

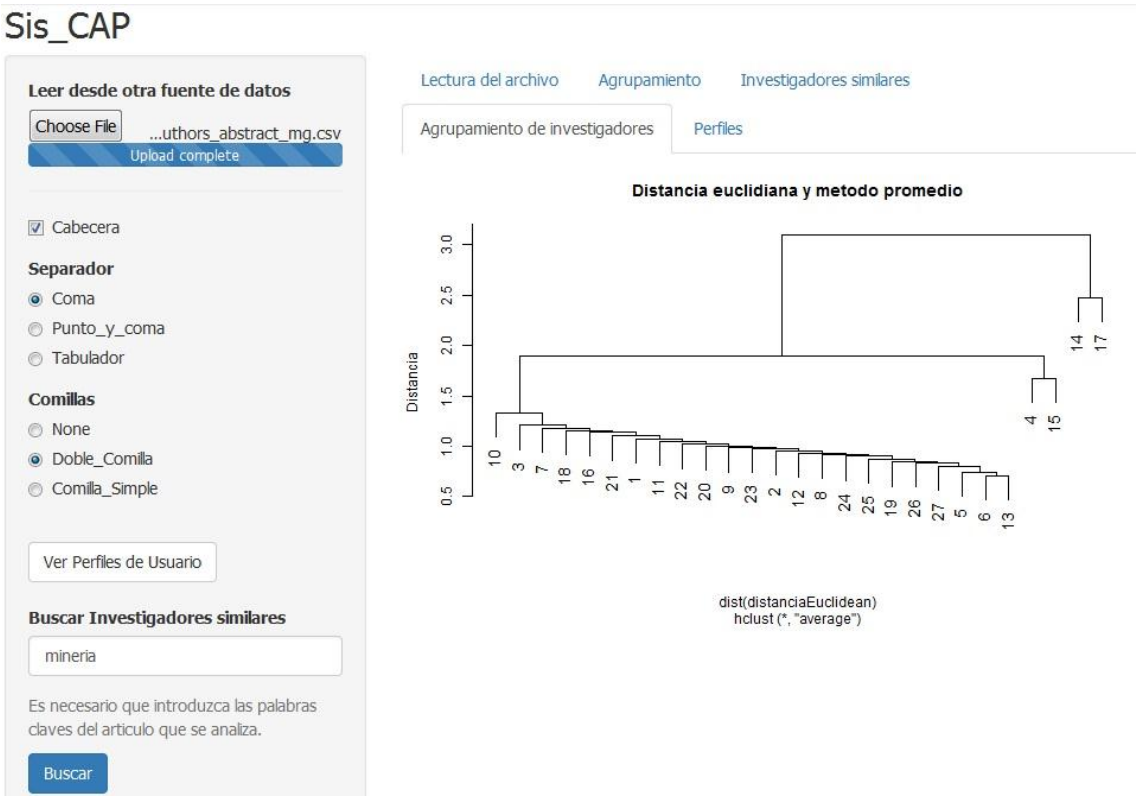


Figura 2.21 Agrupamiento de PU según tema realizado con éxito

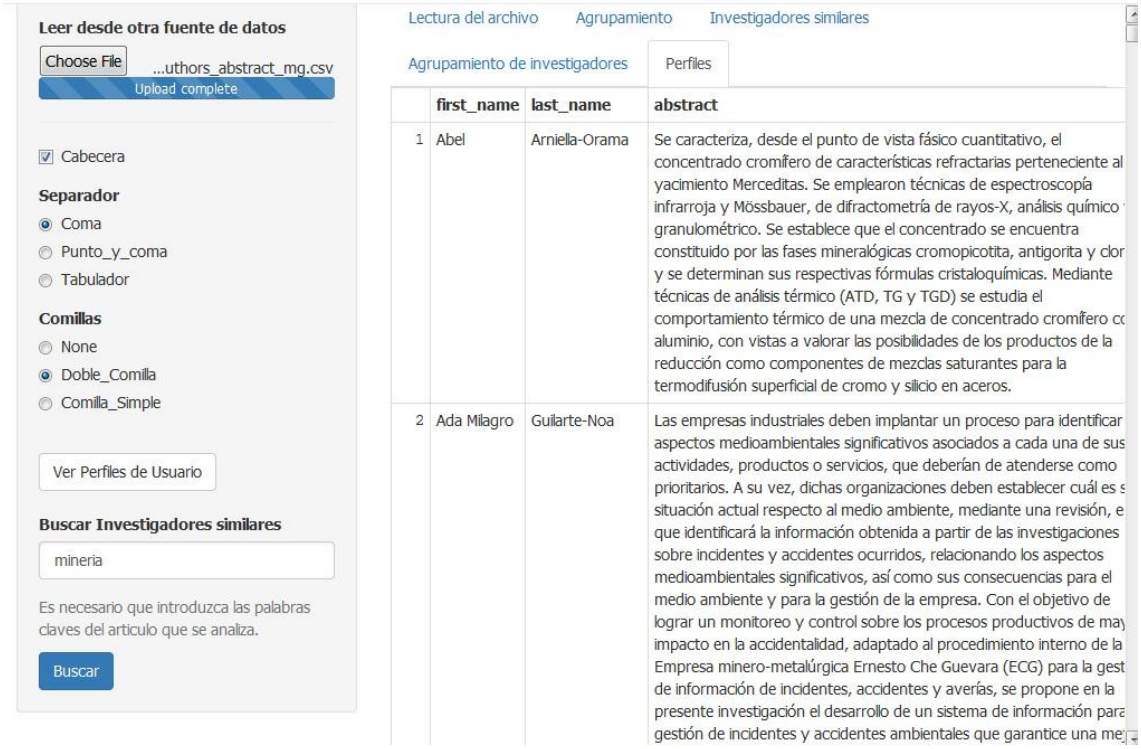


Figura 2.22 Conexión con BD establecida con éxito y muestra de todos los PU del sistema

2.8 Conclusiones del capítulo

En este capítulo se documentó la aplicación de las cuatro primeras fases de la metodología CRIS-DM, donde se establecieron los criterios de éxito del negocio así como los de minería de datos, se observaron resultados en las tareas de transformación de los datos que culminó con la creación de los perfiles de usuario.

La fase de modelado arrojó los modelos necesarios para cumplimentar los objetivos de minería de datos, el uso de la distancia euclidiana y el método de distancia media (*linkage average*) dieron pie al agrupamiento y su visualización por medio de un dendrograma, lo que constituyó un éxito de la evaluación inicial del modelo. La fase de implementación culminó con el proceso de desarrollo de minería de datos, cumpliendo así con el objetivo general propuesto al inicio de esta investigación: el desarrollo de una aplicación web para la clasificación y agrupamiento de perfiles de usuarios de revistas científicas gestionadas con OJS, se realizó la validación funcional de la aplicación utilizando pruebas de software, en este caso la prueba de caja negra.

Capítulo 3. Estudio de Factibilidad.

Introducción.

Después de definir la problemática presente e identificar las causas que ameritan la informatización de los procesos de la conformación y agrupamiento de perfiles de usuario y selección de evaluadores en revistas científicas gestionadas por el Open Journal System, es pertinente realizar un estudio de factibilidad para determinar la infraestructura tecnológica y la capacidad técnica que implica la implantación del sistema en cuestión, así como los costos, beneficios y el grado de aceptación que la propuesta genera.

3.1 Factibilidad Técnica.

La Factibilidad Técnica consiste en realizar una evaluación de la tecnología existente en la organización. Este estudio estuvo destinado a recolectar información sobre los componentes técnicos que se poseen y la posibilidad de hacer uso de los mismos en el desarrollo e implementación del sistema propuesto y de ser necesario, los requerimientos tecnológicos que deben ser adquiridos para el desarrollo y puesta en marcha del sistema en cuestión. De acuerdo a la tecnología necesaria para la implantación de la aplicación para la conformación y agrupamiento de perfiles de usuarios en revistas científicas gestionadas con OJS, se evaluaron enfoques: Hardware y Software.

3.1.1 Hardware

El servidor donde debe estar instalado el sistema propuesto, debe cumplir con los siguientes requerimientos mínimos:

- Procesador Pentium 1.5 Ghz.
- 1 GB de Memoria RAM
- Disco Duro de 40 GB.

Evaluando el hardware existente y tomando en cuenta la configuración mínima necesaria, no se requirió realizar inversión inicial para la adquisición de nuevos equipos, ni tampoco para mejorar o actualizar los equipos existentes. A continuación se muestran las características de red interna con que cuenta actualmente la editorial de la Revista Minería y Geología (revista científica del Instituto Superior Minero Metalúrgico de Moa que se gestiona con OJS):

- Servidor: Profesional HP Proliant ML 350, 2.8 Ghz de velocidad y 2GB RAM.
- Las estaciones de Trabajo: Procesador Pentium 4+, 1+GB en Memoria RAM, Disco Duro 160+GB.
- Concentradores de Puertos RJ-45.

Todas las estaciones de trabajo están conectadas al servidor a través de una red utilizando cable par trenzado. Esta configuración permite que los equipos instalados puedan interactuar con la aplicación web.

3.1.2 Software.

La editorial de la Revista Minería y Geología cuenta con todas las aplicaciones que se emplearan para el correcto funcionamiento del sistema, lo cual no amerita inversión alguna para la adquisición de los mismos. Las estaciones de trabajo, operarán en ambientes MS Windows y GNU/Linux, el servidor se encuentra instalado sobre una plataforma GNU/Linux. Para el uso general de las estaciones en actividades diversas se debe poseer los siguientes requisitos mínimos de software disponibles en el mercado actualmente:

Propiedades	Nombre
Sistemas Operativos	GNU/Linux, Microsoft Windows
Navegador Web	Mozilla Firefox, Internet Explorer, Chrome

Tabla 3.1. Requisitos mínimos de software

Como resultado de este estudio técnico se determinó que la Institución posee la infraestructura tecnológica (Hardware y Software) necesaria para el desarrollo y puesta en funcionamiento del sistema propuesto.

3.2 Factibilidad Económica.

3.2.1 Evaluación de Costo-Beneficio.

Para estudiar la factibilidad de este proyecto se utilizará la Metodología Costo Efectividad (Beneficio), la cual plantea que la conveniencia de la ejecución de un proyecto se determina por la observación conjunta de dos factores:

1. El costo, que involucra la implementación de la solución informática, adquisición y puesta en marcha del sistema hardware/software y los costos de operación asociados.
2. La efectividad, que se entiende como la capacidad del proyecto para satisfacer la necesidad, solucionar el problema o lograr el objetivo para el cual se ideó, es decir, un proyecto será más o menos efectivo con relación al mayor o menor cumplimiento que alcance en la finalidad para la cual fue ideado (costo por unidad de cumplimiento del objetivo). Este puede estar justificado por los beneficios tanto tangibles como intangibles que origina el mismo. En este proceso, se necesita de una selección adecuada de los elementos más convenientes para su evaluación.

3.2.2 Efectos Económicos.

Pueden clasificarse como:

- ✓ Efectos directos.
- ✓ Efectos indirectos.
- ✓ Efectos externos.
- ✓ Intangibles.

3.2.2.1 Efectos directos.

Positivos:

- ✓ Los perfiles de usuario estarán agrupados en orden jerárquico respecto a un tema en particular, lo que facilitará un mejor manejo de los datos, así como mayores potencialidades para la selección.

- ✓ Por su parte el Editor de la revista puede acceder desde cualquier lugar mediante un navegador web al sistema que permite el análisis de los datos y por tanto la identificación de expertos.

Negativos:

- ✓ Para usar la aplicación es vital el uso de un ordenador conectado a la red, aparejado a los gastos de consumo de energía eléctrica.

3.2.2.2 Efecto Indirecto.

Los efectos económicos observados que pudiera repercutir sobre otros mercados no son perceptibles, aunque este proyecto no está construido con la finalidad de comercializarse.

3.2.2.3 Externalidades.

Se contará con una herramienta disponible que facilitará la conformación, clasificación y agrupamiento de perfiles de usuario para la identificación de posibles árbitros en revistas gestionadas con el OJS, optimizando el tiempo y recursos.

3.2.2.4 Intangibles.

En la valoración económica siempre hay elementos perceptibles por una comunidad como perjuicio o beneficio, pero al momento de ponderar en unidades

monetarias esto resulta difícil o prácticamente imposible. A fin de medir con precisión los efectos, deberán considerarse dos situaciones:

Situación sin proyecto

La gestión de artículos científicos, empieza desde el momento en que el autor hace llegar el artículo al editor, quien revisa su adecuación al perfil, a las normas editoriales y a la estructura metodológica del artículo científico; en caso de cumplir con las tres pautas, se pone en cola para asignarlo a un editor de sección o directamente a revisores que evaluarán la calidad científica del trabajo y comentan otros aspectos formales que ayuden al autor a hacer publicable el trabajo; en caso contrario se informa al autor de los problemas del artículo para mejorarlo, o simplemente se rechaza. Una vez que los árbitros consideren que ya el artículo puede publicarse, pasa a manos de los editores de corrección que revisan la gramática y el estilo y se corrige el escrito para su posterior maquetación. El maquetador prepara las maquetas o pruebas de galera, las galeradas finales se preparan en ficheros HTML, PDF. Una vez terminado el proceso editorial, se publica en la web.

El proceso para la elección y el reclutamiento de un revisor en una revista científica puede ser en ocasiones un proceso engorroso por diversas razones, entre las que resalta que el editor debe tener un conocimiento previo de posibles expertos y su estructura jerárquica con sus análogos respecto a una temática en particular, lo que trae consigo que las posibilidades de que un editor logre reclutar verdaderos expertos en una materia determinada sean bajas.

Situación con proyecto

Para el proceso de reclutamiento de expertos, el editor es el encargado de introducir en el sistema las palabras claves del artículo que está siendo analizado, el sistema le mostrará ordenado de forma jerárquica los autores de artículos similares que pudieran servir de expertos.

3.2.3 Beneficios y Costos Intangibles en el proyecto.

Costos:

- Resistencia al cambio.

Beneficios:

- Mayor comodidad para el consejo editorial de las revistas científicas.
- Mayor información visual.
- Mejora la calidad del estudio de posibles evaluadores de artículos.
- Reduce el gasto de materiales de oficina utilizados en estos procesos.

3.2.4 Ficha de Costo.

Para determinar el costo económico del proyecto se utilizará el procedimiento para elaborar Una Ficha De Costo de un Producto Informático [Dra. Ana María Gracia Pérez, UCLV]. Para la elaboración de la ficha se consideran los siguientes elementos de costo, desglosados en moneda libremente convertible y moneda nacional.

Costo en Moneda Libremente Convertible:

Costos Directos:

1. Compra de equipos de cómputo: No procede.

2. Alquiler de equipos de cómputo: No procede.

3. Compra de licencia de Software: No procede.

4. Materiales directos: $0.00.

5. Gasto por consumo de energía eléctrica: No procede

Subtotal: $0.00

Costos Indirectos:

1. Formación del personal que elabora el proyecto: No procede.

2. Gastos en llamadas telefónicas: No procede.

3. Gastos para el mantenimiento del centro: No procede.

4. Know How: No procede.

5. Gastos en representación: No procede.

Subtotal: $0.00

Gastos de distribución y venta.

1. Participación en ferias o exposiciones: No procede.

2. Gastos en transportación: No procede.

3. Compra de materiales de propagandas: No procede.

Subtotal: $0.00

Depreciación de equipo de cómputo.

1. Valor inicial del equipo: 300

2. Valor de depreciación anual: 25 % del valor inicial de compra. (En este caso el valor de depreciación anual es 75)

3. Valor de depreciación en 1 mes de proyecto: 6.25. (75 /12 meses)

4. Valor de depreciación por tiempo completo del proyecto: 37.50

Subtotal: $37.50

Total de Costo en Moneda Libremente Convertible: $37.50

Costo en Moneda Nacional.

Costos Directos.

1. Salario del personal que laborará en el proyecto (en este caso se refiere a estipendio de estudiante universitario de cuarto año de la carrera): $75.00 ($450.00 por 6 meses de trabajo).

2. El 5% del total de gastos por salarios se dedica a la seguridad social: No procede.

3. El 0.09% de salario total, por concepto de vacaciones a acumular: No procede.

4. Gasto por consumo de energía eléctrica: $ 6.00 ($ 36.00 por seis meses Nota: Este valor es un número aproximado, debido a que es imposible proporcionar un valor exacto por medirse el consumo).

5. Gastos en llamadas telefónicas: No procede.

6. Gastos administrativos: No procede.

Costos Indirectos.

1. Know How: No procede.

Subtotal: $ 0.00

Gasto en Distribución y Ventas Subtotal: $ 0.00

Total de Costo en Moneda Nacional: $486.00

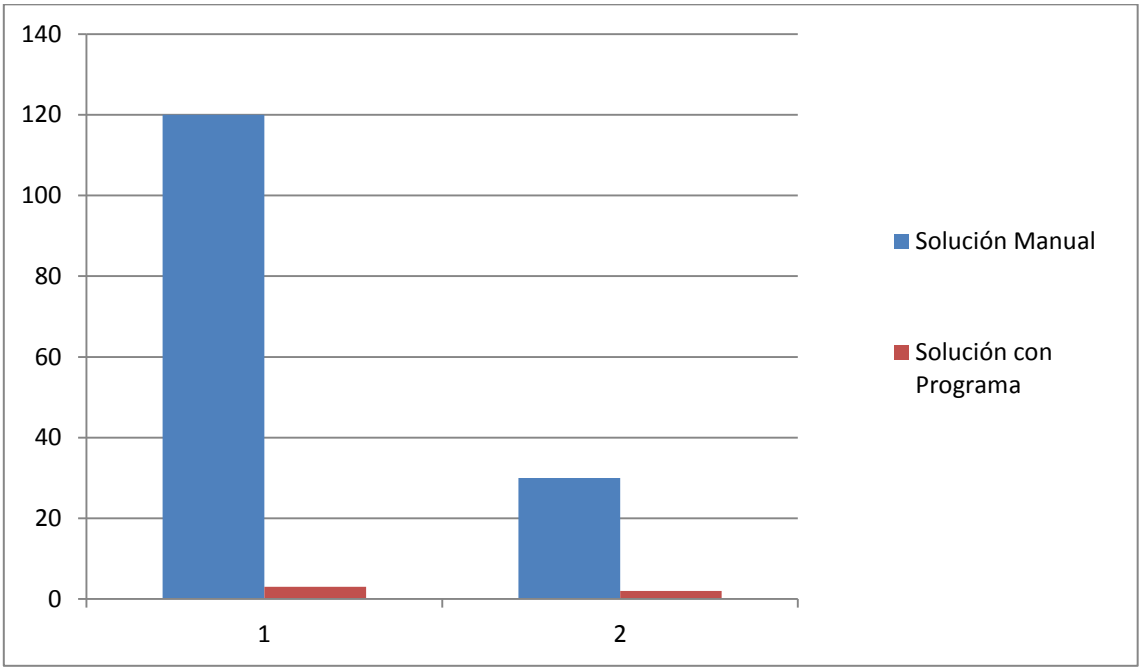
La evaluación económica se efectúa conjuntamente con evaluación técnica del proyecto, que consiste en cerciorarse de la factibilidad técnica del mismo. En el análisis de la Factibilidad Técnica del proyecto, se pudo apreciar que se cuenta con la disponibilidad de hardware/software por lo que se puede inferir que el proyecto es factible técnicamente y no necesita de inversión alguna para su realización, por tanto la decisión de inversión recae en la evaluación económica. Como se hizo referencia anteriormente, la técnica seleccionada para evaluar la factibilidad del proyecto es la Metodología Costo-Efectividad. Dentro de esta metodología la técnica de punto de equilibrio aplicable a proyectos donde los beneficios tangibles no son evidentes, el análisis se basa exclusivamente en los costos. Para esta técnica es imprescindible definir una variable discreta que haga variar los costos. Teniendo en cuenta que el costo para este proyecto es despreciable, tomaremos como costo el tiempo en minutos empleado para realizar la selección de evaluadores en revistas científicas gestionadas por el Open Journal System. Este se divide en 2 pasos:

Valores de las Variables (Solución Manual):

1. El editor tendrá que hacer una búsqueda en los directorios de archivos o en la propia base de datos para encontrar coincidencias en las palabras claves de los artículos. (120 min)
2. El editor deberá ordenar de forma jerárquica los autores con los cuales se establecieron coincidencias. (30 min)

Valores de la variable (Solución con el programa):

1. El sistema le permitirá al Editor introducir las palabras claves del artículo que se analiza. (3 min)
2. El sistema permitirá visualizar los posibles expertos en orden jerárquico, luego de haberlos clasificado y agrupado. (2 min).



Gráfica 3.1. Comparación de solución manual y solución con programa

Teniendo en cuenta los resultados reflejados en las gráficas para cada uno de los casos queda demostrada la factibilidad del sistema, así como la comprobación de la idea a defender planteada en capítulos anteriores y que enuncia lo siguiente: La utilización de la aplicación web para la selección de posibles evaluadores de artículos científicos en revistas gestionadas con el OJS facilitará mayor agilización de este proceso.

3.3 Conclusiones de Capítulo.

En este capítulo se realizó el estudio de factibilidad mediante la Metodología Costo Efectividad (Beneficio), se analizaron los efectos económicos y técnicos necesarios para la realización del software, los beneficios y costos intangibles, además se calculó el costo de ejecución del proyecto mediante la ficha de costo arrojando como resultado $ 37.50 CUC y $ 486.00 MN. Quedó demostrado además que la utilización del software logra agilizar el proceso de selección de expertos en comparación con la solución manual. Con todo este análisis se logró demostrar la factibilidad del proyecto.

Conclusiones

A lo largo de esta investigación, el objetivo general propuesto así como las tareas de investigación fueron cumplidos, proyectándose los siguientes resultados:

- Se optó por la elección y aplicación de la Metodología para proyectos de minería de datos CRISP-DM, ya que se adapta a las necesidades específicas del entorno de la organización. Se identificaron los objetivos, metas y criterios de éxito tanto del negocio como de la propia minería de datos, dejando claras las funcionalidades que debía cumplir el sistema. Se determinaron los pasos a seguir para la implementación del sistema y se aclararon los principales conceptos que debían ser dominados, de manera que se garantizara la correcta implementación de la solución.

- El estudio de factibilidad realizado siguiendo la metodología Costo Efectividad arrojó como resultado los efectos económicos y beneficios, así como el costo de ejecución del proyecto, siendo este $37.50 CUC. y $

486.00 MN demostrándose que es factible el proyecto, tanto económicamente como por los beneficios que aportará su utilización.

- Una vez culminada la implementación se validó la solución mediante la realización de varias pruebas enfocadas principalmente a la funcionalidad del sistema. Gracias a la combinación de varias técnicas de clasificación y agrupamiento, estas pruebas arrojaron resultados satisfactorios en cuanto a la calidad del agrupamiento aunque el cual podría ser mejorado a costa de un estudio más profundo en el proceso de selección de rasgos. También se ven resultados favorables en cuanto al tiempo de procesamiento de los datos, lo que sin dudas garantiza mejoras en el proceso de selección de árbitros.

Bibliografía

ABONY, J. y BALAZS, F. Cluster Analysis for Data Mining and System Identification. Basel: BirkhäuserVerlag AG, 2007.

ARABIE, P. y HUBERT, L. J. An Overview of Combinatorial Data Analysis. En ARABIE, P…. [et.al.], (eds.), Clustering and Classification). New Jersey: World Scientific Publishing, 1996. p. 8-17

ARCO-GARCÍA, L. … [et.al.]. (2007). CorpusMiner 1.0: Herramienta para el agrupamiento de documentos. Revista Cubana de Ciencias Informáticas, 1(2): 18-31.

BARRERA-FERNÁNDEZ, M. *Implantación de un Sistema informático de gestión y publicación para la revista Minería y Geología*. Tesis de Ingeniería en Informática. Departamento de Informática. ISMMM, 2011.

CABRERA, J. A. … [et. al.]. El papel de los expertos en ciencia y tecnología. Revista Madrid, (14): [en línea]. [Consultado: 2016-03-10] Disponible en: http://www.madrimasd.org/revista/revista14/tribuna/tribunas2.asp

CIMIANO, P.; HOTHO, A. y STAAB, S. (2004). Comparing Conceptual, Divise and Agglomerative Clustering for Learning Taxonomies from Text. En LÓPEZ DE MÁNTARAS, R. y SAITTA, L.. (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence* .Valencia: ECAI, 2004).P. 435-439).

CRISP-DM (2007). [en línea]. [Consultado: 2016-02-15]. Disponible en: http://www.crisp-dm.org

DING, W. …[et.al.]. Towards region discovery in spatial datasets.En WASHIO, T.; INOKUCHI.; SUZUKI, E. y TING, K.M. (Eds.). *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD'08).*Osaka: Springer-Verlag, 2008.

DIXON, M. *An Overview of Document Mining Technology*.http://citeseer.ist.psu.edu/ dixon97overview. 1997.

EVERITT, B.; LANDAU, S.; LEESE, M. y STAHL, D. *Cluster Analysis*. Sussex: John Wiley & Sons, Inc, 2011.

FAKES, W.B. *Baeza-Yates Information Retrieval, Data Structures and Algorithms*.Prentice: Hall. 1992.

FERNÁNDEZ, E. *Sistema para el Análisis de Sensibilidad en la plataforma BioSyS.*2011.

FORMAN, G.. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, (3): 1289-1305.

HAN, J. KAMBER, M. Y PEI, J. *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann Publishers Inc, 2012.

HAN, J., KAMBER, M. Y PEI, J. *Data Mining: Concepts Techniques*. San Francisco: Morgan Kaufmann Publishers Inc, 2006.

HASTIE, T.; TIBSHIRANI, R. y FRIEDMAN, J. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag, 2009.

JERIA, V. *Ciencias y tecnologías multidisciplinarias de la Información*. 2007

KARLSSON, C. *Handbook of Research on Cluster Analysis*. London: Edward ElgarPublising Limited, 2008.

KORFHAGE, R.R. *Information storage and retrieval*. New York: wiley, 1977.

LANQUILLON, C. *Enhacing Text Classification to Improve Information Filtering*. Tesis Doctoral. Universidad de Magdeburgo, Alemania, 2001.

LÓPEZ-HERRERA, A.G. *Modelos de Sistemas de Recuperación de Información Documental basados en Información Lingüística difusa*. Granada: [s.n.], 2006.

Microsoft (2007). [en línea]. [Consultado: 2016-02-15]. Disponible en:http://technet.microsoft.com/es-es/library/ms174861.aspx

MANNING, C.; RAGHAVAN, P. y SCHÜTZE, H. *Introduction to Information Retrieval*.New York: Cambridge University Press, 2008.

MOOI, E. y SARDTEDT, M. Cluster Analysis. En MOOI, E. Y MARKO, S. (Edits.). *A Concice Guide to Market Research.The Process, Data, and Methods Usuing IBM SPSS Statistics*. Munich: Springer- Verlag, 2011. P. 9.

NARANJO-VÉLEZ, E.; ÁLVAREZ ZAPATA, D. *Desarrollo de habilidades informativas: una forma de animar a leer.* Medellín: Universidad de Antioquia, Escuela Interamericana de Bibliotecología, 2003. 27 p.

PASCUAL-GONZÁLEZ. *Algoritmos de Agrupamiento basados en densidad y Validación de clusters.* Inédita Tesis Doctoral. Universitat Jaume I, 2010.[Consultado:2016-05-

10]. Disponible en: http://marmota.dlsi.uji.es/WebBIB/papers/2007/1_Pascual-MIA-2007.pdf

R Foundation. [en línea]. [Consultado: 2015-04-21]. Disponible en: http://www.r-project.org/about.html.

RAYMOND J. (2005). *Mooney. CS 378: Intelligent Information Retrieval and Web Searc*h.[en línea].[Consultado: 2016-03-05]. Disponible en: http://www.cs.utexas.edu/users/mooney/.

ROMESBURG, C. *Cluster analysis for researches*. North Carolina: Lulu press, 2004.

SALTON, G. y MCGILL, M. J. *Introduction to Modern Information Retrieval.Computer Science* Series. [S.L.]: McGraw-Hill, 1983.

SALTON, WON y YANG.(1975). A Vector Space Model for Automatic Indexing.*Comunication of the ACM*, 18(11).

SAM, H. y KARYPIS, G. Centroid-Based Document Classification: Analysis & Experimental Results. *En Proceedings of the Fourth European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD) Lyon*, France, 2000, p. 424-431.

SAMAS (2007). [en línea]. [Consultado: 2016-02-18]. Disponible en: http://www.samas.org.ar

SAMPER. (2005). *Estudio y evaluación de un sistema inteligente para la recuperación y el filtrado de información de internet*. Universidad de Granada, Granada.

TAN, A.H. *Text Mining: The state of the art and the challenges*. 1999.

VON-LUXBURG, U. (2007).A tutorial on spectral clustering.*Statistics and Computing* 17(4): 395-416.

Wikipedia (2012). [en línea]. [Consultado: 2016-02-03]. Disponible en: http://es.wikipedia.org/wiki/html.

Anexos

Autora: Nerina Peña Olivero Tutor: Ing. Miguel Barrera Fernández

Anexo 1

Terminologías de Negocio

Arbitro: Experto, referee o persona con influencia en una o ciertas materias porque es considerada una autoridad en ellas.

Maquetas: Artículos corregidos en HTML y/o PDF dependiendo del formato con el que se publique en la revista.

Anexo 2

Terminologías de Minería de Dato

Análisis de series de tiempo (time-series): Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos.

Coeficiente de correlación cofenética: medida muy reconocida dentro de la estadística para medir el grado de fiabilidad con que se puede decir que un dendrograma conserva las distancias en parejas entre los datos originales que no han sido modelados. Este se emplea para evaluar el grado de ajuste de una clasificación a un conjunto de datos y como criterio para evaluar la eficiencia de varias técnicas para obtención de clústeres. También ha sido ampliamente utilizada en estudios de clasificación

Corpus: Conjunto cerrado de textos o de datos destinados a la investigación científica. Recopilación de los escritos de un autor.

Clasificación: Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como "posibilidades de crédito" con valores tales como "Bueno" y "Malo".

Clustering (agrupamiento): Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.

Data Mining: La extracción de información predecible escondida en grandes bases de datos.

Modelo lineal: Un modelo analítico que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes).

Modelo no lineal: Un modelo analítico que no asume una relación lineal en los coeficientes de las variables.

Anexo 3

Entrevistas a miembros del Consejo editorial de la Revista Minería y Geología.

1. ¿Cómo se realiza actualmente el proceso de selección de árbitros en la revista?
2. ¿Qué tiempo demora cada subproceso del proceso de selección de árbitros en la revista?
3. ¿Quiénes son responsables dentro del Consejo editorial de la selección de árbitros?