Universidad Internacional de La Rioja (UNIR)

ESIT

Máster Universitario en Análisis y Visualización de Datos

Comparativa de técnicas para el descubrimiento de conocimiento aplicadas al comportamiento de producción científica

Trabajo Fin de Máster

Presentado por: Galán Mena, Jorge

Directora: Fuentes Lorenzo, Damaris

Ciudad: Cuenca, Ecuador
Fecha: 19/09/2019

Resumen

La creciente actividad científica en las unidades de investigación de los Institutos de Educación Superior (IES) ha generado una gran cantidad de indicadores. Estos indicadores tienen relación con los productos científicos y perfiles de investigadores que son almacenados en Servidores de Repositorios (SDR) de terceros. En este trabajo se presenta una aproximación para contrastar dos técnicas de descubrimiento de conocimiento para la formación de perfiles de investigadores con los que cuenta una IES, ello permite simplificar la categorización de los investigadores a partir de múltiples indicadores de producción científica, obteniendo la información del SDR de Scopus. Las técnicas propuestas parten de la misma información, pero difieren en la conformación del dataset, mediante el uso de métodos tradicionales con tablas y modelado del conocimiento por medio de esquemas semánticos.

Palabras Clave: Descubrimiento de conocimiento, Modelamiento del conocimiento, ACP, K-Medias, Clúster de Louvain

Abstract

The growing scientific activity in the research units of the Institutes of Higher Education (HEI) has generated many indicators. These indicators are related to scientific products and researcher profiles that are stored in third party Repository Servers (RS).

This work presents an approach to contrast two knowledge discovery techniques for the formation of researcher profiles which a HEI has, simplifying the categorization of researchers from multiple indicators of scientific production, obtaining the information from the RS of Scopus. The proposed techniques are based on the same information but differ in the conformation of the dataset, using traditional methods with tables and modelling of knowledge by means of semantic schemes.

Keywords: Knowledge Discovery in Databases, Knowledge Modeling, PCA, K-Means, Louvain Clustering

Índice de contenidos

Índice de tablas

Índice de figuras

1. Introducción

En el campo de la investigación, uno de los resultados más visibles de este proceso intensivo de conocimiento son los productos científicos documentales, que desprenden muchas métricas y forman una huella del comportamiento científico de los investigadores. Con esta premisa, es inevitable caracterizar el potencial de investigación que una IES posee y también diagnosticar los cambios que se dan ante las decisiones gerenciales en estas comunidades científicas, por lo cual este estudio tratará las problemáticas y posibles soluciones que el presente TFM abordará a lo largo de sus contenidos.

1.1 Justificación

El creciente aumento en la producción científica por parte de las comunidades de investigadores de cada Institución de Educación Superior (IES) a nivel mundial propone un reto de gobernabilidad, esto ha provocado que las personas que lideran estos procesos en las comunidades científicas muevan dichas comunidades en torno a indicadores propuestos por la misma institución y por terceros. Múltiples aproximaciones alrededor de indicadores de desempeño de la producción científica han sido propuestas por organizaciones reconocidas dentro de la comunidad científica, las cuales proponen indicadores como: número de publicaciones de una institución en revistas de acceso abierto, número de publicaciones en revistas de alto impacto y número de coautores de instituciones de otros países; entre estas organizaciones que buscan medir el impacto científico de las IES se encuentran: Global University Ranking, Webometrics Ranking of World University, World Best University Ranking y otros.

El problema se suscita cuando se pierde el entendimiento de la complejidad de los ecosistemas de investigación, basando las decisiones en la elección de un solo indicador, el cual puede tener una relación muy fuerte con otros indicadores que no se han tomado en cuenta, y a su vez, como en el caso de los indicadores de terceros, pueden no perseguir los objetivos de la institución. Todas las IES tienen diferentes metas, objetivos y visiones hacia los que los esfuerzos de cada nivel de una institución se destinan a alcanzar, por lo cual no se pueden desvirtuar de indicadores propuestos en *rankings* que cuentan con la aprobación de numerosas personas estudiadas en el tema y grandes comunidades científicas; sin embargo, no deben ser los ejes centrales sino una mediación entre indicadores de *rankings* y los propios indicadores propuestos por las IES, siendo estos indicadores internos los que deben primar sobre el resto.

El problema del entendimiento de la complejidad de un ecosistema de investigación toma su relevancia debido a que los múltiples *rankings* se ajustan de una manera global para

categorizar a las IES según su comportamiento de producción científica pero no se ajustan a las realidades de cada una de las IES para poder diagnosticar dichos ecosistemas y, de esta manera, evaluar las decisiones tomadas a partir de los indicadores.

1.2 Planteamiento del trabajo

La producción científica de los investigadores de las IES se encuentra generalmente fuera de las bases de datos y de sus repositorios internos, por lo cual dicha información se almacena en repositorios de terceros, denominados repositorios de datos científicos (SDR), como es el caso de: IEEE Explore, Scopus, Scielo, Latindex, entre otros. Estos SDR la mayoría de las veces ofrecen APIs para el consumo de sus metadatos [1] en donde se exponen los comportamientos de la producción científica de los investigadores, facilitando la obtención de estos metadatos para realizar análisis de diferente índole, como las redes colaborativas que se forman a partir de la producción de papers [2] o análisis con tesauros para la caracterización de los contenidos como es el caso de SciVal [3].

Este trabajo convierte la información ofrecida por uno de los SDR de mayor popularidad a nivel de la comunidad científica, como es el caso de Scopus, ofreciendo la posibilidad de utilizar su API [4] y con estos metadatos realizar un análisis de la información de los investigadores de la Universidad Politécnica Salesiana de Ecuador por medio de dos métodos diferentes basados en descubrimiento de conocimiento, contrastando los beneficios del análisis de patrones por medio de técnicas de minería de datos para encontrar características comunes (tipo de producción científica, cuartil de publicación, área de investigación, número de coautores internos y externos a la institución) que permitan encontrar clústeres, caracterizados, a través de estadística descriptiva, para, de esa manera, diagnosticar el ecosistema de investigación, posibilitando el entendimiento de los perfiles de investigadores con los que cuenta la institución.

1.3 Estructura de la memoria

A continuación se detalla la organización de los capítulos del trabajo de fin de máster (TFM) descritos de una manera breve.

1.3.1. Introducción

La introducción relata la problemática que fundamenta el porqué de la relevancia del presente TFM, describe las causas y aborda de manera general la solución que se encontró para afrontar la problemática descrita. Por último, se describen los contenidos que se abordan a lo largo del presente trabajo.

1.3.2. Contexto y estado del arte

Para esta sección del marco teórico se abordan las diferentes aproximaciones que existen sobre el descubrimiento del conocimiento como valor agregado para la toma de decisiones, también se presenta literatura científica que propone la resolución de diferentes problemáticas por medio del uso de clústeres y el rol de los modelos semánticos en estas técnicas. Por último, se describe la estructura de los modelos semánticos en la web semántica y sus diferentes axiomas.

1.3.3. Metodología

En esta parte se describe el objetivo general, así como los objetivos específicos del TFM, además de la metodología de trabajo, con el detalle de los pasos que se van a utilizar, cómo se van a medir los resultados y a nivel de contraste entre las técnicas.

1.3.4. Descubrimiento del conocimiento de los perfiles de investigadores, aproximación con técnicas tradicionales.

En esta sección se aborda las diferentes fases de un proceso de descubrimiento de conocimiento sobre el dataset, compuesto por información de los investigadores de la UPS obtenidos del repositorio de Scopus, por medio de la utilización de técnicas de clustering, diferentes comunidades de investigadores que están caracterizados por diversos comportamientos científicos, que se traducen en una huella de perfiles de investigadores con la que cuenta la institución.

1.3.5. Descubrimiento del conocimiento de los perfiles de investigadores, aproximación basada en esquemas semánticos.

El capítulo utiliza las mismas fases que la primera metodología de descubrimiento de conocimiento de los investigadores de la UPS abordada en el capítulo anterior, pero utiliza técnicas de modelamiento del conocimiento, empleando ontologías y aplicación de similitud semántica para caracterizar la investigación de los trabajos científicos en torno al tesauro de la UNESCO, aportando una fase extra de agregación de conocimiento al dataset.

1.3.6. Comparativa de técnicas de descubrimiento de conocimiento de perfiles.

Esta parte arroja una comparativa de los clústeres encontrados en las dos metodologías, mediante un contraste de las métricas internas de la formación de los clústeres, así como también un contraste de los resultados finales a nivel cualitativo, para alcanzar los propósitos de los objetivos planteados en el TFM.

1.3.7. Conclusiones y trabajos futuros.

En esta parte final, se presenta una síntesis de las metodologías abordadas para el descubrimiento de patrones en los comportamientos de producción científica de la UPS. También se realiza una reflexión sobres las comparativas de las dos técnicas y los trabajos futuros que surgen en torno al TFM.

2. Contexto y estado del arte

El descubrimiento del conocimiento sobre bases de datos, más conocido en el mercado en inglés como *Knowledge Discovery in Databases* (KDD), se presenta como uno de los productos finales de Tecnologías de la Información (TI) de mayor valor para las organizaciones que lo poseen. El proceso permite extraer el conocimiento a partir de un proceso de identificación, de patrones válidos y nuevos sobre un dataset grande [5]. El proceso matemático sobre el cual se fundamenta el descubrimiento de patrones se denomina minería de datos, y agrupa un conjunto de técnicas basadas en algoritmos de inferencia que exploran los datos.

El KDD está compuesto de nueve pasos que son iterativos, lo que significa que en cada paso que necesite algún tipo de corrección, seguramente se requiere retroceder al paso anterior para realizar dicho cambio. Para aplicar correctamente el proceso se necesita un profundo entendimiento tanto de KDD como también del problema a resolver para aplicar los siguientes pasos como se puede ver en la Figura 1:

1. Comprensión del Dominio y Metas del KDD: El primer paso se basa en el entendimiento del dominio de la aplicación, puesto que la persona que lleva a cabo los procesos de KDD tiene que entender los objetivos del proyecto, para preparar un escenario en el cual se tomarán algunas decisiones que tienen que ser entendidas completamente.

2. Selección y Adición: Formar el dataset sobre el cual se va a realizar el KDD, seleccionando los datos de fuentes disponibles para completar e integrar el dataset con los atributos necesarios para llevar a cabo el proceso, puesto que es la materia prima sobre la cual se va a realizar la minería de datos y, al no considerar todos los atributos relevantes, podría comprometer el éxito del proceso de KDD. También se tiene que considerar el equilibrio en la tarea de construcción del dataset, puesto que, si se construye un dataset muy complejo, conlleva más esfuerzo por lo cual se debe tener muy en cuenta este balance.

3. Preprocesamiento: En esta etapa se mejora la fiabilidad de los datos, removiendo el ruido y lidiando con los datos faltantes. Dependiendo del caso se aplicarán desde técnicas de Minería de Datos para predecir datos faltantes hasta la posibilidad de no hacer nada con la información.

4. Transformación: El objetivo de este paso es mejorar los datos antes de aplicar un proceso de Minería de Datos, por lo cual se aplican métodos de reducción de

dimensiones y transformación de atributos como la discretización de atributos numéricos o el cambio de valores ordinales a valores numéricos.

5. Selección de la tarea de Minería de Datos: A partir del dataset se tiene que elegir el tipo de Minería de Datos más apropiada a lo que se busca, teniendo en cuenta que se puede tener dos grandes tipos de Minería de Datos: predicción que generalmente utilizan algoritmos supervisados y descriptivos que utilizan algoritmos no supervisados. Generalmente se construyen modelos explícitos o implícitos, utilizando técnicas de Minería de Datos basadas en aprendizaje inductivo.

6. Selección del algoritmo de Minería de Datos: Al definir el tipo de Minería de Datos, se tiene que elegir el método a aplicar. Teniendo en cuenta que se busca un resultado que brinde más comprensión o precisión, se debe optar por diferentes aproximaciones como las redes neuronales o técnicas como los árboles de decisión.

7. Implementación del algoritmo de Minería de Datos: En esta etapa se implementa el algoritmo para descubrir los patrones sobre el dataset y se varían los parámetros según el tipo de algoritmo y el número de épocas de entrenamiento, precisión, bias, entre otros, con el propósito de alcanzar un resultado óptimo de acuerdo con las metas planteadas.

8. Evaluación e interpretación: En este paso se interpretan los patrones descubiertos de los resultados y se evalúan según diferentes métricas, como la confiabilidad, para posteriormente comprender y encontrar la utilidad del modelo.

9. Descubrimiento del conocimiento: A partir del cumplimento de las anteriores etapas, es posible replicar el modelo en otros sistemas en entornos dinámicos con los retos de no tener las condiciones iniciales del laboratorio.

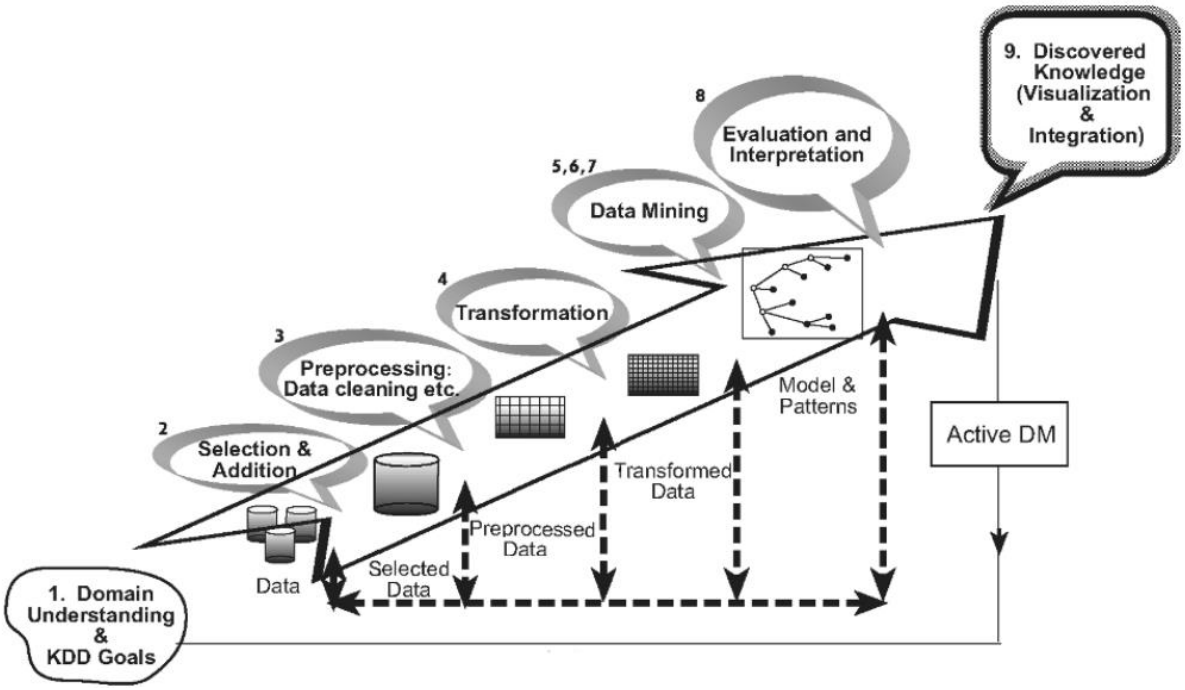


Figura 1. El proceso del Descubrimiento de Conocimiento en Bases de Datos. *[5]*

Una de las técnicas más utilizadas para lidiar con la reducción de dimensionalidad en la fase de transformación en el proceso de KDD es el Análisis de Componentes Principales (ACP), el cual sirve para representar la información de una manera más compacta. El ACP mira las relaciones que existen entre las variables de un objeto y estas son llamadas factores, obteniendo los componentes de los datos y capturando las características de estos datos [6]. Este método de reducción de dimensionalidad se ha utilizado en conjunto con técnicas de clústeres, como señala Kenekayoro [7], proponiendo un análisis de las páginas de home en sitios web de diferentes grupos de investigación para agruparlos según sus áreas de interés, posibilitando un futuro análisis para la formación de equipos de trabajo en el sector de la política. En el campo del análisis científico, Bollen [8] realiza un ACP sobre 39 indicadores de impacto académico obtenidos sobre el cálculo de diferentes bases de citas para investigar cómo estas medidas se relacionan entre sí.

Las IES en las comunidades científicas se suelen caracterizar en rankings de terceros según su producción científica, posicionando a cada entidad alrededor de varios indicadores. La organización SIMAGO propone un ranking ibero americano llamado SIR Iber [9], el cual mide el comportamiento científico de 1761 IES con base en los artículos en el SDR de Scopus. SIMAGO utiliza 11 indicadores para medir el desempeño de las IES en el campo de la investigación, estos indicadores son: la colaboración internacional, excelencia y liderazgo, artículos indizados en Scopus, proyectos de investigación, publicaciones en revistas que no pertenezcan a las IES, número de revistas de las IES, artículos con el 10% de citación más

alta en los campos científicos, artículos de revistas Q1, colaboración internacional, artículos publicados en revistas de acceso abierto, impacto normalizado y número de autores de diferentes instituciones que han participado como coautores.

En la literatura se pueden encontrar trabajos que buscan agrupar la actividad científica de los investigadores por medio de la aplicación de técnicas de mapeo y clústeres, basados en la modularidad de las redes bibliométricas que se forman, agrupando de esta manera a los investigadores de acuerdo con sus publicaciones científicas [10]. Múltiples aproximaciones a la literatura bibliométrica se han aplicado para encontrar áreas de investigación, basándose principalmente en las relaciones que se crean en función de las citaciones, formando diversas redes científicas de las cuales se pueden realizar análisis para encontrar patrones por medio de técnicas de partición de grafos [11].

La semántica se ha aplicado como una técnica que aprovecha el modelado del conocimiento y la explicitación de sus significados como base para diversas técnicas de aprendizaje de máquina aplicada a diversos estudios que utilizan técnicas de agrupamiento jerárquico apoyados en los metadatos de una ontología [12]. Otra investigación cuyo propósito era la de agrupar documentos de una manera significativa basado en sus contenidos, utilizó las ontologías para agregar conocimiento de fondo en conjunto con técnicas de clústeres basadas en precategorización, obteniendo mejores resultados de agrupación [13] y una ventaja al momento de resolver temas de incompatibilidad del lenguaje aplicado localmente para el agrupamiento [14].

La Web Semántica nace a partir de los problemas detectados en los inicios de Internet por parte de la W3C; Las páginas web estaban orientadas al entendimiento de las personas, pero no necesariamente para un fácil entendimiento de las máquinas. Aunque existen varias técnicas de scraping y análisis de texto para la captación de información de una página web, mantiene un porcentaje de error debido a que no es la manera más conveniente de exponer la información a robots que soliciten el contenido [15]. Es aquí cuando interviene la Web Semántica como un marco estandarizado de intercambio de información en un contexto universal de los contenidos de los datos.

Para hablar de Web Semántica y modelado de conocimiento, primero se debe partir de que el conocimiento posee un contexto específico que se interrelaciona con otros conocimientos adquiridos y posee un significado; Nonaka y Takeuchi lo definen como “un proceso humano dinámico de justificación de la creencia personal en busca de la verdad”, que cumple con una espiral en donde el conocimiento, a lo largo de su proceso, pasa de implícito a explícito y viceversa [16].

El conocimiento ya se ha instrumentalizado desde los lógicos matemáticos que propusieron una forma de representación declarativa antes de que se pudiera implementar en la parte informática. Con la informática y la Inteligencia Artificial (IA) se trabaja con el conocimiento explícito para crear modelos que formalicen estos acuerdos contextuales de representación, dichos modelos aún no han llegado a ser lo suficientemente expresivos ante las necesidades de imitar los comportamientos humanos, mientras que los matemáticos no tomaron en cuenta la posibilidad de la creación de un razonamiento autónomo [17].

La Web Semántica propuesta por Tim Berners-Lee se divide en varias capas construidas sobre las capas de la Web tradicional [18] como se presenta en la Figura 2:

1. Capa XML: El XML (eXtensible Markup Language) derivado de SGML (Standar Generalized Markup Language), es un metalenguaje que utiliza etiquetas de apertura y cierre para formar elementos. A diferencia de HTML, que está más orientado a la visualización de los elementos por parte de los usuarios, con XML se forman estructuras anidadas que definen los datos y tiene la capacidad de especificar restricciones. La gran limitación que posee es que al enfocarse en las estructuras de sus etiquetas carece de interpretación del significado.

2. Capa RDF y RDF Schema: Esta capa introduce un modelo semántico sobre la capa de XML, que aporta un modelo de datos con significado. Este modelo está representado por tripletas formadas de un sujeto, predicado y objeto. Aunque RDF utiliza al metalenguaje XML para su representación no es el único lenguaje sobre el cual se pueden representar documentos RDF, existen notaciones de documentos como N-Triple, Notation 3 RDF, Turtle, entre otras [19]. RDF Schema es el que provee el modelo de axiomas que organiza al modelo de datos, estos axiomas son reglas características que le brindan una gran potencia al momento de definir los esquemas, como es el caso de las subclases y subpropiedades que representadas por medio de las propiedades subClassOf y subPropertyOf, también tienen la posibilidad de generar restricciones tanto de rangos cuanto de dominios.

3. Ontología: Las ontologías en la rama de la IA son utilizadas como instrumentos para formalizar el conocimiento. Muchos autores, a lo largo de la historia, han tratado de darle sentido a las ontologías. Una de las definiciones más importantes la postuló Gruber en 1993, definiendo que una ontología constituye “una especificación explícita formal de una conceptualización compartida” [20]. Se puede decir que una ontología es una explicitación del conocimiento representado por un modelo formal de una conceptualización de la realidad, la cual se representa por medio de un lenguaje para ser manipulada por un sistema autónomo que puede generar inferencia a partir del

conocimiento almacenado. Aunque RDF y RDFS proponen varios axiomas, la W3C creó tres lenguajes que extendían los axiomas de estas estructuras semánticas para tener una mayor potencia a nivel de inferencias, estos lenguajes son: OWL Full, OWL DL y OWL Lite.

4. Lógica: Esta capa permite generar nuevo conocimiento a partir de los axiomas propuestos en las capas inferiores, para de esta manera, gestionar los datos que la mayoría de las veces se lo hace en la capa de negocios. Existen algunos tipos de lenguajes que ayudan a implementar reglas sobre la capa ontológica como son: RuleML (Rule Markup Language), SWRL (Semantic Web Rule Language) y SPIN (SPARQL Inference Notation).

5. Capa de comprobación y confianza: La capa de comprobación verifica que las capas de menores niveles estén correctas, realizando validaciones, mientras que la capa de confianza se orienta más al aseguramiento de la fuente de publicación de los datos cifrados y su información para verificar. Se puede decir que estas últimas dos capas son las menos explotadas en el campo de investigación de la Web Semántica.

El campo del descubrimiento de conocimiento visto en el proceso de KDD brinda una metodología interactiva e iterativa para encontrar patrones de datasets grandes y, de esta manera, generar valor para una institución que los aplica. Para el presente trabajo, el descubrimiento de información tiene que estar sustentado sobre un marco referencial fuerte como lo es KDD para, de esta manera, realizar un análisis descriptivo de la información de los investigadores con base en sus publicaciones científicas en el SDR de Scopus.

La novedad que suscitó el manejo de modelos de conceptualización por parte de la Web Semántica como un camino para mitigar la ambigüedad de la representación de información, añade valor a cualquier modelo de datos, sumando que en el ámbito de la bibliometría tiene una fuerte aceptación al momento de categorizar documentación con base en su contenido, añadiendo una capa de significado más fuerte a la descripción de los objetos, por cuanto, apoyados de técnicas de clustering, realizan análisis descriptivos que aportan conocimiento sobre las áreas de investigación y producción científica, lo cual sugiere utilizar técnicas de modelamiento del conocimiento sobre el dataset para mejorar el proceso de KDD.

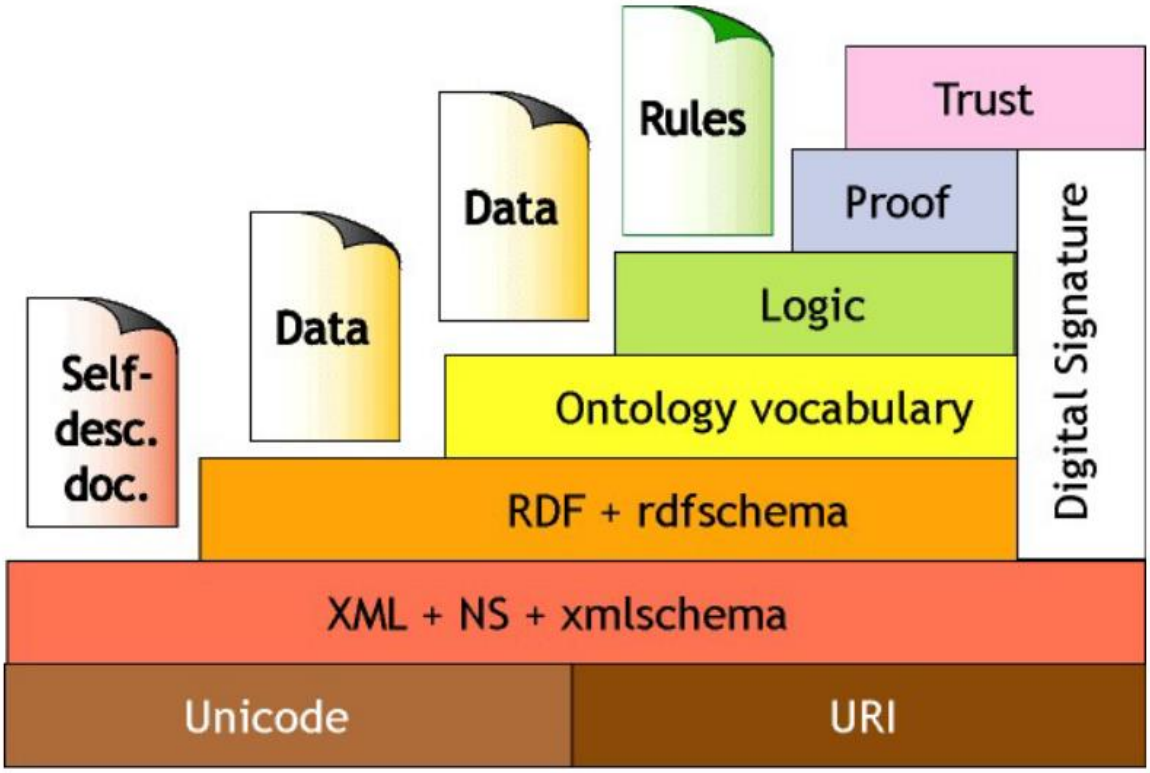


Figura 2. Capas de la Web Semántica. [18]

Las técnicas aplicadas en la fase de transformación de los datos para reducción de dimensionalidad, vistas anteriormente, ayudan a manejar de una manera más fácil el conjunto de variables que representan a un objeto como es el caso de ACP. Al combinar técnicas de reducción de dimensionalidad en conjunto con técnicas de aprendizaje no supervisado, como es el caso de K-Means, se obtienen resultados favorables con una optimización del proceso de agrupamiento al momento de ejecutar el algoritmo en el computador. Estás técnicas en conjunto se han utilizado en procesos similares cuando se manejan indicadores de producción científica, como se revisó en el estado de arte actual y son un punto referencial muy fuerte para la elección de estos algoritmos en la etapa de Minería de Datos del KDD.

3. Objetivos concretos y metodología de trabajo

Este capítulo describe los alcances del TFM**,** así como también la metodología que se va a abordar para el desarrollo del presente trabajo.

3.1. Objetivo general

El objetivo principal de este proyecto es realizar la comparativa entre dos técnicas de descubrimiento de conocimiento, aplicadas al análisis de patrones de comportamiento de los investigadores de la UPS.

3.2. Objetivos específicos

Para alcanzar el objetivo general de la comparativa metodológica, primero se deben cumplir los siguientes objetivos específicos:

- Crear el dataset del comportamiento científico de los investigadores de la UPS, obteniendo la información por medio del API de Scopus.
- Definir las metodologías a abordar para el descubrimiento de patrones en el dataset de la producción científica de la UPS.
- Limpiar los registros del dataset para asegurar la calidad de los datos a ser analizados durante las siguientes etapas.
- Transformar el dataset a una versión más manejable y estandarizada para la fase de clústeres.
- Crear los clústeres que representan las comunidades de investigadores que se autodescriben para el entendimiento de la fuerza laboral en producción científica de la UPS.
- Realizar las comparativas de métricas internas de formación de clústeres en las dos metodologías.
- Realizar el contraste de la coherencia a nivel funcional del descubrimiento de patrones aplicados al dataset de investigadores por medio de las dos metodologías.

3.3. Metodología del trabajo

En este capítulo se determina el caso de estudio sobre el cual, de manera sistemática, se presentan los dos caminos de metodologías aplicadas al descubrimiento de conocimiento sobre un dataset de investigadores. Después, se describe las métricas utilizadas para hacer la comparativa y contraste de las técnicas.

3.3.1. Caso de estudio

En el caso de estudio se considera la producción científica de los investigadores de la Universidad Politécnica Salesiana del Ecuador, contando con información de la producción científica de la institución desde 1998. El estudio se realizó en torno a la publicación científica de 600 investigadores de diversos centros de investigación con más de 800 artículos. La información se obtuvo de la base de datos de Scopus, de la cual se obtuvieron los perfiles de investigadores y producción científica, complementando esta información con el sistema institucional de investigación de la UPS, para completar los perfiles de los investigadores con el número de proyectos de investigación en los que participaron.

3.3.2. Metodología

La comparativa se centra en dos metodologías de KDD aplicadas sobre la información de los investigadores de la UPS como se presenta en la Figura 3; (a) La primera metodología define los pasos tradicionales de KDD partiendo desde la comprensión del dominio y metas del KDD, selección y adición de información, remoción de los investigadores que sean outliers o carezcan de ciertas variables, reducción de la dimensionalidad por medio de ACP, K-Means sobre la matriz de factores de ACP para obtener los clústeres y estadística descriptiva sobre los clústeres que lleven a interpretar la información; (b) El segundo método utiliza las mismas etapas pero antes de la fase de selección y adición realiza una fase de modelado de conocimiento basada en la metodología NeOn [21], la cual tiene su propia fase de selección de datos no ontológicos, transformación de los documentos no ontológicos a ontológicos, carga de documentos ontológicos, aplicación de modelos de similitud por medio de Vector Space Model (VSM) y de esta manera, apoyados por la base de datos de tripletas GraphDB, se consulta esta información con conocimiento agregado como entrada de la fase de selección y adición. En la metodología desarrollada se toma el capítulo 1 de Introducción como la fase de comprensión del dominio y objetivos del KDD para las dos metodologías.

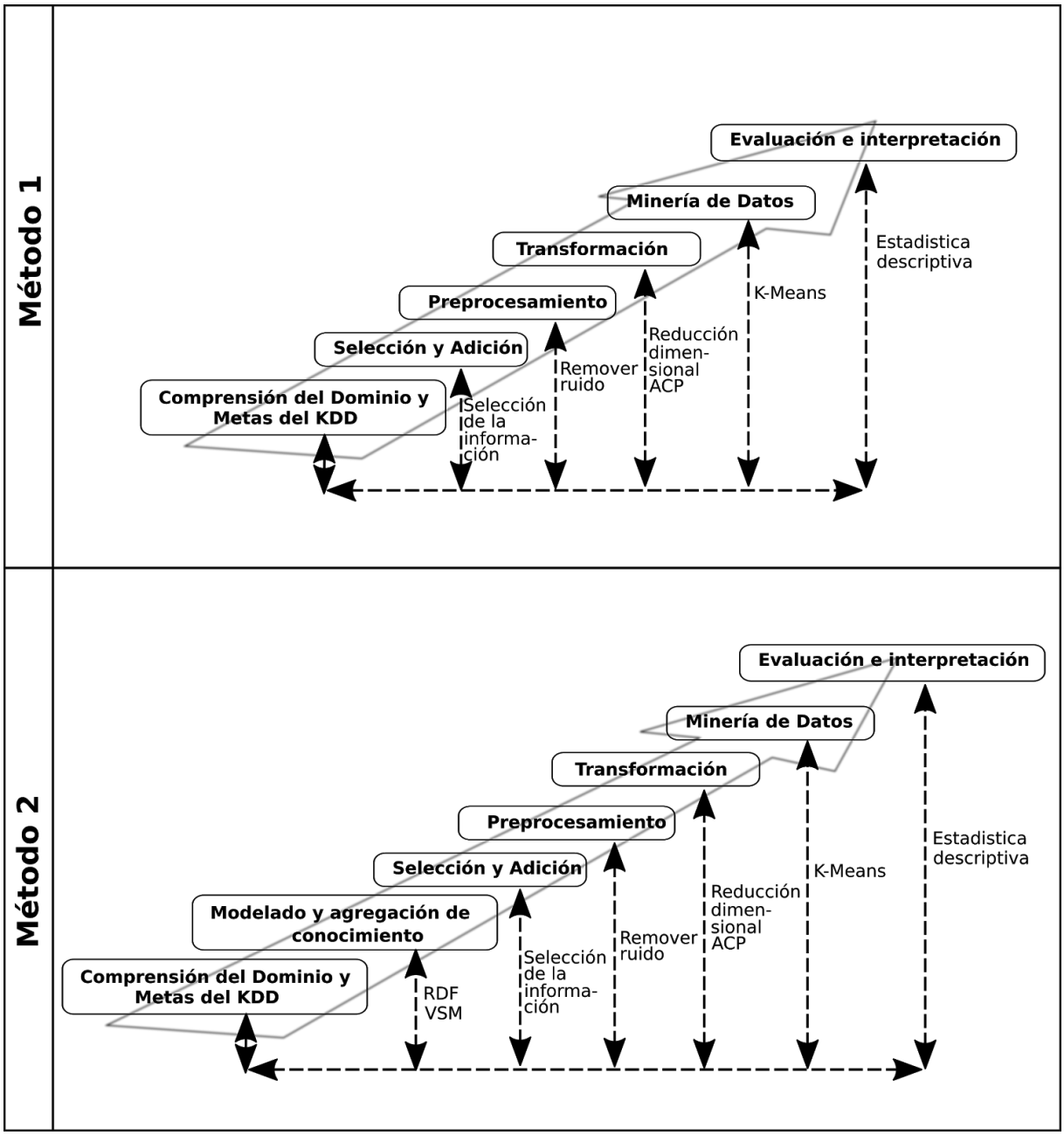


Figura 3. Metodologías de KDD (a) figura superior, metodología tradicional de KDD (b) figura inferior, metodología de KDD apoyado por modelado de conocimiento.

4. Descubrimiento del conocimiento de los perfiles de investigadores, aproximación con técnicas tradicionales

En este capítulo se describen los procesos de la metodología de KDD expuesta en el capítulo 2 de Marco Teórico en donde se abordan las fases de Selección y Adición, Preprocesamiento, Transformación, Minería de Datos y la Evaluación e interpretación de los resultados. Esta metodología tiene como finalidad encontrar patrones sobre la información del dataset de producción científica de los investigadores de la UPS para generar conocimientos nuevos acerca de las comunidades que se forman en el ecosistema de investigación y, de esta manera, explicar qué caracteriza a dichas comunidades; de esta manera, se logra un instrumento para visibilizar la complejidad sistémica ante la toma de decisiones.

4.1. Selección y adición

En esta fase se procede a seleccionar la información de las bases de datos internas y externas, relacionadas con la producción científica de los investigadores de la UPS, como se presenta en la Figura 4. También se adiciona información faltante que no se encuentra en los registros obtenidos a partir de las bases de datos. Como resultado final de esta fase, se obtiene un dataset con toda la información relevante para los demás procesos de KDD.

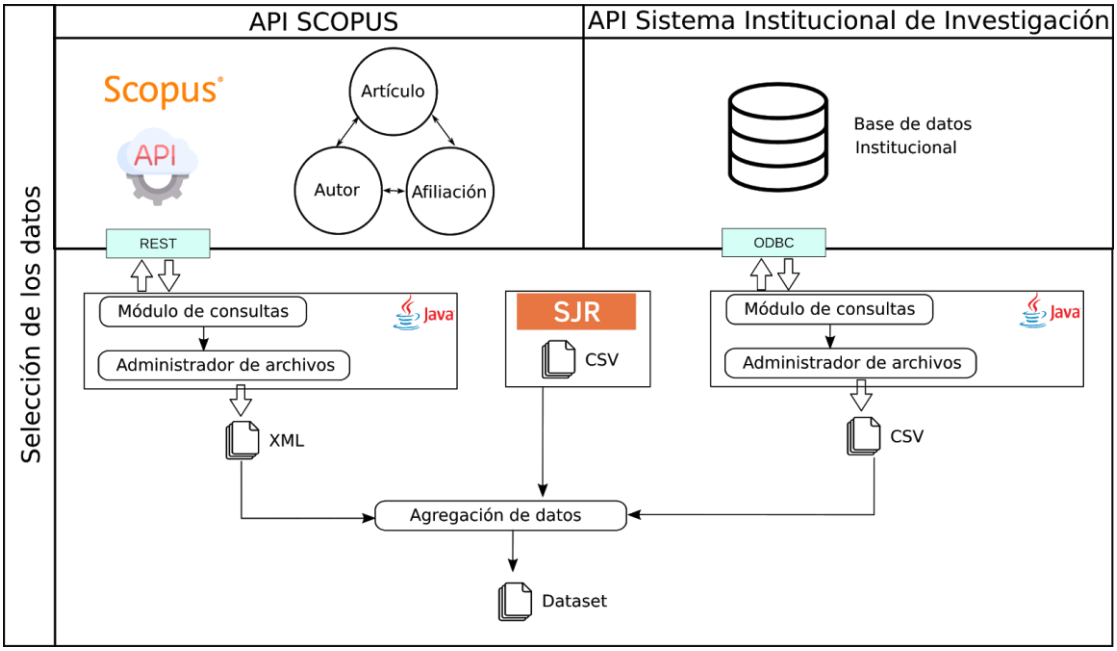


Figura 4. Esquema de proceso de selección y agregación.

4.1.1. Selección de los datos

La investigación en las IES es un proceso de conocimiento intensivo que arroja, como parte de sus resultados, productos documentales que se encuentran dispersos en repositorios de diferentes organizaciones, dichos repositorios albergan metadatos que se desprenden del proceso de investigación como son las citaciones, redes de colaboración de investigadores y afiliaciones, palabras clave, *abstract*, título, entre otras. Uno de los repositorios más importantes de documentación científica es Scopus, la cual basa su modelo de datos en tres grandes grupos que son: documentos, autores y afiliaciones. Entre las ventajas que ofrece Scopus a los analistas de datos y desarrolladores, se encuentra el acceso a sus servicios de publicación de metadatos por medio de un API; por esta razón, con la gran aceptación y reconocimiento por la comunidad científica, se tomó esta fuente de datos para construir el dataset de investigadores.

Para consultar la información se utilizó el lenguaje Java con el fin de consultar dichos servicios por medio del consumo de servicios REST publicados por Scopus, en donde se realizó una consulta, por medio del Scopus Search API, de los códigos de afiliación (116454887, 116541188, 60105797, 60105798, 60105693) de la UPS como se muestra a continuación:

https://api.elsevier.com/content/search/scopus?query=AF-ID(116454887,116541188, 60105797, 60105798, 60105693) & apiKey=<apikey>

Esta búsqueda retornó 808 objetos en 33 documentos XML con una paginación de 25 documentos por consulta. De dicha información se extrajeron los campos más relevantes, como se puede apreciar en la Tabla 1:

Tabla 1. Tabla descriptiva de los campos seleccionados de los metadatos del servicio Scopus Search Retrieval.

Campo	**Descripción**
Tipo de contenido	Tipo de documento científico como, por ejemplo: revista, paper de conferencia, entre otros.
Fecha de cubierta	Fecha de la portada del documento científico.
Id Scopus	Identificador de documento científico manejado por Scopus.
Código de fuente	Código de revista que alberga el documento

A partir de estos documentos se realizó una consulta al Autor Retrieval API de Scopus para obtener la información pertinente a los autores de los documentos científicos recuperados:

https://api.elsevier.com/content/author/author_id/<author-id>?apiKey=<apikey>

Esta búsqueda arrojó la información de los perfiles de los investigadores de entre los cuales se escogieron los valores más importantes como se presenta en la Tabla 2.

Tabla 2. Tabla descriptiva de los campos utilizados a partir de la consulta de Author Retrieval.

Campo	**Descripción**
Conteo de citaciones	El total de citaciones de los documentos científicos del autor.
Conteo de coautores	Conteo de coautores con los que ha colaborado para producir artículos científicos.
Conteo de documentos	Conteo de artículos científicos producidos por el autor.
Scopus Id	Id del perfil del autor en la base de datos de Scopus.
Índice H	Índice H mide la calidad de investigador en función de la producción de artículos científicos y las citaciones que tiene.
Publicación final	Última fecha de publicación de un artículo científico por parte del autor.
Publicación inicial	Fecha de primera publicación de artículo científico por parte del autor.

Otra de las fuentes utilizadas para la consulta de datos es un sistema institucional de la UPS, el cual recopila información en bases de datos internas acerca de la organización de los grupos de investigación, la producción científica de los investigadores y los proyectos de investigación sobre los cuales se trabaja en la institución. Utilizando el lenguaje Java y un conector a la base de datos, se realizó la consulta a la base para obtener más información sobre el perfil del investigador que se arma para el dataset, aportando con datos como el rol que el investigador desempeña en un grupo de investigación o proyecto como se presenta en la Tabla 3.

Tabla 3. Tabla de resumen de los campos de interés a partir de la consulta del perfil de investigador en el sistema de investigación institucional.

Campo	**Descripción**
Id	Identificador en el sistema institucional de investigación.
Conteo proyectos	Número de proyectos de investigación en el que ha participado el investigador.
Rol	Rol del investigador en el grupo de investigación.
Id grupo de investigación	Identificador del grupo de investigación al que pertenece el perfil del investigador.
Id Scopus	Identificador del perfil del investigador en Scopus

4.1.2. Adición

Los datos que se consultaron tanto de las bases de datos institucionales como del repositorio de Scopus aportan mucho valor para las siguientes etapas del KDD; sin embargo, para aprovechar al máximo todos los parámetros, se añadió un dataset del Scimago Journal Ranking (SJR) con el fin de clasificar los códigos fuentes de los documentos que se extrajeron, de esta manera se puede agregar el cuartil de impacto que tiene esa publicación. El indicador de cuartil de impacto científico propuesto por SJR, realiza una ponderación de los artículos de revistas académicas en el SDR de Scopus, a través de categorizaciones en donde el primer cuartil es de mayor impacto y el último es de menor impacto, basándose en las citaciones que tenga la revista y en función de la red de citación heterogénea en la que se encuentre [22].

De esta manera se construyó, a partir de toda la información seleccionada, un dataset que une tanto los perfiles de Scopus como los perfiles de los sistemas institucionales de investigación de la UPS como se muestra en la Tabla 4.

Tabla 4. Campos del dataset final de la fase de Selección y Adición.

Campo	**Descripción**
Id	Identificador en el sistema institucional de investigación.
Conteo proyectos	Número de proyectos de investigación en el que ha participado el investigador.

Rol	Rol del investigador en el grupo de investigación.
Id grupo de investigación	Identificador del grupo de investigación al que pertenece el perfil del investigador.
Id Scopus	Identificador del perfil del investigador en Scopus.
Conteo de citaciones	El total de citaciones de los documentos científicos del autor.
Conteo de coautores	Conteo de coautores con los que ha colaborado para producir artículos científicos.
Conteo de documentos	Conteo de artículos científicos producidos por el autor.
Índice H	Índice H mide la calidad de investigador en función de la producción de artículos científicos y las citaciones que tiene.
Publicación final	Última fecha de publicación de un artículo científico por parte del autor.
Publicación inicial	Fecha de primera publicación de artículo científico por parte del autor.
Conteo de Documentos Q1	Documentos científicos publicados en una revista de impacto Q1.
Conteo de Documentos Q2	Documentos científicos publicados en una revista de impacto Q2.
Conteo de Documentos Q3	Documentos científicos publicados en una revista de impacto Q3.
Conteo de Documentos Q4	Documentos científicos publicados en una revista de impacto Q4.

4.2. Transformación

En la transformación de datos se realizaron los procesos de discretización de atributos y reducción de dimensionalidad para ser utilizados en la posterior técnica de clustering.

La discretización de atributos se realizó en el lenguaje de programación Python, transformando la columna de "Id Grupo de Investigación" para el set de grupos de investigación en diferentes columnas, las cuales podrán tener un valor de 0 si el investigador no pertenece a la columna del grupo, y 1 si el investigador pertenece al grupo de investigación. En esta tarea se adicionaron 45 columnas pertenecientes a los grupos de investigación como se presenta en la Tabla 5 desde la columna i16 a la columna i60; desde la columna i1 a i15 se encuentran codificadas las diferentes variables de la etapa de selección.

Tabla 5. Campos del dataset final de la fase de Transformación.

Atributo	Descripción	Tipo de valor
I1	Conteo de citaciones	Entero
I2	Conteo de citaciones al autor	Entero
I3	Conteo de coautores	Entero
I4	Conteo de proyectos de investigación	Entero
I5	Conteo de artículos sin cuartil en SJR	Entero
I6	Conteo de artículos Q1 en SJR	Entero
I7	Conteo de artículos Q2 en SJR	Entero
I8	Conteo de artículos Q3 en SJR	Entero
I9	Conteo de artículos Q4 en SJR	Entero
I10	Rol de coordinador de grupo de investigación	Binario (0 o 1)
I11	Rol de estudiante investigador	Binario (0 o 1)
I12	Rol de investigador	Binario (0 o 1)
I13	Rol de investigador externo	Binario (0 o 1)
I14	Rol de responsable de comunicación	Binario (0 o 1)
I15	Rol de técnico docente	Binario (0 o 1)
I16-I60	Grupos de investigación de la Universidad Politécnica Salesiana a los que pertenece el investigador	Binario (0 o 1)

La tarea de reducción de dimensionalidad se hizo mediante la herramienta Orange 3; en la Figura 5 se aprecia el flujo de lectura del dataset y aplicación del algoritmo ACP para reducción de dimensiones. Se seleccionó un total de 39 componentes principales que representan el

90% de la cobertura de covarianza de las muestras como se puede observar en la Figura 6. Esta tarea redujo el dataset de un total de 60 dimensiones a 39 dimensiones para su posterior uso en la fase de minería de datos.

Figura 5. Captura de Orange 3 del flujo de tareas para la reducción de dimensiones.

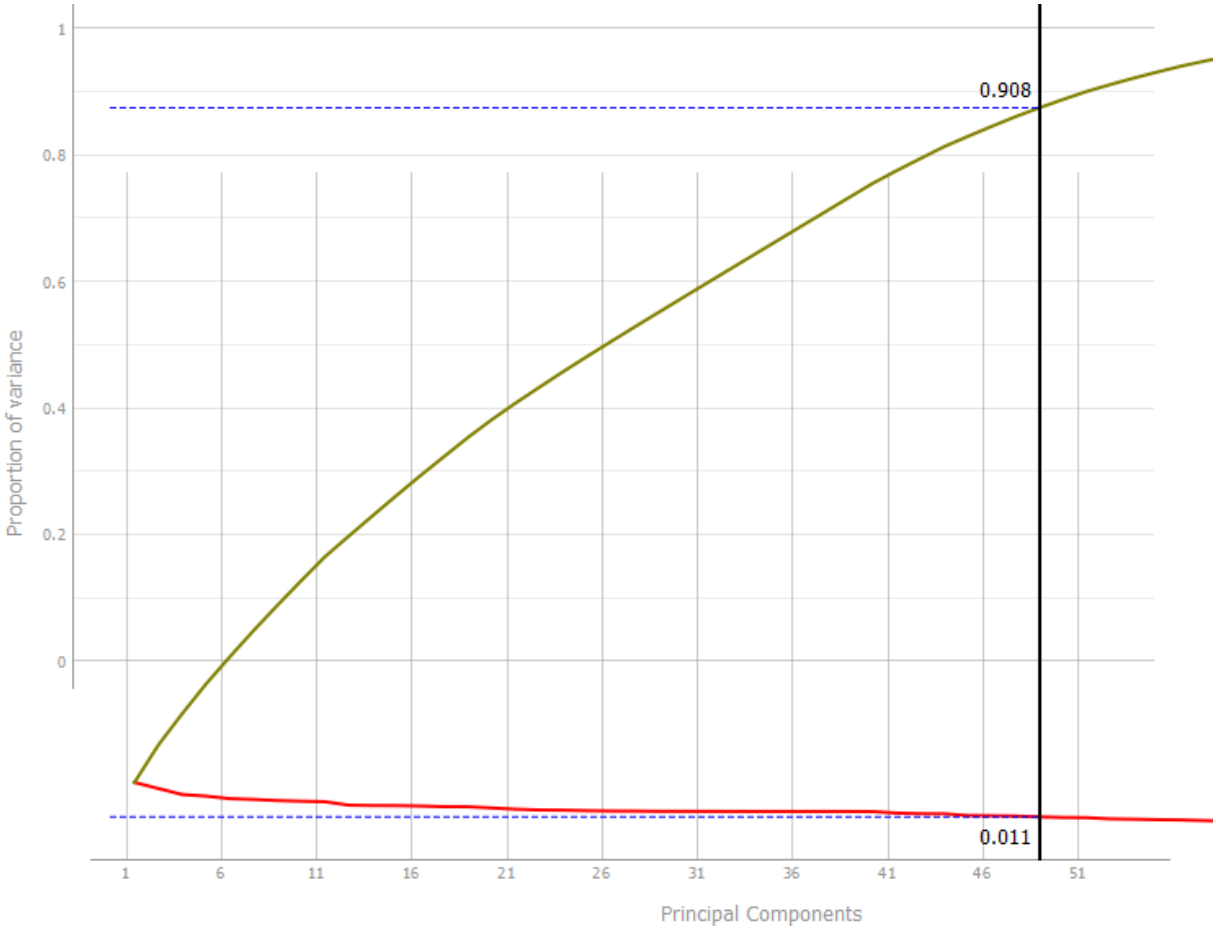


Figura 6. Gráfica de proporción de varianza de los componentes principales.

4.3. Minería de datos

En la etapa de minería de datos se tomó como entrada el dataset de 39 dimensiones y 221 registros de la etapa de transformación, escogiendo dos tareas para la descripción de la

información del dataset, esto se hizo a través de los algoritmos no supervisados de: K-Means y Clustering de Louvain, en el software de Orange 3.

4.3.1. K-Means

La tarea de K-Means tuvo una inicialización de los centroides aleatoria para obtener una convergencia más rápida del algoritmo, se utilizaron 500 iteraciones y se utilizó la medida de la silueta para detectar el K-Optimo de clústeres a formarse. En la Figura 7 se presenta el flujo de conexión de la tarea mediante la herramienta Orange 3.

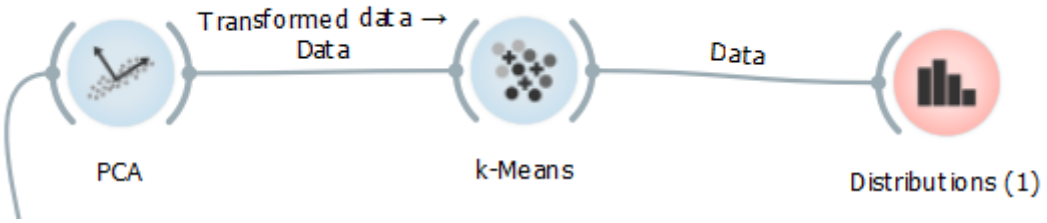


Figura 7. Captura de flujo de tareas de K-Means

Al aplicar la tarea de K-Means, el K-Óptimo fue de 29 grupos con una puntuación de silueta de 0.363. En la Figura 8 se presenta la gráfica de distribución de investigadores en torno a los grupos formados en los clústeres.

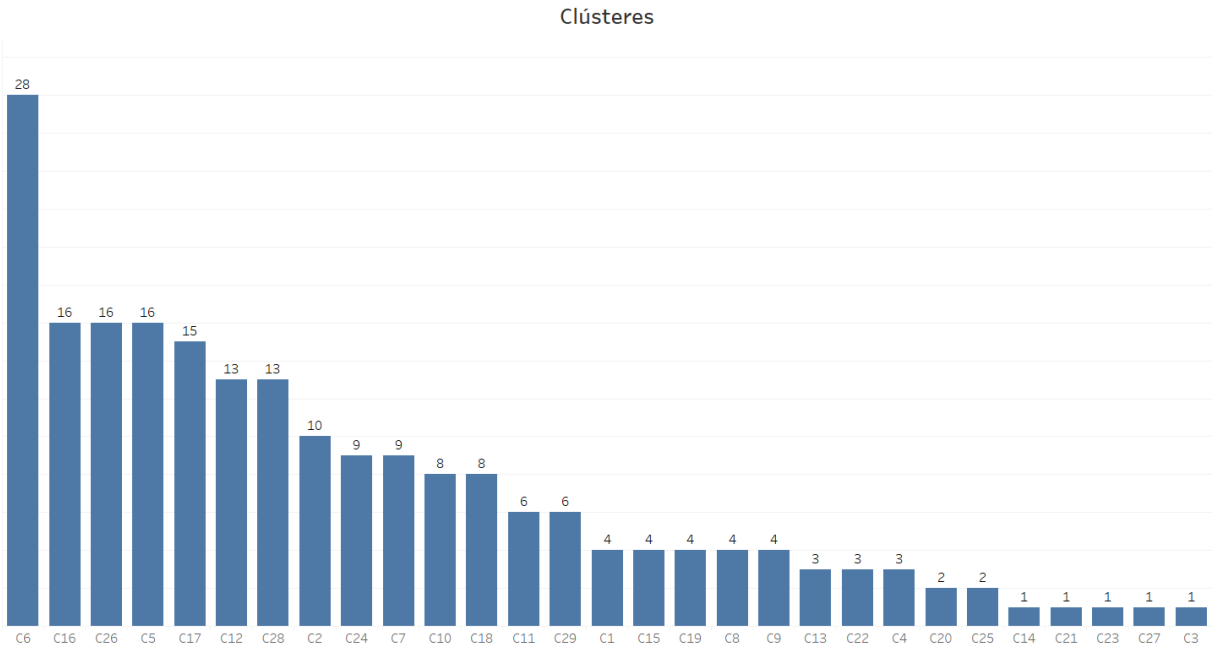


Figura 8. Diagrama de distribución de los clústeres formados con K-Means

4.3.2. Clustering de Louvain

La tarea del clúster de Louvain sirve como un contraste a la primera técnica, el algoritmo es utilizado para el reconocimiento de comunidades de redes detectando la modularidad. Se utilizó la medida Euclídea como medida de distancia con una resolución 1. En la Figura 9 se presenta el flujo de conexión de la tarea mediante la herramienta Orange 3.

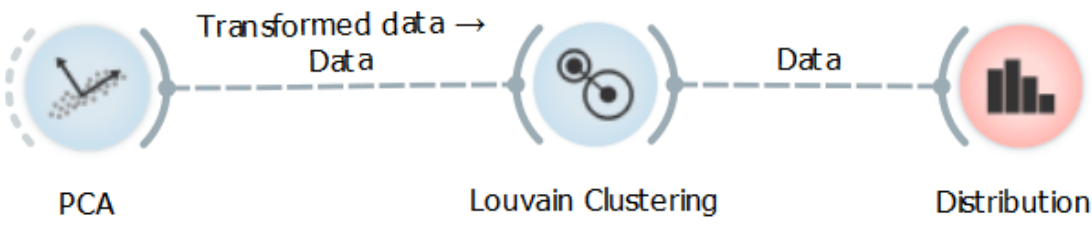


Figura 9. Captura de flujo de tareas de Clustering Louvain

Aplicando el Clustering de Louvain sobre el dataset se formaron siete clústeres, como se presenta en la Figura 10 en la gráfica de distribución.

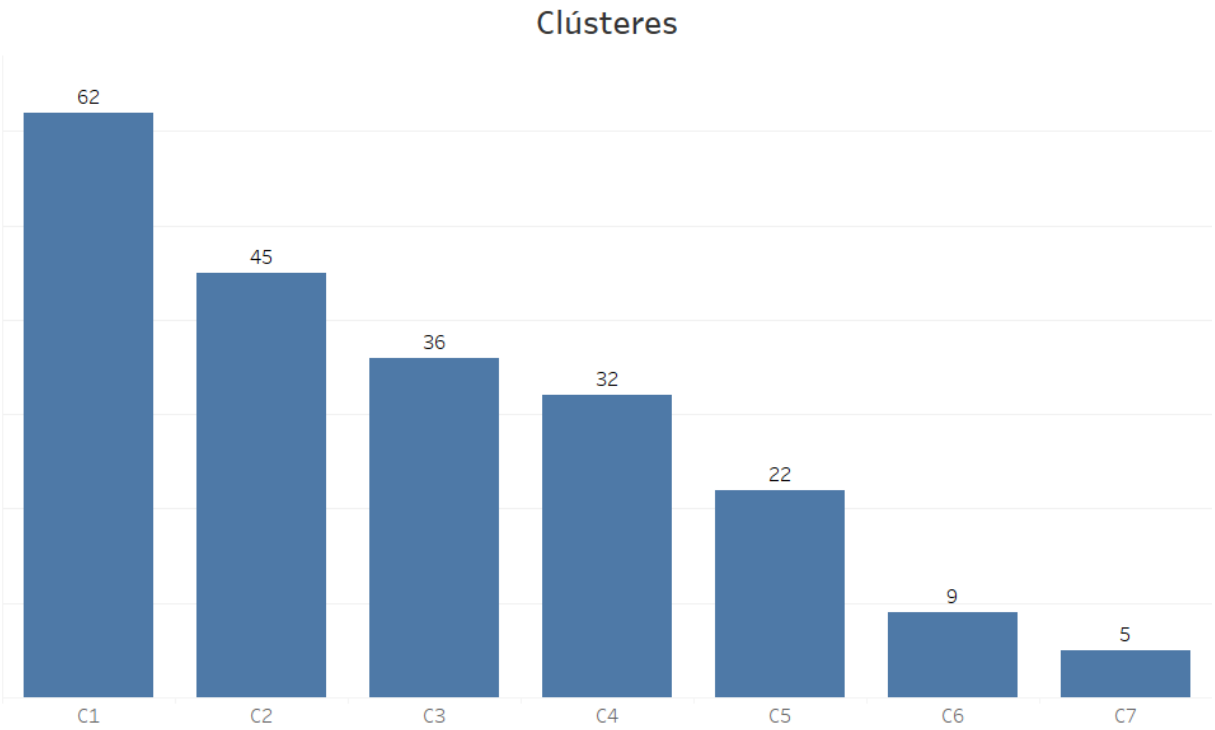


Figura 10. Diagrama de distribución de los clústeres formados con Clustering Louvain

4.4. Evaluación e interpretación

La evaluación de los clústeres e interpretación se realizó mediante la aplicación de estadística descriptiva sobre las 60 dimensiones consideradas y detalladas en la Tabla 11, las cuales, después de los dos procesos de minería de datos, se caracterizaron en distintos clústeres. A continuación, se presenta la evaluación e interpretación para los dos métodos de minería de datos tanto K-Means y Louvain, aplicando la medida de posición de la mediana debido a la gran dispersión de los datos en los clústeres y utilizando la distribución en cuartiles para mostrar la dispersión de los datos y tendencias de centralidad.

4.4.1. Métricas

Para la experimentación en el análisis descriptivo se utilizaron las medidas de tendencia central de media y mediana, también la medida de dispersión de la desviación típica.

Media

La media es la medida que indica el valor que se encuentra en la mitad de una distribución de datos.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde:

n es el número de muestras

x_i es el elemento del conjunto *X* en la posición *i*

Mediana

La mediana es la medida que encuentra el valor central de una distribución ordenada de números en la cual se aplican los siguientes casos:

1. Si el número de datos es impar, se toma el número central de los números.
2. Si el número de datos es par, se toman los dos números centrales y se realiza el promedio de los dos números.

Desviación típica

La desviación típica expresa la dispersión de la distribución en las unidades de la variable.

$$S = \sqrt{\frac{\sum_{j=1}^{n} \left(X_j - \text{Media}(X)\right)^2}{n}}$$

Donde:

n es el número de muestras

x_j es el elemento del conjunto *X* en la posición *j*

4.4.2. Evaluación e interpretación clustering con K-Means

Aplicando la mediana como medida de centralidad a las diferentes dimensiones de los clústeres, se realizó el análisis para explicar las características de los diferentes clústeres formados. En el análisis se observa que el clúster 10 está conformado por investigadores del grupo de investigación 10 que aportan artículos científicos de impactos Q1 y Q2 en SJR, los cuales tienen la mayor tasa de citación. El clúster 17 que se conforma por 15 personas del grupo de investigación 35, tiene una alta tasa de citación. Los clústeres 5, 9 y 17 están conformados por investigadores de los grupos de investigación 30, 41 y 35 respectivamente, las personas dentro de estos clústeres producen artículos Q2 en SJR los cuales son de alto impacto, teniendo un total de 35 personas. Los clústeres 4, 24 y 25 están conformados por investigadores que tienen una alta media de proyectos de investigación, teniendo un total de 14 personas.

Existen dos clústeres que al estar compuestos por una sola persona se vuelven casos especiales a analizar, debido a su despunte de valores en las medianas. El primer caso corresponde al clúster 3, debido a que tiene los mayores valores de conteo de coautores, artículos científicos sin cuartil, artículos científicos en Q2, artículos científicos en Q3 y una buena tasa de citación, se despunta de los otros investigadores posicionándolo como el mejor investigador con respecto a estas dimensiones. El segundo caso corresponde al clúster 14, el cual tiene la mayor tasa de artículos Q1 en SJR y un alto trabajo con otros autores por lo cual despunta como el investigador con mayor producción en impacto científico.

El análisis completo de los 29 clústeres formados a partir de la técnica de K-Means se presenta a continuación, utilizando como referencia la medida de centralidad de la mediana.

Clúster 1:

- Tiene 4 observaciones.
- La mitad del clúster está conformado por coordinadores de grupos.
- Todos los integrantes son miembros del Grupo de Investigación 18.

Clúster 2:

- Tiene 10 observaciones.

- Todos los integrantes son responsables de comunicación.
- Todos los integrantes son miembros del Grupo de Investigación 13.

Clúster 3:

- Tiene 1 observación.
- Tienen una buena mediana de citación de 67 y conteo de citación al autor de 58 con respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de coautores con 163 respecto al resto de grupos.
- Tienen la mayor mediana de conteo de proyectos de investigación con 20 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos científicos sin cuartil en SJR con 44 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos Q2 en SJR con 44 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos Q3 en SJR con 2 respecto al resto de clústeres.
- Todos los integrantes son coordinadores de grupos.
- Todos los integrantes son miembros de los Grupos de Investigación 39 y 41.

Clúster 4:

- Tiene 3 observaciones.
- Tienen una buena mediana de conteo de proyectos de investigación con 7 respecto al resto de clústeres.
- Está conformado enteramente por miembros del Grupo de Investigación 5.

Clúster 5:

- Tiene 16 observaciones.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 5 respecto al resto de clústeres.
- Todos los integrantes son miembros del Grupo de Investigación 30.

Clúster 6:

- Tiene 28 observaciones.

- Todos los integrantes son responsables de comunicación.
- Todos los integrantes son miembros del Grupo de Investigación 41.

Clúster 7:

- Tiene 9 observaciones.
- Todos los integrantes son miembros del Grupo de Investigación 44.

Clúster 8:

- Tiene 4 observaciones.
- La mitad del clúster son investigadores.
- La mitad del clúster son responsables de comunicación.
- Todos los integrantes son miembros del Grupo de Investigación 8.

Clúster 9:

- Tiene 4 observaciones.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 5 respecto al resto de clústeres.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 41.

Clúster 10:

- Tiene 8 observaciones.
- Tienen la mediana de citación de 129 y conteo de citación al autor de 106.5, que son las más alta de todos los clústeres.
- Tienen una buena mediana de conteo de artículos Q1 en SJR con 3.5 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 5 respecto al resto de clústeres.
- La mitad del clúster son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 6.

Clúster 11:

- Tiene 6 observaciones.

- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 45.

Clúster 12:

- Tiene 13 observaciones.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 40.

Clúster 13:

- Tiene 3 observaciones.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 21.

Clúster 14:

- Tiene 1 observación.
- Tienen una buena mediana de conteo de coautores con 46 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos Q1 en SJR con 14 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 4.
- Todos los integrantes son miembros del Grupo de Investigación 33.

Clúster 15:

- Tiene 4 observaciones.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 13.

Clúster 16:

- Tiene 16 observaciones.
- Todos los integrantes son investigadores.

Clúster 17:

- Tiene 15 observaciones.

- Tienen una buena mediana de citación de 44 y conteo de citación al autor de 28 con respecto al resto de clústeres.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 4 respecto al resto de clústeres.
- Todos los integrantes son miembros del Grupo de Investigación 35.

Clúster 18:

- Tiene 8 observaciones.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 34.

Clúster 19:

- Tiene 4 observaciones.
- La mitad del clúster son coordinadores de grupos.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros de los Grupos de Investigación 16 y 39.
- La mitad del clúster son miembros del Grupo de Investigación 41.

Clúster 20:

- Tiene 2 observaciones.
- La mitad del clúster son coordinadores de grupos.
- La mitad del clúster son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 42.

Clúster 21:

- Tiene 1 observación.
- Todos los integrantes son responsables de comunicación.
- Todos los integrantes son miembros de los Grupo de Investigación 13 y 30.

Clúster 22:

- Tiene 3 observaciones.
- Todos los integrantes son coordinadores de grupos.

Clúster 23:

- Tiene 1 observación.
- Todos los integrantes son coordinadores de grupos.
- Todos los integrantes son miembros de los Grupos de Investigación 22, 25, 33 y 38.

Clúster 24:

- Tiene 9 observaciones.
- Tienen una buena media de conteo de proyectos de investigación con 6 respecto al resto de clústeres.

Clúster 25:

- Tiene 2 observaciones.
- Tienen una buena mediana de conteo de proyectos de investigación con 6 respecto al resto de clústeres.
- Todos los integrantes son coordinadores de grupos.
- La mitad del clúster son miembros de los Grupos de Investigación 4 y 8.
- Todos los integrantes son miembros del Grupo de Investigación 20.

Clúster 26:

- Tiene 16 observaciones.
- Todos los integrantes son responsables de comunicación.
- Todos los integrantes son miembros del Grupo de Investigación 30.

Clúster 27:

- Tiene 1 observación.
- Todos los integrantes son miembros de los Grupos de Investigación 6, 15, 13 y 25.

Clúster 28:

- Tiene 13 observaciones.
- Todos los integrantes son miembros del Grupo de Investigación 41.

Clúster 29:

- Tiene 6 observaciones.
- Todos los integrantes son miembros del Grupo de Investigación 26.

En la Figura 11 se puede corroborar, mediante la gráfica de bigotes, como sobresalen los clústeres 10 y 17, los cuales tienen una alta dispersión, pero la medida de centralidad de la media los ubica por encima del resto de clústeres.

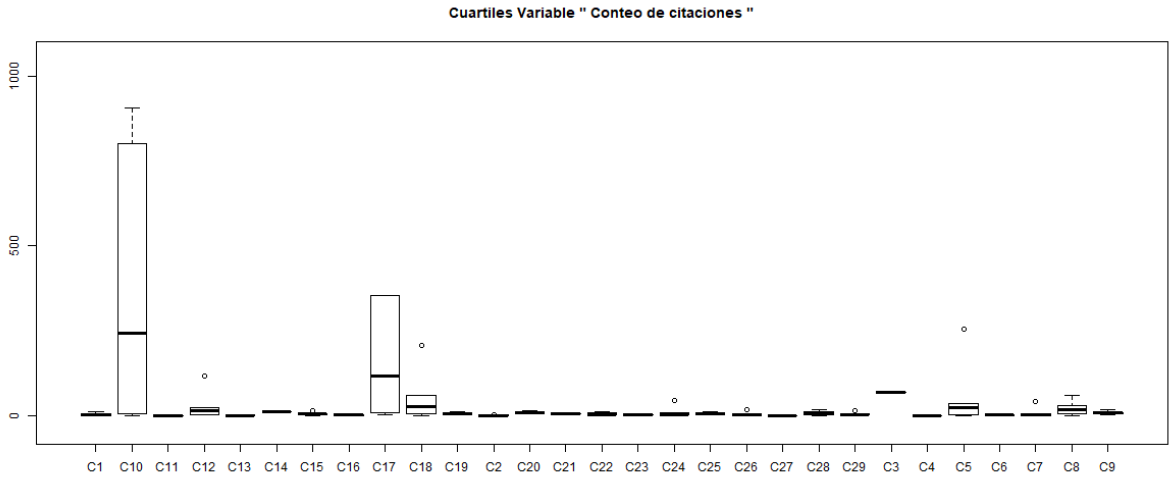


Figura 11. Cuartiles de la variable de conteo de citaciones aplicando K-Means en la técnica 1.

En la Figura 12 se puede apreciar cómo la dispersión del clúster 10 es muy alta por lo cual, aunque la media indique un valor de 5, la mediana muestra que la medida de centralidad es de 3.5 artículos. El clúster 14 presenta una media más alta, pero al estar conformado por una sola observación tiene una media igual al máximo y mínimo valor de su cuartil.

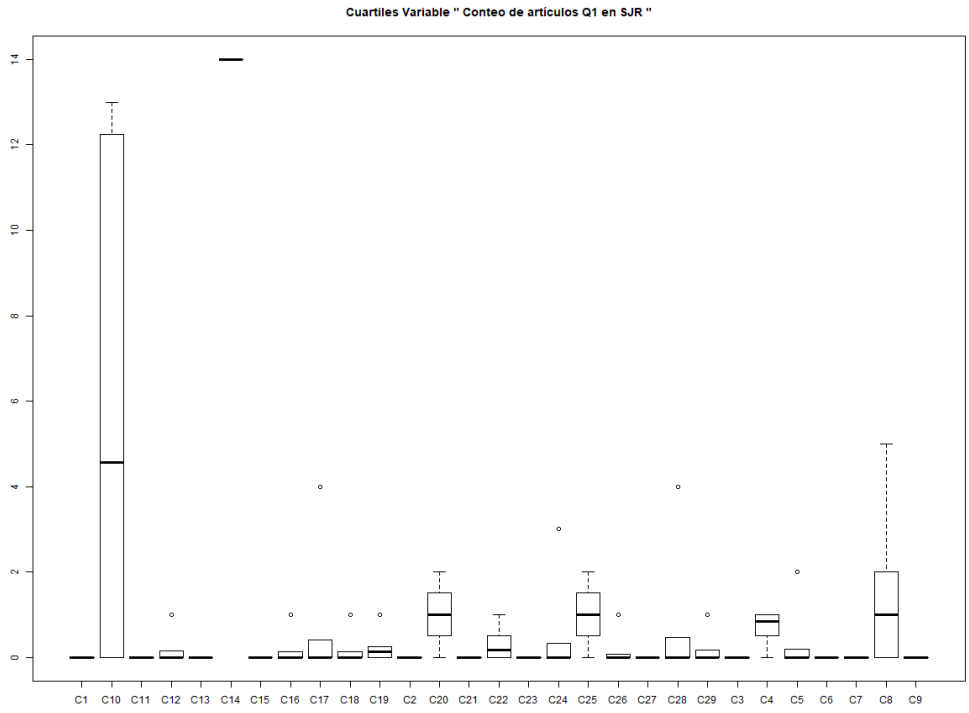


Figura 12. Cuartiles de variable de conteo de artículos Q1 en SJR aplicando K-Means en la técnica 1

4.4.2. Evaluación e interpretación clustering con Louvain

Al igual que en el análisis del clúster con K-Means, se utilizó la medida de centralidad de la mediana, por la dispersión de los datos y posteriormente para aclarar ciertas características importantes de los clústeres, se utilizó un diagrama de bigotes para mostrar los cuartiles y la distribución de las dimensiones.

En el análisis estadístico descriptivo se caracterizaron algunos los comportamientos de los clústeres con respecto a las dimensiones. El clúster 3 se posiciona como el de mayor rendimiento, debido a que tiene la mayor mediana en citas y citas al autor, además de conteo de coautores, artículos sin cuartil y con cuartil Q2 en SJR. También todos los miembros del clúster son investigadores con una buena tasa de conteo de proyectos por investigador.

El clúster 1 y 2 tienen buenas medianas de conteo de coautores, producción de artículos en cuartiles Q2 y sin cuartil en SJR, pero los miembros del clúster 2 tienen una mediana de proyectos de las más altas con respecto al resto de clústeres, mientras que el clúster 1 tiene una mediana de 0 en el conteo de proyectos. El clúster 4 está conformado por 22 investigadores que tiene buen trabajo con respecto al número de coautores con los que trabajan y también tienen una buena mediana de proyectos. Los clústeres 5 y 6 tienen un buen número de citaciones y citaciones al autor con respecto al resto de clústeres, también tienen la mediana más alta con respecto a proyectos de investigación. El clúster 7 tiene la mayor tasa de producción en artículos en cuartil Q3 de SJR y un alto número de coautores por investigador.

El análisis completo de los diferentes clústeres formados a partir de la técnica de Louvain se presenta a continuación, utilizando como referencia la medida de centralidad de la mediana.

Clúster 1:

- Tiene 62 observaciones.
- Tienen una buena mediana de conteo de coautores con 7 respecto al resto de clústeres.
- La mediana de proyectos de investigación es de 0.
- Tienen una buena mediana de conteo de artículos sin cuartil en SJR con 2 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 2 respecto al resto de clústeres.

Clúster 2:

- Tiene 45 observaciones.
- Tienen una buena mediana de conteo de coautores con 6 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de proyectos de investigación con 2 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de artículos sin cuartil en SJR con 2 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de artículos Q2 en SJR con 2 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son de los Grupo de Investigación 30 y 41.

Clúster 3:

- Tiene 36 observaciones.
- Tienen la mayor mediana de conteo de citaciones con 12.5 y conteo de citaciones al autor con 11 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de coautores con 13 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de proyectos de investigación con 1 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos sin cuartil SJR con 4 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de artículos Q2 en SJR con 4 respecto al resto de clústeres.
- Todos los integrantes son investigadores.

Clúster 4:

- Tiene 32 observaciones.
- Tienen una buena mediana de conteo de coautores con 5 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de proyectos de investigación con 1 respecto al resto de clústeres.
- Tienen una baja mediana de conteo de artículos sin cuartil en SJR con 1 respecto al resto de clústeres.

- Todos los integrantes son investigadores.

Clúster 5:

- Tiene 22 observaciones.
- Tienen una buena mediana de conteo de citaciones con 5 y conteo de citaciones al autor con 5 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de coautores con 6.5 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de proyectos de investigación con 2 respecto al resto de clústeres.
- Tienen una baja mediana de conteo de artículos sin cuartil en SJR con 1 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son del Grupo de Investigación 40.

Clúster 6:

- Tiene 9 observaciones.
- Tienen una buena mediana de conteo de citaciones con 7 y conteo de citaciones al autor con 7 respecto al resto de clústeres.
- Tienen una buena mediana de conteo de coautores con 8 respecto al resto de clústeres.
- Tienen la mayor mediana de conteo de proyectos de investigación con 2 respecto al resto de clústeres.
- Tienen una baja mediana de conteo de artículos sin cuartil en SJR con 1 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son del Grupo de Investigación 34.

Clúster 7:

- Tiene 5 observaciones.
- Tienen una buena mediana de conteo de coautores con 5 respecto al resto de clústeres.
- La mediana de artículos sin cuartil en SJR es de 0.

- Tienen la mayor mediana de conteo de artículos Q3 en SJR con 1 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son del Grupo de Investigación 28.

El clúster 3 presenta una mediana de 13 coautores, pero se puede observar como en la Figura 13 aunque es el que tiene la media más alta también es el clúster que presenta la dispersión más alta en sus observaciones de esta dimensión. También se puede ver que el resto de los clústeres presentan una media muy similar entre sí, con dispersiones muy bajas.

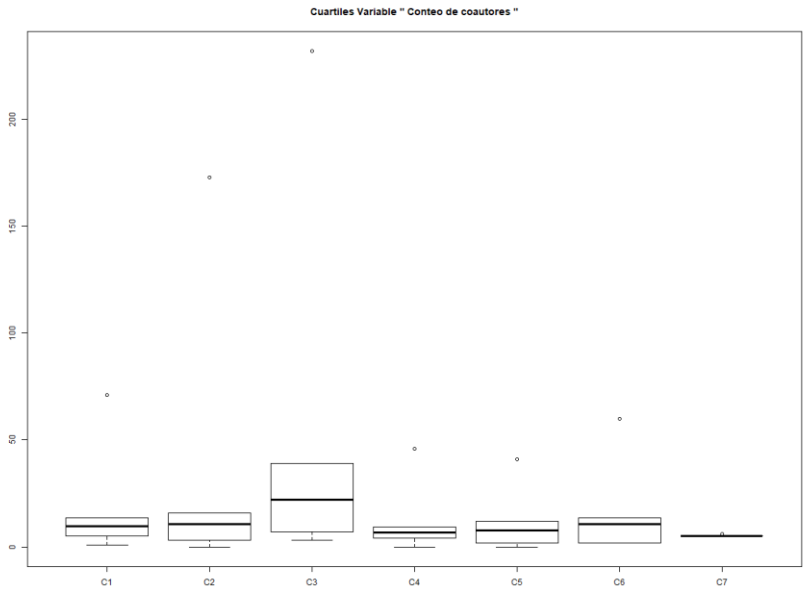


Figura 13. Cuartiles de la variable de conteo de coautores aplicando Louvain en la técnica 1.

Para el caso del clúster 7 en la Figura 14 se presenta el gráfico de bigotes de la dimensión de conteo de artículos en cuartil Q3 en SJR, en donde este clúster presenta la mayor media y una baja dispersión de sus observaciones, pero esto responde también a que es el grupo con menos observaciones con respecto al resto de clústeres.

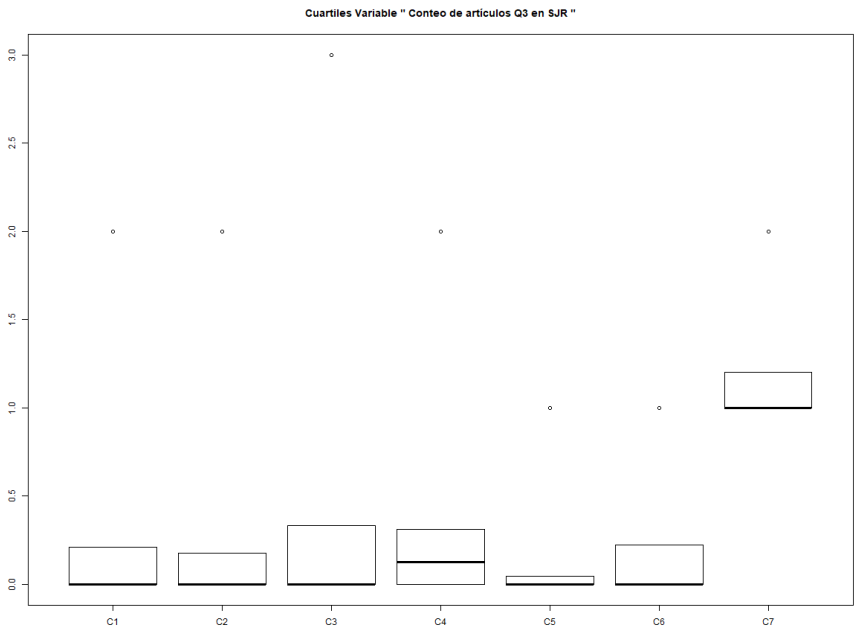


Figura 14. Cuartiles de la variable de conteo de artículos Q3 en SJR aplicando Louvain en la técnica 1.

5. Descubrimiento del conocimiento de los perfiles de investigadores, aproximación basada en esquemas semánticos

El presente capítulo desarrolla la técnica de KDD para encontrar patrones sobre la información de los investigadores de la UPS, realizando una etapa extra de adición de conocimiento, utilizando ontologías y técnicas de similitud con VSM para categorizar los contenidos de la producción científica de los investigadores, para, de esta manera, crear conocimiento que no se encontraba en el dataset original.

5.1. Modelado y agregación de conocimiento

El modelado del conocimiento es un proceso en el que se busca construir un modelo que conceptualiza una abstracción de una parte de la realidad. Basados en algunas partes del marco de desarrollo de ontologías de NeOn [21] se proponen algunos procesos como se presenta en la Figura 15, para la construcción de la red ontológica y base de conocimiento. NeOn propone un conjunto de actividades de ingeniería ontológica y un árbol de diferentes casos que guían al ingeniero de conocimiento a través de un flujo de actividades.

En la Figura 15 se observa las líneas que guían el modelado del conocimiento, partiendo desde el proceso de elicitación de requerimientos que entrega el documento de elicitación de requerimientos ontológicos. Los recursos ontológicos por utilizar en la base de conocimiento se obtienen del API de Scopus, el tesauro de la UNESCO y mediante el proceso de selección de elementos ontológicos, el cual busca un esquema que se ajuste a los requerimientos de la ontología. Los recursos no ontológicos se obtienen de tres diferentes fuentes, la base de datos del Sistema Institucional de Investigación de la UPS, términos no cubiertos de los requerimientos y el documento de revistas del SJR, aplicando un proceso de reingeniería de elementos no ontológicos a cada uno. El modelamiento culmina con la integración de los elementos del proceso de reingeniería de elementos no ontológicos y los elementos ontológicos seleccionados, formando una base de conocimiento en un repositorio mediante la herramienta GraphDB.

La agregación de conocimiento proporciona una capa de información extra sobre el dominio de conocimiento actual. En la base de conocimiento se realiza la agregación de conocimiento, utilizando un proceso de similitud, uniendo los artículos científicos con los términos proporcionados en el tesauro UNESCO. Al realizar este proceso se posibilita la

caracterización de los productos científicos; de esta manera, se descubre el área de producción de los investigadores de la UPS.

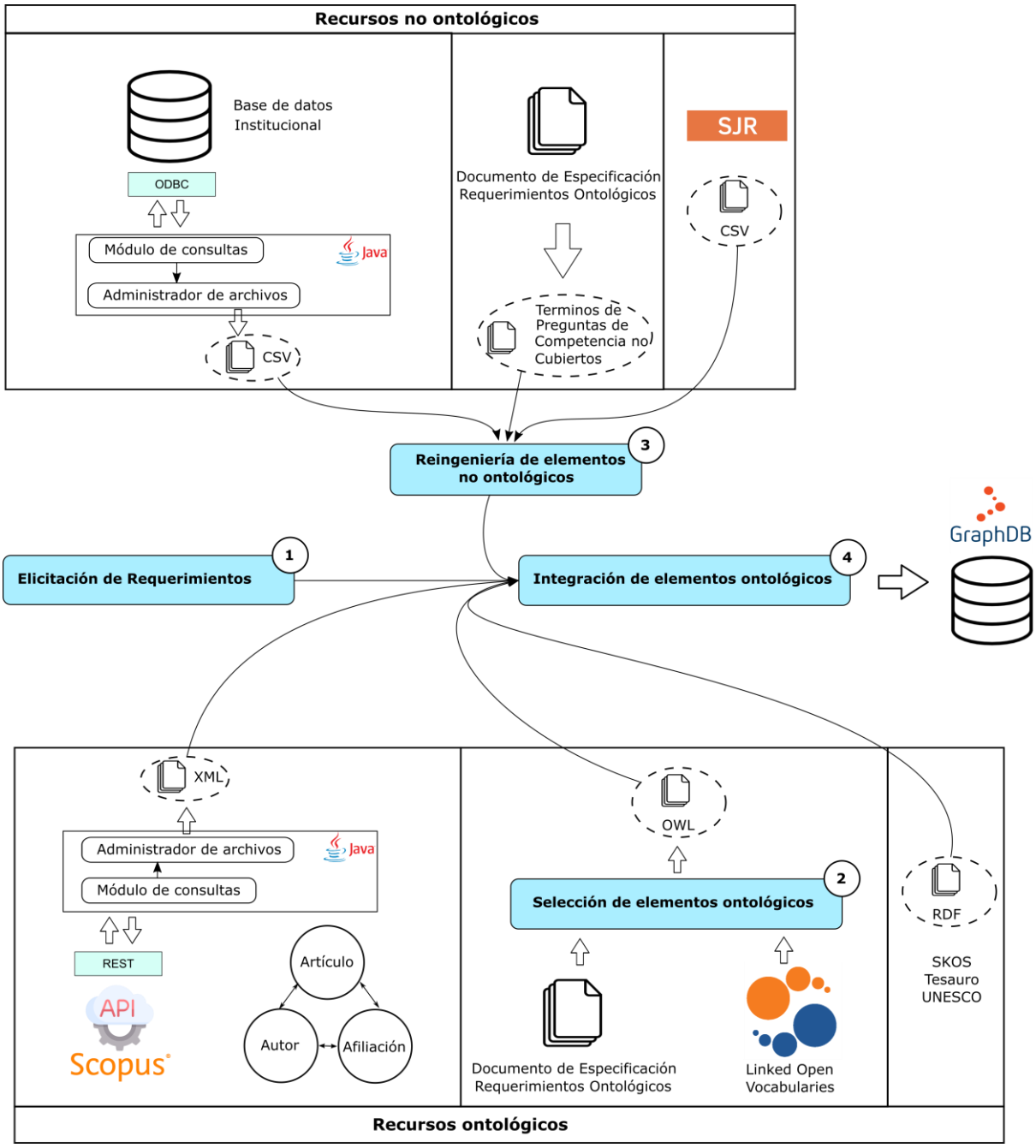


Figura 15. Esquema de proceso de modelado de conocimiento.

5.1.1. Elicitación de requerimientos ontológicos

El proceso de elicitación de requerimientos define el propósito de la construcción de la ontología o red ontológica, por lo cual se utilizó el marco referencial de tareas propuestos en la metodología NeOn [23]. Estás tareas parten desde la identificación del propósito, alcance,

lenguaje de implementación, identificar usuarios finales, identificar usos, requerimientos, agrupar requerimientos, priorizar requerimientos y extraer la terminología a usar.

El resultado de las diferentes tareas del proceso de elicitación de requerimientos tuvo como resultado el Documento de Especificación de Requerimientos Ontológicos (DERO) que se presenta en la Tabla 6. A fin de responder a las Preguntas de Competencia (PC) para la construcción del DERO, se utilizó como información de entrada los esquemas e instancias de los esquemas de las bases de datos institucionales y del SDR de Scopus.

Tabla 6. Documento de elicitación de requerimientos ontológicos.

1	**Propósito**
La red ontológica tiene el propósito de modelar el conocimiento de los comportamientos científicos de los investigadores de la UPS, en torno a su producción científica y caracterizar dicha producción científica en torno a un tesauro de términos.	
2	**Alcance**
La red ontológica está centrada alrededor de la información plasmada en la base de datos institucional de la UPS y de producción científica que se encuentra en el repositorio de Scopus.	
3	**Lenguaje de implementación**
Los diferentes componentes de la red ontológica se implementarán en el lenguaje OWL.	
4	**Usuarios finales previstos**
Usuario 1. Profesores y estudiantes investigadores. Usuario 2. Administrativos del departamento de investigación.	
5	**Usos previstos**
Uso 1. Caracterizar el contenido de los artículos científicos en torno a conceptos más generales. Uso 2. Representar la información de los perfiles de los investigadores y unidades de investigación. Uso 3. Representar la información de artículos científicos de los investigadores. Uso 4. Representar las estructuras jerárquicas de un tesauro de conocimientos.	
6	**Requerimientos de la ontología**
	a. Requerimientos no funcionales
	RNF 1. La ontología debe soportar al menos el lenguaje inglés en las etiquetas de sus clases y propiedades. RNF 2. El tesauro de términos debe estar bajo la ontología SKOS.
	b. Requerimientos funcionales
	GPC1. Investigador

<table>
<tr><td></td><td colspan="4">PC1. ¿Quiénes son los actores del ecosistema de investigación?
PC2. ¿Cuáles son los items que describen a un investigador?
GPC2. Organización de investigación
PC3. ¿Cuáles son las unidades organizativas de investigación?
PC4. ¿Cuáles son los items que describen a una organización de investigación?
GPC3. Documento de investigación
PC5. ¿Cuáles son los productos científicos de los investigadores?
PC6. ¿Cómo categorizan el impacto según Scopus?
PC7. ¿Cuáles son los items que describen un producto documental científico?</td></tr>
<tr><td>7</td><td colspan="4">Pre-Glosario de términos</td></tr>
<tr><td></td><td colspan="4">a. Objetos + Preguntas de competencia</td></tr>
<tr><td></td><td>Investigadores</td><td>PC1</td><td>Artículos de revista</td><td>PC5</td></tr>
<tr><td></td><td>Investigadores externos</td><td>PC1</td><td>Review</td><td>PC5</td></tr>
<tr><td></td><td>Estudiante investigador</td><td>PC1</td><td>Sin impacto</td><td>PC6</td></tr>
<tr><td></td><td>Coordinador de grupo</td><td>PC1</td><td>Q1</td><td>PC6</td></tr>
<tr><td></td><td>Responsable de comunicación</td><td>PC1</td><td>Q2</td><td>PC6</td></tr>
<tr><td></td><td>Técnico docente</td><td>PC1</td><td>Q3</td><td>PC6</td></tr>
<tr><td></td><td>Nombres</td><td>PC2</td><td>Q4</td><td>PC6</td></tr>
<tr><td></td><td>Apellidos</td><td>PC2</td><td>Titulo</td><td>PC7</td></tr>
<tr><td></td><td>Rol</td><td>PC2</td><td>Resumen</td><td>PC7</td></tr>
<tr><td></td><td>Grupo de investigación</td><td>PC3</td><td>Palabras clave</td><td>PC7</td></tr>
<tr><td></td><td>Universidad</td><td>PC3</td><td>DOI</td><td>PC7</td></tr>
<tr><td></td><td>Nombre</td><td>PC4</td><td>ISSN</td><td>PC7</td></tr>
<tr><td></td><td>País</td><td>PC4</td><td>ISBN</td><td>PC7</td></tr>
<tr><td></td><td>Estado</td><td>PC4</td><td>Páginas</td><td>PC7</td></tr>
<tr><td></td><td>Ciudad</td><td>PC4</td><td>URL</td><td>PC7</td></tr>
<tr><td></td><td>Dirección</td><td>PC4</td><td>Volumen</td><td>PC7</td></tr>
<tr><td></td><td>Capítulos de libro</td><td>PC5</td><td>Año</td><td>PC7</td></tr>
<tr><td></td><td>Libros</td><td>PC5</td><td>Fecha</td><td>PC7</td></tr>
<tr><td></td><td>Artículos de conferencia</td><td>PC5</td><td></td><td></td></tr>
</table>

El DERO que se presenta en la Tabla 6, describe las PC que se clasifican en los Grupos de Preguntas de Competencia (GPC) de investigador, documento de investigación y organización de investigación. Las PC se contestaron mediante los esquemas de la base de datos institucional de la UPS y el SDR de Scopus, generando los términos del pre-glosario de términos que servirán para las posteriores etapas de la construcción de la red ontológica.

5.1.2. Selección de los elementos ontológicos a reutilizar

En la selección de elementos a integrar a la red ontológica se tomó como entrada del DERO, utilizando como requerimientos de búsqueda los términos encontrados en el pre-glosario. Para instrumentalizar la búsqueda partiendo de los términos del DERO, se convirtió cada uno de los elementos de grupo, preguntas de competencia y términos a instancias de una ontología, reutilizando la ontología Simple Knowledge Organization System (SKOS) [24]. El motivo de manejar los requerimientos como instancias de una ontología posibilita la búsqueda de conceptos relacionados con las ontologías candidatas, descubriendo de manera objetiva el factor de cobertura de requerimientos de los axiomas de dichas ontologías candidatas.

Para la construcción del esquema ontológico de mapeo de términos y PC del DERO se utilizó como base el esquema de SKOS, el cual está compuesto, como se observa en la Figura 15, por clases que detallan conceptos, esquemas de conceptos y colecciones. Partiendo del esquema SKOS se construyó la Ontología de Elicitación de Requerimientos Ontológicos denominada OER, presentada en la Figura 16, la cual toma el concepto skos:Concept y crea tres subclases propias denominadas oer:term, oer:competency_question y oer:competency_question_group, las cuales simbolizan las clase de los términos, preguntas de competencia y grupos, respectivamente.

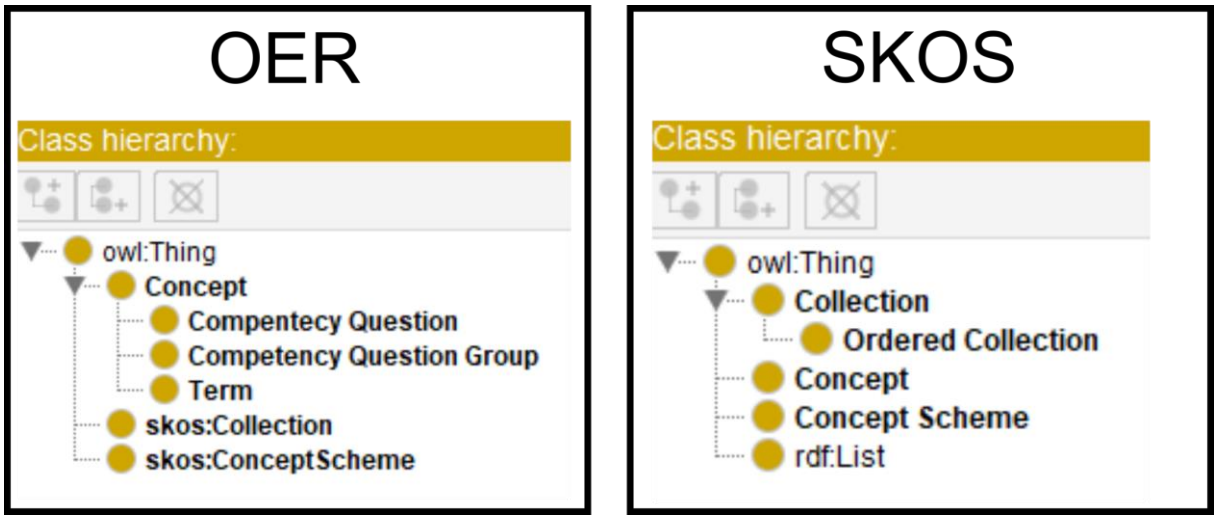
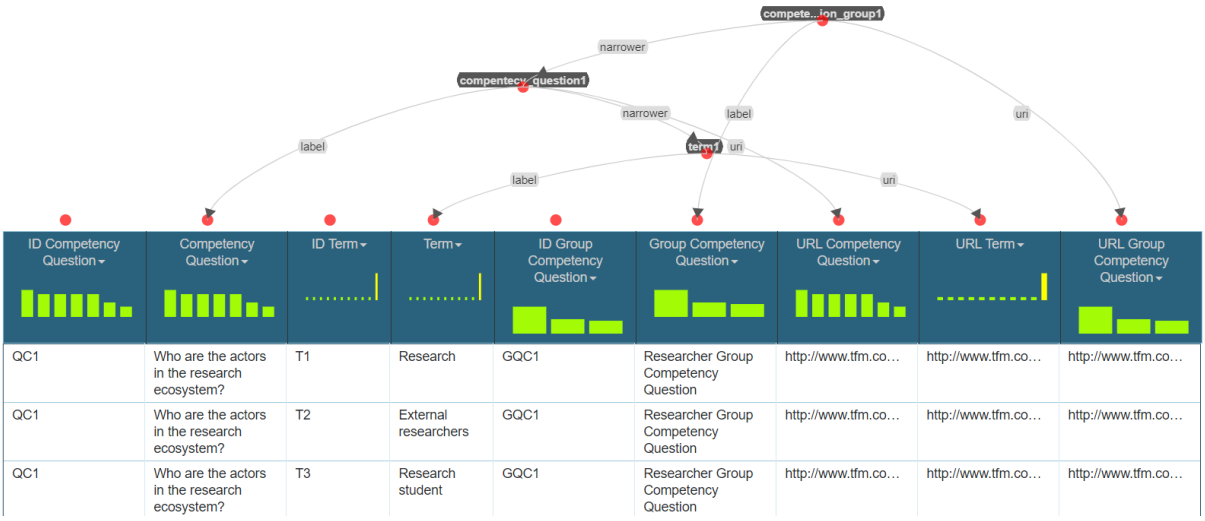


Figura 16. Captura del esquema basado en SKOS. (a) A la izquierda se encuentra el esquema OER. (b) A la derecha se presenta el esquema SKOS.

La población de las instancias de la ontología OER se obtuvieron a partir de la tabla formada por las columnas de GPC, PC y término en formato CSV. Tomando como entrada el documento en formato CSV más las ontologías SKOS y OER, se introdujeron al programa Karma [25] para convertir la estructura en tabla a tripletas. En la Figura 17 se presenta el

esquema de mapeo de la información en formato CSV a instancias de tripletas, lo cual produce un documento de tripletas en notación Turtle.



ID Competency Question	Competency Question	ID Term	Term	ID Group Competency Question	Group Competency Question	URL Competency Question	URL Term	URL Group Competency Question
QC1	Who are the actors in the research ecosystem?	T1	Research	GQC1	Researcher Group Competency Question	http://www.tfm.co…	http://www.tfm.co…	http://www.tfm.co…
QC1	Who are the actors in the research ecosystem?	T2	External researchers	GQC1	Researcher Group Competency Question	http://www.tfm.co…	http://www.tfm.co…	http://www.tfm.co…
QC1	Who are the actors in the research ecosystem?	T3	Research student	GQC1	Researcher Group Competency Question	http://www.tfm.co…	http://www.tfm.co…	http://www.tfm.co…

Figura 17. Captura del esquema de mapeo de los requerimientos del DERO en Karma.

El documento de tripletas producido a partir de la tarea de mapeo de datos forma parte de la población de la base de conocimiento para seleccionar los elementos ontológicos. La base de conocimiento utiliza el repositorio de tripletas GraphDB [26], en el cual se cargaron las ontologías SKOS y OER que contienen la parte del esquema. En la Figura 18 se presentan las instancias de los requerimientos, tanto término, PC y GPC que se cargaron a la base de conocimiento en GraphDB.

La selección de los elementos ontológicos que van a ser candidatos para su reutilización se escogió a partir de Linked Open Vocabularies (LOV) [27] como se presenta en la Figura 19. LOV es un buscador de vocabularios de elementos ontológicos de acuerdo con las necesidades del usuario, el cual posibilita reutilizar dichos elementos ontológicos. Se realizó la búsqueda en LOV de las palabras en inglés research, researcher, journal, article y organization research, identificando tres elementos ontológicos candidatos para su posible selección y reutilización.

Las ontologías candidatas para el proceso de selección fueron Discovery Vocabulary (DISCO) [28], que es una ontología para el descubrimiento de información alrededor de la investigación; Funding, Research Administration and Projects Ontology (FRAPO) [29], una ontología que modela la parte administrativa de proyectos de investigación y financiamiento; VIVO Ontology [30], una ontología del área de investigación y academia. Acotando que las tres ontologías cumplen con RNF1 al tener sus etiquetas de literales en el lenguaje inglés.

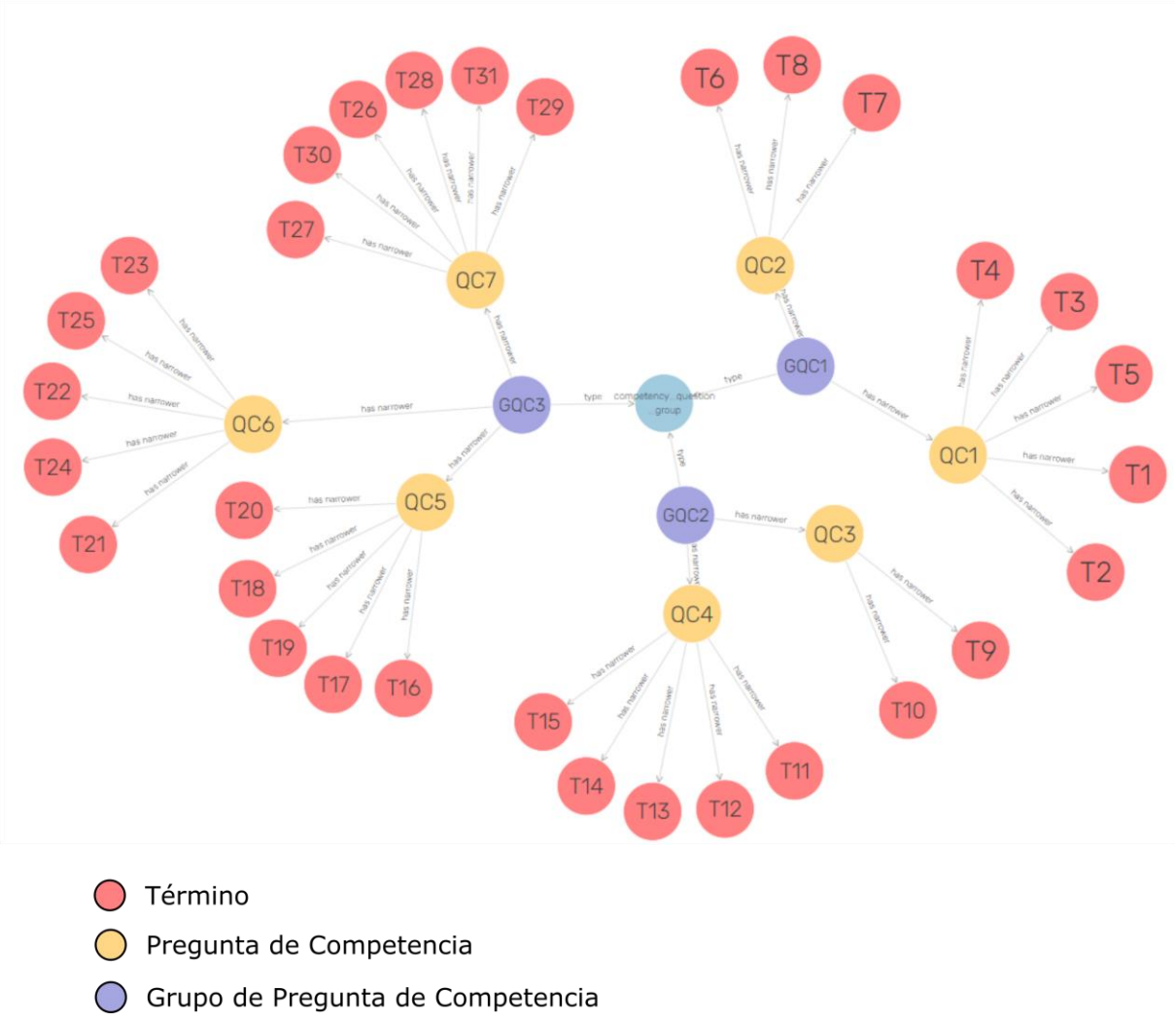


Figura 18. Instancias de DERO en la base de conocimiento en GraphDB.

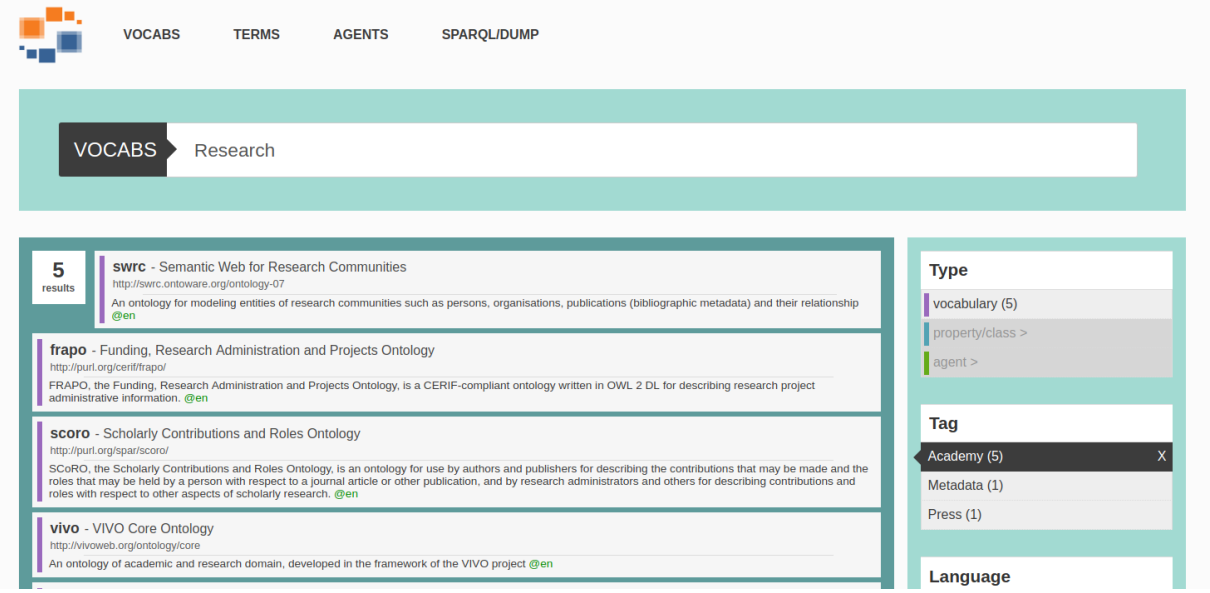


Figura 19. Captura de búsqueda del término “Research” en Linked Open Vocabulary.

Con la base de conocimiento completa se procedió a utilizar el plugin de GraphDB de búsqueda de similitud semántica para encontrar el grado de similitud que existen entre objetos de las tripletas. Para utilizar el plugin de similitud se desarrolló una consulta SPARQL que entregó como salida un objeto y un texto que pertenece a dicho objeto.

Realizando una consulta al índice de similitud, entre las ontologías candidatas y las instancias de los requerimientos, se obtuvo una matriz de cruce entre el glosario de términos y las clases y propiedades que tenían textos similares a los términos, como se presenta en la Tabla 7, utilizando un umbral de similitud de 0.6 para realizar la operación.

Tabla 7. Matriz de similaridad de términos y ontologías candidatas.

			DISCO		**FRAPO**		**VIVO**			**Total**
GPC	**PC**	**Termino**	**C**	**PO**	**C**	**PD**	**C**	**PD**	**PO**	
GPC1	PC1	External researchers	0	1	0	0	0	0	0	1
		Group Coordinator	0	1	2	0	3	1	0	7
		Research	0	0	2	0	5	2	3	12
		Research student	0	0	0	0	4	0	0	4
		Responsible for communication	0	0	0	0	1	0	0	1
	PC2	First name	0	0	0	0	1	1	0	3
		Last name	0	0	0	0	1	1	0	3
		Role	0	0	0	0	10	1	0	11
GPC2	PC3	Research group	0	1	2	0	3	0	0	6
		University	1	1	1		1	0	0	4
	PC4	Address	0	0	1	2	2	1	1	7
		City	0	0	0	1	0		0	1
		Country	0	0	0	2	1	4	1	8
		Name	0	0	0	1	2	4	2	11
		State	0	0	0	0	1	0	0	1
GPC3	PC5	Book	0	0	0	0	3	2	0	5
		Chapter	0	0	0	0	1	1	0	2

		Conference paper	0	0	1	0	4	0	0	5
		Journal article	0	0	0	0	5	0	0	5
		Review	0	0	0	0	3	0	2	5
	PC6	Non-impact	0	0	0	0	0	0	0	0
		Q1	0	0	0	0	0	0	0	0
		Q2	0	0	0	0	0	0	0	0
		Q3	0	0	0	0	0	0	0	0
		Q4	0	0	0	0	0	0	0	0
	PC7	Abstract	0	0	0	0	1	1	0	2
		DOI	0	0	0	0	0	1	0	1
		ISBN	0	0	0	0	0	2	0	2
		Keyword	0	0	0	1	0	1	0	2
		Title	0	0	0	1	1	2	1	5
Totales			1	4	9	8	53	25	10	114
Clase (C) Propiedad de objeto (PO) Propiedad de datos (PD)										

En la Tabla 9 se observa que tanto VIVO como FRAPO tienen al menos una coincidencia dentro de los 3 GCG, mientras que DISCO no tiene en su dominio ninguno de los términos del GCP de Documento de Investigación. Analizando la cobertura de términos de manera más granular, la ontología VIVO cubre 23 de los 31 términos, FRAPO cubre 11 términos y DISCO cubre solamente 4 términos. Con respecto a la cobertura a nivel de clases y propiedades de las ontologías que se relacionan con los términos del glosario, VIVO tuvo 53 clases relacionadas y 35 propiedades, FRAPO tuvo 9 clases y 8 propiedades de dato, por último, DISCO tuvo 1 clase y 4 propiedades de objeto relacionadas. Con el análisis realizado, se puede afirmar que VIVO es la ontología que mejor se acopla a los requerimientos funcionales planteados en el DERO y que fue reutilizada para la construcción de la base de conocimiento.

Para cumplir con RNF2, se seleccionó la ontología SKOS UNESCO [31] [32] que define una lista estructurada de términos separados con base en diferentes áreas de conocimiento, cuenta con 4408 conceptos distribuidos en 95 micro tesauros.

Parte de la información para la población de la base de conocimiento está en el SDR de Scopus, el cual posee tanto información de los autores cuanto de las publicaciones científicas indizadas dentro de su repositorio. Scopus ofrece dentro de sus servicios de API, la posibilidad de hacer peticiones REST en formato application/rdf+xml que contienen instancias en forma

de tripletas de los artículos y autores buscados. El formato RDF-XML ofrecido por el API de Scopus es considerado un elemento ontológico.

En este proyecto se descargaron los artículos mediante un programa construido en el lenguaje de programación Java, el cual mediante la librería de GraphDB, permitió cargar toda esta información a la base de conocimiento en una instancia de la base de datos GraphDB de manera local. Los componentes más importantes de la estructura de tripletas que presenta Scopus para describir sus artículos científicos y autores se presentan en la Tabla 8, detallando propiedades de datos y objetos con los prefijos dc perteneciente a <http://purl.org/dc/elements/1.1/> y api perteneciente a <http://www.elsevier.com/xml/svapi/rdf/dtd/>.

Tabla 8. Descripción de las propiedades de dato y objeto de las instancias de Scopus.

Articulos Scopus		**Autores Scopus**	
Propiedades de Datos	**Propiedades de Objeto**	**Propiedades de Datos**	**Propiedades de Objetos**
dc:identifier	api:writesAbout	dc:identifier	api:reference
api:authorId	api:journal	prism:issn	dc:subject
api:eid		prism:publicationName	dc:creator
api:preferredName		prism:contentType	
api:documentCount		prism:doi	
api:citedbyCount		dc:title	
api:citationCount		prism:aggregationType	
api:coauthorCount		prism:pageRange	
api:h-index		prism:coverDate	
api:publicationStart		prism:copyrightYear	
api:publicationEnd		api:openaccess	
		prism:keyword	

5.1.3. Reingeniería de elementos no ontológicos

Realizando un análisis a los axiomas que posee la ontología VIVO, se describen las clases y propiedades más importantes a utilizar para el dominio, en función a los datos que se tienen, como se presentan en la Tabla 9.

Tabla 9. Axiomas más relevantes de la ontología VIVO.

Clase	**Propiedad de Objetos**	**Propiedad de Datos**
Person	bearer of	DOI
Researcher Role	Inheres in	ISBN
Leader Role	date/time value	ISSN
Research Organization	participates in	label
Article	has participant	SCOPUS ID
Book	date/time interval	
Journal Article	relates	
Conference Paper	related by	
Journal		
Date/Time Value		
Date/Time Interval		

En los axiomas de la Tabla 8 y la matriz de la Tabla 9, se observa que no se cubren los requerimientos de los cuartiles de producción científica de Scopus presentados en la PC 6. Al no ser elementos complejos de modelado, se construyó un pequeño módulo que extendía VIVO para tomar en cuenta estos términos del glosario.

A partir de los términos de PC6 se construyó una ontología modular denominada VIVO Quartiles (VIVO-Q) como se presenta en la Figura 20, la cual está conformada por una clase superior vivoq:quartile con sub clases vivoq:quartile_1, vivoq:quartile_2, vivoq:quartile_3 y vivoq:quartile_4. El módulo VIVO-Q está alineado con la ontología VIVO para ser integrado y que funcione como parte de la red ontológica, esta alineación se lo hizo al poner como clase padre de vivoq:quartile a la clase Information Content Entity que está dentro de la ontología de VIVO. La propiedad de objeto utilizada para hacer referencia a una instancia de vivoq:quartile es vivoq:has_quartile.

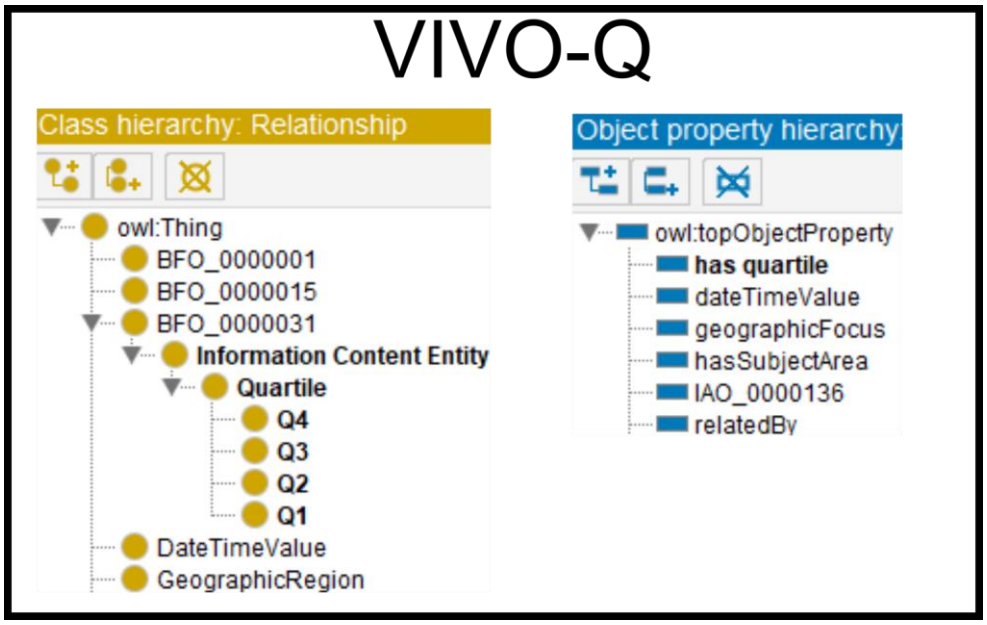


Figura 20. Esquema del módulo ontológico VIVO Quartile.

Al entender los axiomas de la ontología VIVO y el módulo VIVO-Q se puede mapear la información obtenida de la base de datos de Scopus al igual que de la base de datos del sistema de investigación institucional, para posteriormente poblar la base de conocimiento. En el proceso de mapeo se utilizó la información del Sistema de Investigación de la UPS, el cual se presenta en la Tabla 3, este fue utilizado en la etapa de integración y adición de la técnica tradicional. Esta información se encuentra como una tabla, por lo cual se utilizó la herramienta Karma para realizar el mapeo de la información a tripletas, tomando como base de consulta al Sistema de Investigación y utilizando la ontología VIVO. En la Figura 21 se presenta el modelo de mapeo, utilizando la herramienta Karma, la cual produce tripletas para la población de la base de conocimiento.

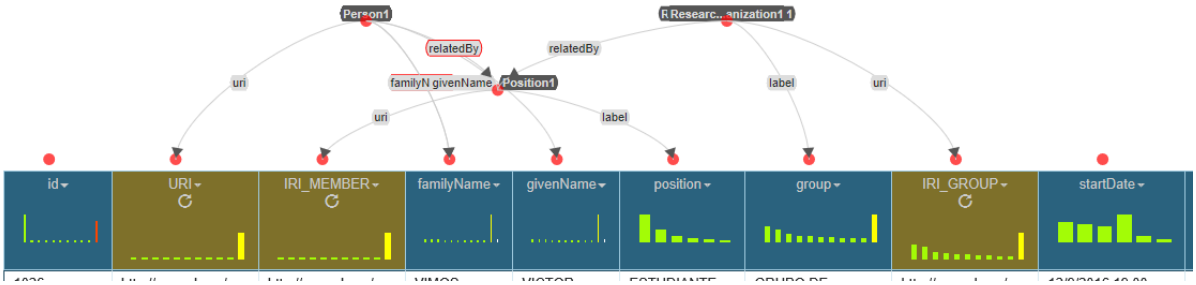


Figura 21. Esquema de mapeo de los datos del Sistema Institucional de la UPS en Karma.

Un elemento no ontológico para saber cuál es el impacto de las publicaciones de los investigadores es el SJR, el cual comparte sus métricas a través de un documento en formato

CSV. Para el mapeo del CSV de SJR se debe emplear las ontologías de esquema VIVO y VIVO-Q, mapeando los datos a manera de table de SJR por medio del software Karma. En la Figura 22 se presenta el esquema de mapeo de Karma, el cual produjo las tripletas para la población de la base ontológica.

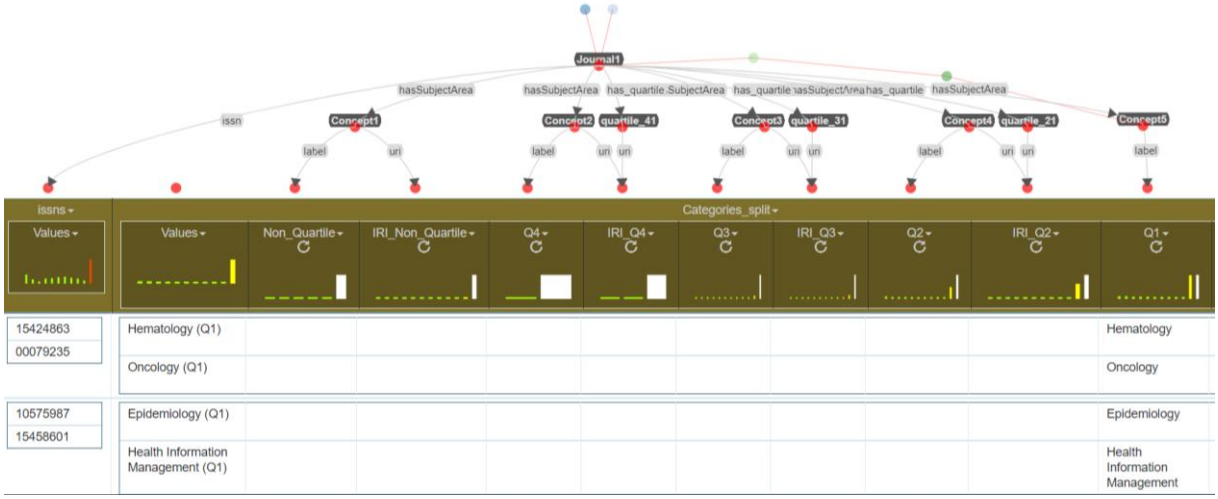


Figura 22. Esquema de mapeo del documento de SJR en Karma.

Al finalizar las tareas de reingeniería de elementos no ontológicos se obtuvo como salida tres elementos ontológicos que fueron posteriormente cargados a la base de conocimiento. El módulo VIVO-Q es un esquema ontológico que extiende y está alineado con VIVO para cubrir con los términos de los requerimientos del glosario de DERO. El documento de tripletas con las instancias de las tablas consultadas del Sistema de Investigación de UPS y el documento de tripletas del SJR que ayuda para saber el impacto de los artículos científicos de Scopus que están dentro de las revistas.

5.1.4. Integración de elementos ontológicos a la base de conocimiento

La integración de los elementos ontológicos es el proceso en el cual se unen los elementos previamente alineados y se comprueba el funcionamiento en conjunto. La integración parte desde el acoplamiento de los esquemas para formar la red ontológica, dichas ontologías cubren los requerimientos del DERO y fueron la salida del proceso de selección de elementos ontológicos. Las tres ontologías que forman el esquema ontológico VIVO, VIVO-Q y SKOS UNESCO se distribuyen en diferentes contextos de acuerdo con el dominio que tienen dentro de la base de conocimiento. Como se presenta en la Figura 23, el contexto de investigación está definido por las ontologías de VIVO y VIVO-Q, mientras que el contexto del tesauro UNESCO esta únicamente conformado por la ontología SKOS UNESCO, está distribución en

contextos permite una mayor flexibilidad a cambios en los esquemas y mantener un orden de las tripletas en contenedores de acuerdo con su dominio. Mientras que la vista a nivel de contextos permite un mayor orden, la vista de cómo se acoplan estos esquemas, presentada en la parte inferior izquierda de la Figura 23, evidencia una jerarquización de las ontologías en la red ontológica, observando que VIVO es la ontología principal que extiende su dominio a través de las ontologías SKOS UNESCO y VIVO-Q, las cuales cumplen la función de ontologías modulares dentro de la red.

La población de la base de conocimiento se realizó a partir del dominio del esquema, es por esto por lo que el proceso de reingeniería de elementos no ontológicos se hizo a partir de los esquemas de la fase de selección de elementos ontológicos. Los archivos de tripletas de la base de datos institucional de la UPS, tripletas de SJR, tripletas de los artículos y perfiles de investigadores de Scopus están distribuidos de acuerdo con el dominio en la base de conocimiento. Las instancias de la base de datos institucional e instancias de Scopus están en el contexto de investigación, mientras que las instancias del SJR se encuentran en el Contexto de SJR. El caso de la ontología SKOS UNESCO es especial puesto que, aunque extienda el esquema SKOS, está conformada en su mayoría por instancias de los conceptos del tesauro de UNESCO, por lo cual aporta tanto con instancias para la población cuanto con axiomas para el esquema de la base de conocimiento.

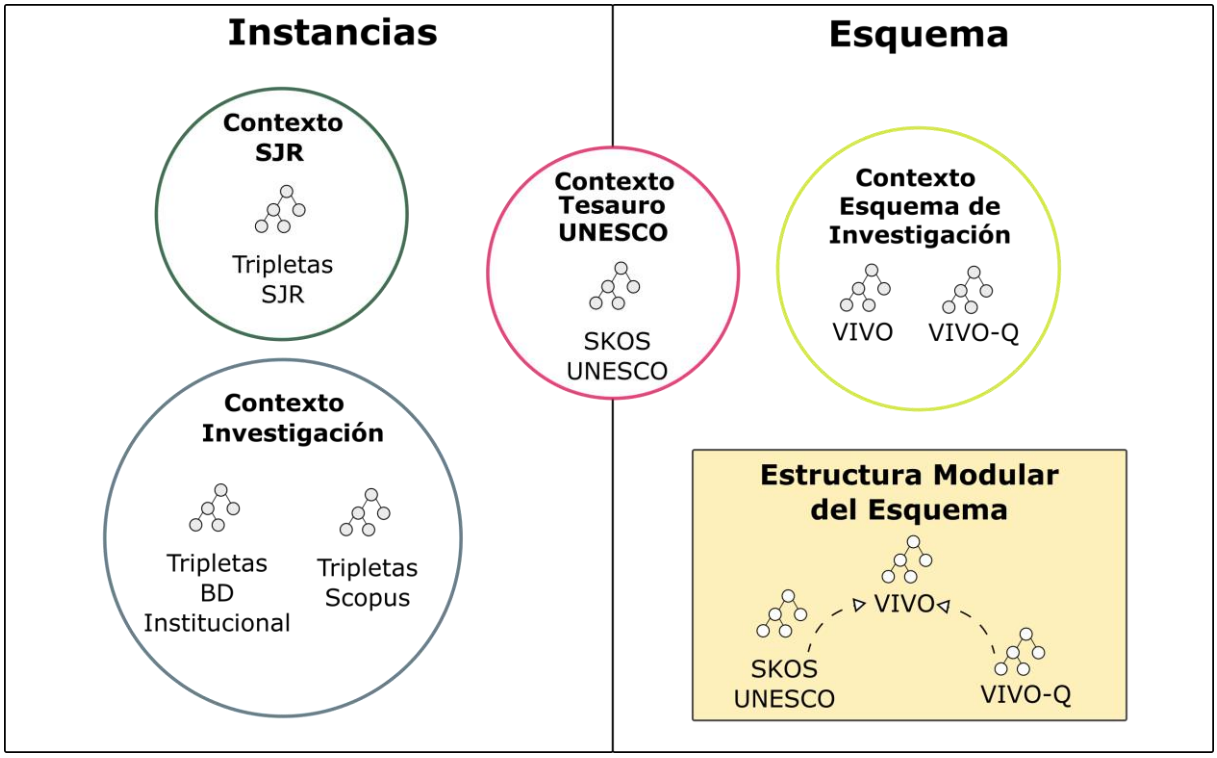


Figura 23. Esquema de distribución de los elementos ontológicos.

Para comprobar que los axiomas de las ontologías del esquema de la base de conocimiento estén integrados, se puede utilizar la herramienta Protégé para importar las tres ontologías en un mismo proyecto, como se presenta en la Figura 24; al importar los tres módulos con la herramienta Protégé se observa cómo se acoplan los elementos en la red ontológica. La ontología principal es VIVO, cuya composición tiene una ontología de alto nivel llamada Basic Formal Ontology (BFO) [33] que permite la clasificación de otras ontologías dentro de las clases de alto nivel que posee. Con VIVO como ontología principal, se integran a manera de módulos tanto la ontología de VIVO-Q y SKOS UNESCO, comprobando su integración en la Figura 24, en donde se activó el motor de reglas Hermit para comprobar que no existieran conflictos entre los axiomas.

Al comprobar la integración de los componentes del esquema de la base de conocimiento, se cargaron los módulos a la base de conocimiento en GraphDB en los diferentes contextos de acuerdo con la distribución presentada en la Figura 23.

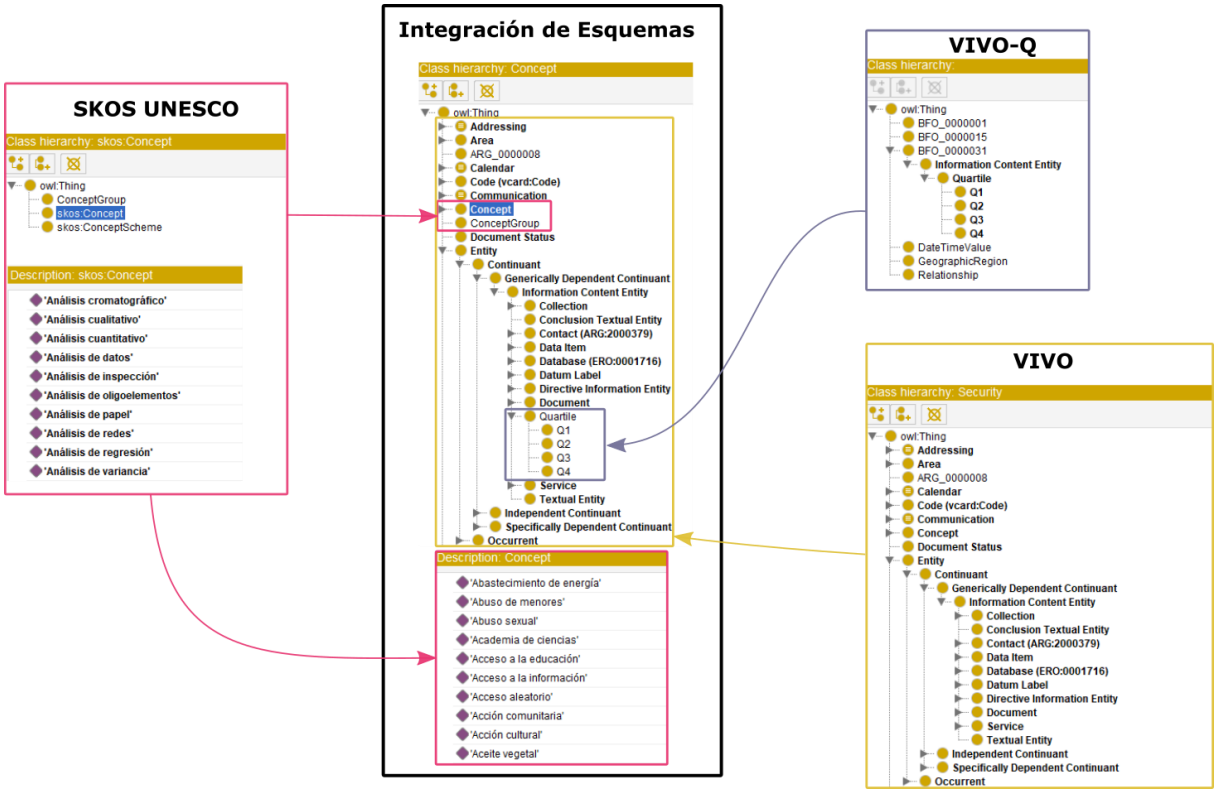


Figura 24. Integración de los esquemas en la herramienta Protégé.

Al proceder con la carga de datos al repositorio en GraphDB se presenta en la Figura 25, cómo la información de los elementos no ontológicos que se mapearon se conecta con los elementos ontológicos que se introdujeron. Para este caso el grupo, rol y persona son parte de los elementos no ontológicos que se mapearon y el nodo 5791891952, que se muestra en

la gráfica, representa el nodo descargado del api de Scopus con la información de conteo de citas, conteo de coautores y demás datos que ofrece Scopus.

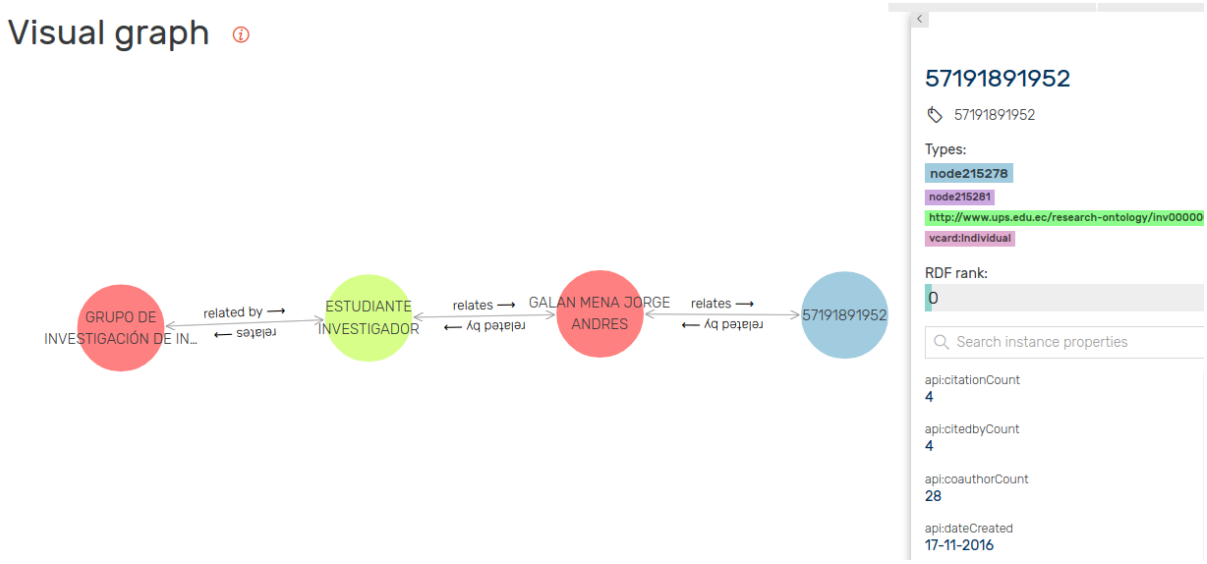


Figura 25. Captura del visor de nodos de GraphDB.

5.1.5. Agregación de conocimiento

Los artículos científicos que se encuentran en la base de conocimiento tienen información implícita, es así como los textos del título, palabras clave y resumen tienen contenido de tipo semántico. Esta semántica en los textos se puede categorizar para saber de qué rama de la ciencia se está hablando en cada uno de los documentos científicos.

La base de conocimiento cuenta con un tesauro de términos de la UNESCO, el cual categoriza en tres niveles diferentes a las diferentes ramas de la ciencia. Para la agregación de conocimiento se utiliza en el proyecto justamente el tesauro de la UNESCO y el contexto de investigación que cuenta con los artículos científicos de los investigadores de la UPS; aprovechando estos dos contextos, se aplica un proceso de similitud mediante técnicas de vectores de textos en los nodos de las tripletas. El proceso de similitud entrega una matriz de distancias entre los diferentes nodos de las ontologías de los dos contextos, marcando un puntaje normalizado de 0 a 1 de los valores de similitud entre dos nodos, agregando conocimiento al categorizar los artículos científicos con base en una clasificación de términos en un tesauro como se presenta en la Figura 26.

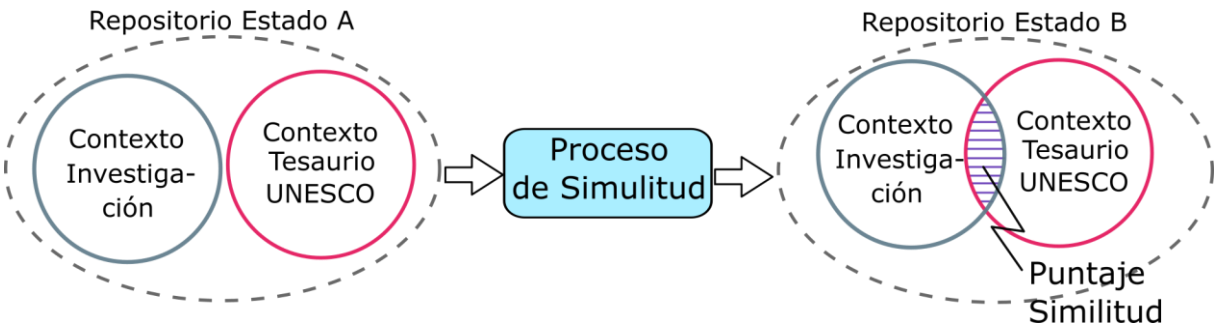


Figura 26. Proceso de similitud de contextos.

El proceso de similitud de vectores de texto entre los contextos de investigación y el tesauro de la UNESCO se realiza mediante el plugin de GraphDB de similitud. El plugin pide como entrada una consulta SPARQL con los nodos y el texto de esos nodos que van a estar en las matrices de distancia. De esta manera, con un umbral de distancia de 0.6, se encontraron 1518 coincidencias de 505 artículos en la base de conocimiento.

5.2. Selección y adición

Tomando como entrada la base de conocimiento construida anteriormente, se refleja que se tienen todos los datos en un mismo repositorio, tanto la información del Sistema Institucional, información de los artículos del SDR de Scopus y la información de los cuartiles de SJR, al contrario de la etapa de selección y adición del método anterior en la cual se tenía que tratar con información dispersa de varias fuentes.

En esta etapa se realizó un SPARQL a la base de conocimiento para cruzar los contextos de SJR e Investigación, adicionando la información del contexto del tesauro UNESCO mediante el plugin de similitud de GraphDB que se construyó en la tarea de agregación de conocimiento. Con el SPARQL se formó el dataset con los campos presentados y detallados en la Tabla 10, obteniendo los mismos resultados que en la etapa de selección y adición de la técnica anterior, pero sumando el conocimiento agregado de los contenidos de la UNESCO.

Tabla 10. Campos del dataset final de la fase de Selección y Adición.

Campo	**Descripción**
Id	Identificador en el sistema institucional de investigación.
Conteo proyectos	Número de proyectos de investigación en el que ha participado el investigador.
Rol	Rol del investigador en el grupo de investigación.

Id grupo de investigación	Identificador del grupo de investigación al que pertenece el perfil del investigador.
Id Scopus	Identificador del perfil del investigador en Scopus.
Conteo de citaciones	El total de citaciones de los documentos científicos del autor.
Conteo de coautores	Conteo de coautores con los que ha colaborado para producir artículos científicos.
Conteo de documentos	Conteo de artículos científicos producidos por el autor.
Índice H	Índice H mide la calidad de investigador en función de la producción de artículos científicos y las citaciones que tiene.
Publicación final	Última fecha de publicación de un artículo científico por parte del autor.
Publicación inicial	Fecha de primera publicación de artículo científico por parte del autor.
Conteo de Documentos Q1	Documentos científicos publicados en una revista de impacto Q1.
Conteo de Documentos Q2	Documentos científicos publicados en una revista de impacto Q2.
Conteo de Documentos Q3	Documentos científicos publicados en una revista de impacto Q3.
Conteo de Documentos Q4	Documentos científicos publicados en una revista de impacto Q4.
Países y agrupación de países	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Países y agrupación de países”.
Cultura	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Cultura”.
Educación	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Educación”.

Información y comunicación	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Información y comunicación”.
Política, derecho y economía	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Política, derecho y economía”.
Ciencia	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Ciencia”.
Ciencias Sociales y humanas	Producción científica del investigador relacionado con el campo amplio de la UNESCO de “Ciencias Sociales y humanas”.

5.3. Transformación

La tarea tomó como entrada el dataset formado en la fase de selección y adición de información. En esta etapa se realizaron las mismas tareas que en la fase de transformación de la técnica tradicional; tanto la tarea de discretización de datos sobre la columna de “Id Grupo de Investigación” cuanto la reducción de dimensionalidad son iguales. En la Tabla 11 se presenta los resultados que son similares a la Tabla 5 de la técnica tradicional, pero desde el atributo i61 al i67 se aumentan las 7 áreas generales de conocimiento propuestas en el Tesauro UNESCO.

Tabla 11. Campos del dataset de la fase de Transformación.

Atributo	Descripción	Tipo de valor
i1	Conteo de citaciones	Entero
i2	Conteo de citaciones al autor	Entero
i3	Conteo de coautores	Entero
i4	Conteo de proyectos de investigación	Entero
i5	Conteo de artículos sin cuartil en SJR	Entero
i6	Conteo de artículos Q1 en SJR	Entero
i7	Conteo de artículos Q2 en SJR	Entero
i8	Conteo de artículos Q3 en SJR	Entero
i9	Conteo de artículos Q4 en SJR	Entero
i10	Rol de coordinador de grupo de investigación	Binario (0 o 1)
i11	Rol de estudiante investigador	Binario (0 o 1)
i12	Rol de investigador	Binario (0 o 1)

i13	Rol de investigador externo	Binario (0 o 1)
i14	Rol de responsable de comunicación	Binario (0 o 1)
i15	Rol de técnico docente	Binario (0 o 1)
i16-I60	Grupos de investigación de la Universidad Politécnica Salesiana a los que pertenece el investigador	Binario (0 o 1)
i61	Países y agrupación de países	Decimal
i62	Cultura	Decimal
i63	Educación	Decimal
i64	Información y comunicación	Decimal
i65	Política, derecho y economía	Decimal
i66	Ciencia	Decimal
i67	Ciencias sociales y humanas	Decimal

La tarea de reducción de dimensiones, al igual que en la técnica tradicional, tiene el flujo de procesos presentado en la Figura 5, obteniendo una reducción de 67 dimensiones a 39 dimensiones, como se puede observar en la Figura 27.

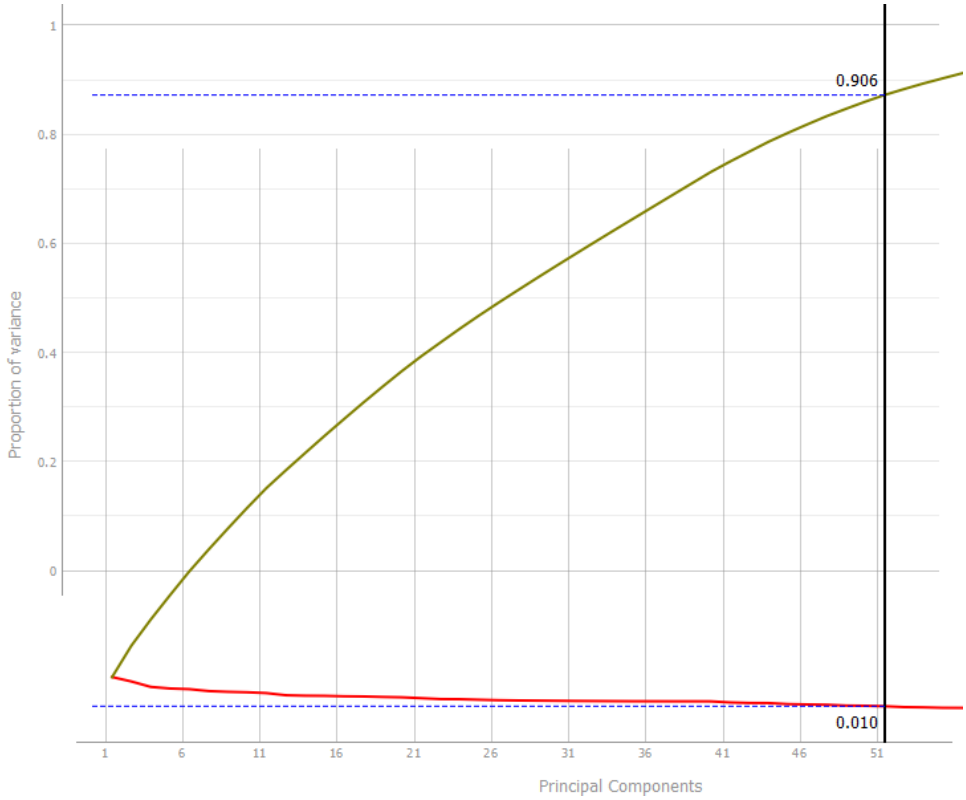


Figura 27. Gráfica de proporción de varianza de los componentes principales.

5.4. Minería de datos

En la etapa de minería de datos se tomó como entrada el dataset de 41 dimensiones y 221 registros de la etapa de transformación, escogiendo dos tareas para la descripción de la información del dataset, esto se hizo a través de los algoritmos no supervisados de: K-Means y Clustering de Louvain, en el software de Orange 3.

5.4.1. K-Means

En la tarea de K-Means se definieron las mismas condiciones que en la tarea de K-Means de técnica tradicional, utilizando 500 iteraciones y la medida de la silueta para detectar el K-Optimo de clústeres a formarse, con el flujo de conexiones de tareas presentado en la Figura 7. Los resultados de la tarea de K-Means dieron un K-Optimo de 28 grupos con una medida de silueta de 0.282, los cuales se pueden observar en el diagrama de distribución de la Figura 28.

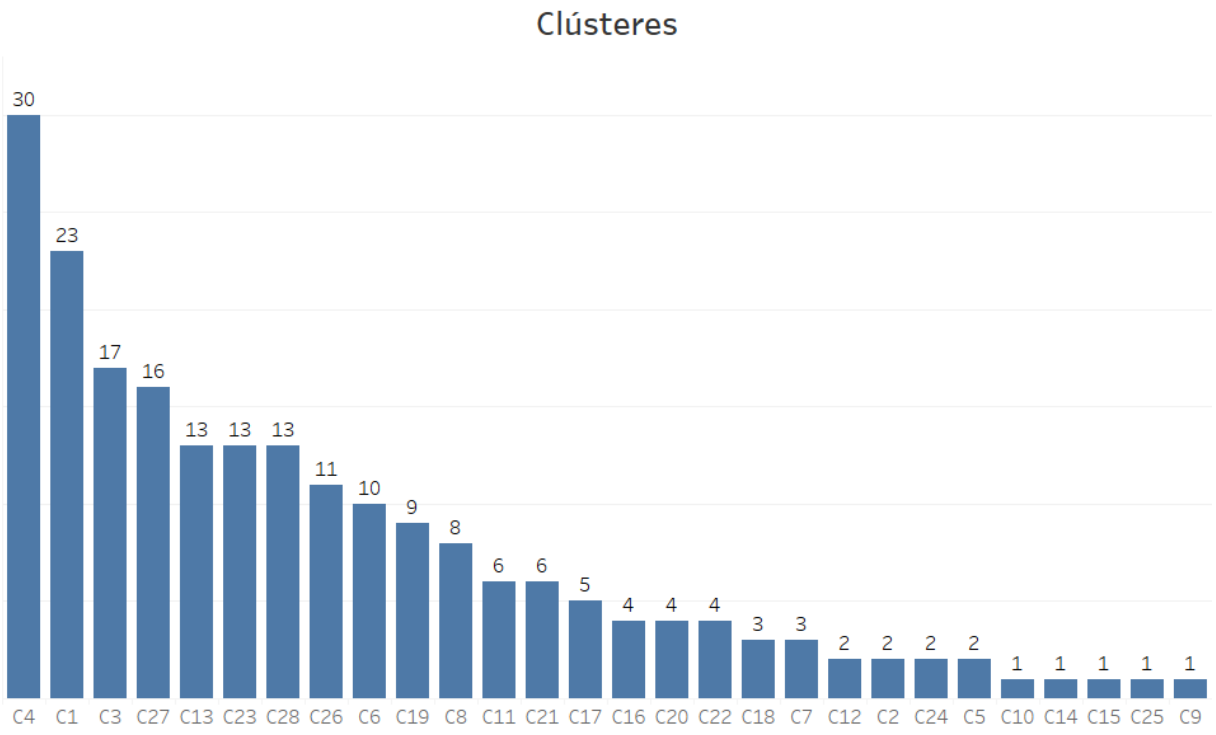


Figura 28. Diagrama de distribución de los clústeres formados con K-Means.

5.4.2. Clustering de Louvain

La tarea del clúster de Louvain se realizó bajo las mismas condiciones de la técnica tradicional. Se utilizó la medida Euclídea como medida de distancia con una resolución 1. En la Figura 9 se presenta el flujo de conexión de la tarea mediante la herramienta Orange 3.

Al realizar la tarea del Clustering de Louvain sobre el dataset se formaron 6 clústeres que se presentan en la Figura 29 de la gráfica de dispersión de los investigadores en los clústeres.

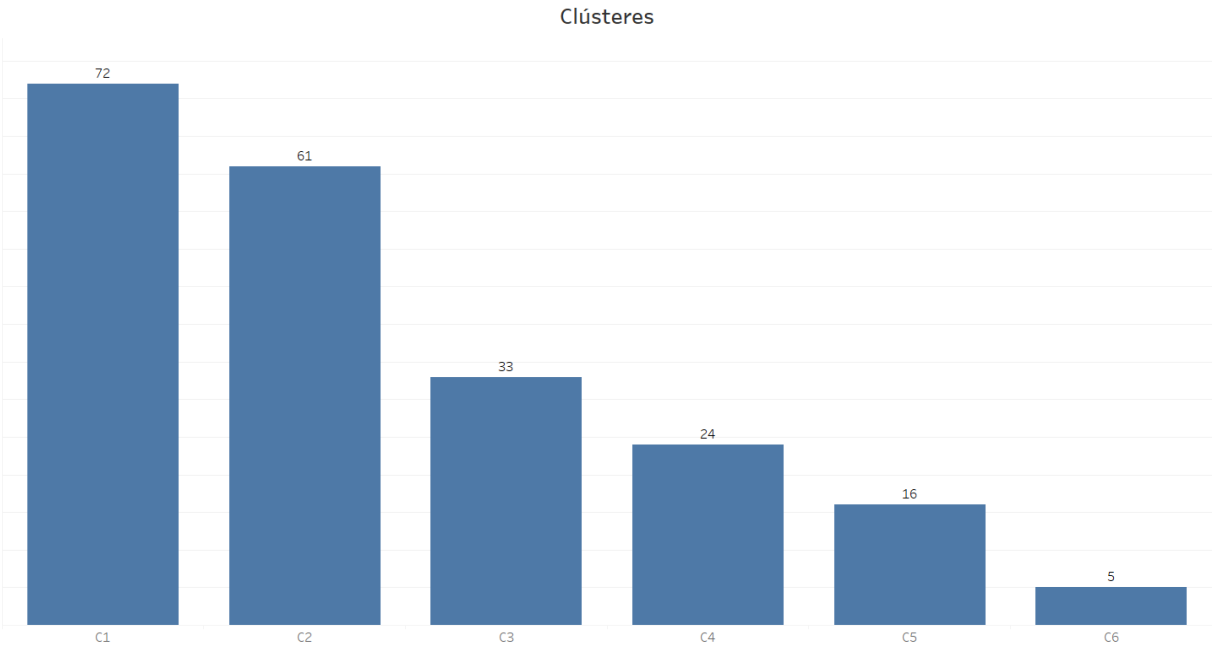


Figura 29. Diagrama de distribución de los clústeres formados con Clustering Louvain

5.5. Evaluación e interpretación

La evaluación de los clústeres e interpretación se realizó mediante la aplicación de estadística descriptiva sobre las 67 dimensiones consideradas y detalladas en la Tabla 11, las cuales después de los dos procesos de minería de datos, se caracterizaron en distintos clústeres. A continuación se presentó la evaluación e interpretación para los dos métodos de minería de datos tanto K-Means y Louvain, aplicando la medida de posición de la mediana, debido a la gran dispersión de los datos en los clústeres y cuartiles para mostrar la dispersión de los datos y tendencias de centralidad al igual que en la técnica 1.

5.5.1. Evaluación e interpretación clustering con K-Means

La evaluación de los clústeres que se formaron utilizando K-Means fueron evaluados con la medida de centralidad de la mediana al igual que en la técnica 1. Está medida de centralidad caracterizó alguno de los comportamientos más importantes de los clústeres.

El clúster 8 fue el de mayor rendimiento en función a las dimensiones de citación, conteo de artículos de cuartil Q1 y Q2 de SJR, además de una alta tasa de conteo de proyectos de investigación y coautores por investigador. Al análisis se añade que tiene una alta tasa de

producción de artículos categorizados bajo el área de Países y agrupaciones de países con un 21%.

El clúster C1 se caracteriza por su buen rendimiento en las dimensiones de citas, citas al autor, número de coautores, producción de artículos en cuartiles Q2 y sin cuartil en SJR. El clúster C6 tiene el más alto rendimiento en la dimensión de conteo de artículos de cuartil Q2 en SJR y un buen rendimiento en las variables de conteo de coautores, además de que sus autores están categorizados con 29% en el área de Política, derecho y economía. El clúster 17 tiene el más alto rendimiento en conteo de artículos de cuartil Q3 y sus investigadores tienen una producción de 100% categorizada en Información y comunicación.

Existen cuatro casos especiales de clústeres que están conformados solamente por un investigador y que tienen un alto rendimiento en las dimensiones. El clúster 9 que tiene el más alto rendimiento en el conteo de artículos de cuartil Q3 y un buen rendimiento de las variables de conteo de artículos Q1 y proyectos de investigación, además de tener el más alto porcentaje de artículos relacionados con el área de Cultura con el 23.5% y con el área de la Ciencia con un 17.5%. El clúster 10 tiene el más alto rendimiento en relación con producción de artículos Q3 y un alto índice de artículos categorizados en las áreas de Cultura y Educación con 20%. El clúster 15 que es el de mayor rendimiento, tiene los valores más altos en las variables de conteo de coautores, proyectos de investigación, producción de artículos científicos Q2 y Q3, además de tener una categorización de su producción científica de un 19% con respecto al área de Educación. Por último, el clúster 25 tiene el más alto rendimiento del conteo de artículos científicos Q3 y una categorización de la producción científica de sus artículos en las áreas de Educación, Ciencia e Información y comunicación con el 20% cada una.

Otros clústeres se caracterizan no por tener altos índices de producción de artículos en cuartiles o varias citaciones, sino por tener alta afinidad con algunas de las áreas UNESCO. El clúster 5 tienen el más alto porcentaje de producción en el área de Educación con 20% y un 33% en Información y comunicación, el clúster 21 tiene los más altos índices de producción en las áreas de País y agrupación de países con 50% y Política, derecho y economía con 50%.

El análisis completo de los diferentes clústeres formados a partir de la técnica de K-Means se presenta a continuación, utilizando como referencia la medida de centralidad de la mediana.

Clúster 1:

- Tiene 23 observaciones.

- Tienen una buena mediana respecto del resto de clústeres en el conteo de citaciones con 21, conteo de citaciones al autor con 20, conteo de coautores con 23, conteo de artículos sin cuartil en SJR con 7 y conteo de artículos Q2 en SJR con 7.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 2.

Clúster 2:

- Tiene 9 observaciones.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 25.
- Tienen la más alta mediana del porcentaje de artículos afines a Ciencias sociales y humanas con 65.4% respecto al resto de clústeres.

Clúster 3:

- Tiene 17 observaciones.
- Todos los integrantes son técnicos docentes.

Clúster 4:

- Tiene 30 observaciones.
- Tienen buena mediana del porcentaje de artículos afines a Información y comunicación con 19.5% respecto al resto de clústeres.
- Todos los integrantes son estudiantes investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 43.

Clúster 5:

- Tiene 2 observaciones.
- Tienen la más alta mediana del porcentaje de artículos afines a Educación con 20.85% respecto al resto de clústeres.
- Tienen buena mediana del porcentaje de artículos afines a Información y comunicación con 33.35% respecto al resto de clústeres.
- La mitad del clúster son coordinadores de grupos, investigadores y técnicos docentes.
- La mitad de los integrantes son miembros del Grupo de Investigación 22.
- Todos los integrantes son miembros del Grupo de Investigación 38.

Clúster 6:

- Tiene 10 observaciones.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de coautores con 18.5 y porcentaje de artículos afines a Política, derecho y economía con 29.7%
- Tienen la mayor mediana de conteo de artículos Q2 en SJR con 6 respecto al resto de clústeres.
- Todos los integrantes son investigadores externos.
- Todos los integrantes son miembros del Grupo de Investigación 14.

Clúster 7:

- Tiene 3 observaciones.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de proyectos de investigación con 5, conteo de coautores con 22, conteo de artículos Q1 en SJR con 2 y porcentaje de artículos afines a Política, derecho y economía con 46.2%.
- Tienen la mayor mediana respecto al resto de clústeres del conteo de artículos Q3 en SJR con 1 respecto al resto de clústeres.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 23 y 45.

Clúster 8:

- Tiene 8 observaciones.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de citaciones con 129, conteo de citaciones al autor con 106.5, conteo de artículos Q1 en SJR con 3.5 y conteo de artículos Q2 en SJR con 5.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de coautores con 19, conteo de artículos sin cuartil SJR con 5, porcentaje de artículos afines a Países y agrupación de países con 28.40% y porcentaje de artículos afines a Ciencias sociales y humanas con 26.95%.
- La mitad del clúster son técnico docente.
- Todos los integrantes son miembros del Grupo de Investigación 44.

Clúster 9:

- Tiene 1 observaciones.

- Tienen una buena mediana respecto al resto de clústeres en el conteo de proyectos de investigación con 5, conteo de artículos Q1 en SJR con 1 y en el porcentaje de artículos afines a Ciencia con 17.6%.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de artículos Q3 en SJR con 1 y en el porcentaje de artículos afines a Cultura con 23.50%.
- Todos los integrantes son investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 26.

Clúster 10:

- Tiene 1 observación.
- Tienen la mayor mediana de conteo de artículos Q3 en SJR con 1 respecto al resto de clústeres.
- Tienen buena mediana respecto al resto de clústeres en el porcentaje de artículos afines a Cultura con 20% y en el porcentaje de artículos afines a Educación con 20%.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros de los Grupos de Investigación 28 y 34.

Clúster 11:

- Tiene 6 observaciones.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 17.

Clúster 12:

- Tiene 2 observaciones.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de artículos Q3 en SJR con 1 y en el porcentaje de artículos afines a Países y agrupación de países con 33.35%.
- La mitad del clúster son investigadores y técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 40.

Clúster 13:

- Tiene 13 observaciones.
- Todos los integrantes son técnicos docentes.

Clúster 14:

- Tiene 1 observación.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de proyectos de investigación con 3, porcentaje de artículos afines a Política, derecho y economía con 42.9% y en el porcentaje de artículos afines a Ciencias sociales y humanas con 28.6%.
- Todos los integrantes son técnico docente.
- Todos los integrantes son miembros del Grupo de Investigación 1.

Clúster 15:

- Tiene 1 observación.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de citaciones con 67, conteo de citación al autor de 58, conteo de artículos sin cuartil SJR con 44 y en el del porcentaje de artículos afines a Educación con 19.5%.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de coautores con 163, conteo de proyectos de investigación con 20, conteo de Q2 en SJR con 44 y en el conteo de Q3 en SJR con 2.
- Todos los integrantes son investigadores y técnicos docentes.
- Todos los integrantes son miembros de los Grupos de Investigación 32 y 43.

Clúster 16:

- Tiene 4 observaciones.
- Tienen buena mediana respecto al resto de clústeres del porcentaje de artículos afines a Países y agrupación de países con 38.65% y en el porcentaje de artículos afines a Educación con 19.5%.
- La mitad del clúster está integrada por investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 22.

Clúster 17:

- Tiene 5 observaciones.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de artículos Q3 en SJR con 1 y en el porcentaje de artículos afines a Información y comunicación con 100%.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 4

Clúster 18:

- Tiene 3 observaciones.
- Tienen la mayor mediana del porcentaje de artículos afines a Países y agrupación de países con 50% respecto al resto de clústeres.
- Todos los integrantes son técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 31.

Clúster 19:

- Tiene 9 observaciones.
- Todos los integrantes son miembros del Grupo de Investigación 18.

Clúster 20:

- Tiene 4 observaciones.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de citación con 14.5, conteo de citación al autor de 12.5, conteo de coautores con 16.5 y en el porcentaje de artículos afines a Política, derecho y economía con 35%.
- La mitad del clúster son estudiantes investigadores y técnicos docentes.
- Todos los integrantes son miembros del Grupo de Investigación 5.

Clúster 21:

- Tiene 6 observaciones.
- Tienen la mayor mediana respecto al resto de clústeres en el porcentaje de artículos afines a Países y agrupación de países con 50% y en el porcentaje de artículos afines a Política, derecho y economía con 50%.
- Todos los integrantes son técnicos docentes.

Clúster 22:

- Tiene 4 observaciones.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de coautores con 20, porcentaje de artículos afines a Educación con 19.3%, del porcentaje de artículos afines a Ciencias sociales y humanas con 27.4% y en el porcentaje de artículos afines a Ciencia con 14.6%.
- Todos los integrantes son técnico docente.
- Todos los integrantes son miembros de los Grupos de Investigación 11 y 37.

Clúster 23:

- Tiene 13 observaciones.
- Todos los integrantes son técnico docente.
- Todos los integrantes son miembros del Grupo de Investigación 13.

Clúster 24:

- Tiene 2 observaciones.
- Tienen buena mediana del porcentaje de artículos afines a Educación con 17.15% respecto al resto de clústeres.
- La mitad del clúster son investigadores.
- La mitad de los integrantes son miembros de los Grupos de Investigación 13 y 32.
- Todos los integrantes son miembros del Grupo de Investigación 36.

Clúster 25:

- Tiene 1 observación.
- Tienen la mayor mediana respecto al resto de clústeres en el conteo de artículos Q3 en SJR con 1 y en el porcentaje de artículos afines a Ciencia con 20%.
- Tienen buena mediana respecto al resto de clústeres en el porcentaje de artículos afines a Educación con 20% y en el del porcentaje de artículos afines a Información y comunicación con 20%.
- Todos los integrantes son investigadores y técnicos docentes.
- Todos los integrantes son miembros de los Grupos de Investigación 6, 16, 39 y 41.

Clúster 26:

- Tiene 11 observaciones.
- Tienen buena mediana del porcentaje de artículos afines a Ciencias sociales y humanas con 37.5% respecto al resto de clústeres.
- Todos los integrantes son estudiantes investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 37.

Clúster 27:

- Tiene 16 observaciones.
- Tienen buena mediana del porcentaje de artículos afines a Ciencias sociales y humanas con 43.75% respecto al resto de clústeres.
- Todos los integrantes son estudiantes investigadores.

- Todos los integrantes son miembros del Grupo de Investigación 2.

Clúster 28:

- Tiene 13 observaciones.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de coautores con 21, conteo de artículos sin cuartil SJR con 4 y en el porcentaje de artículos afines a Educación con 17.30%.
- Todos los integrantes son miembros del Grupo de Investigación 43.

En la Figura 30 se puede ver cómo el conteo de citaciones del clúster 8 tiene una gran dispersión, pero su medida de centralidad usando la media es cercana a 200, mientras que la mediana es 129, siendo superior al resto de clústeres en esta dimensión. También el clúster 1 y 6 tienen un mayor despunte frente al resto de clústeres, pero entre más despunta en esta dimensión, más crece la dispersión.

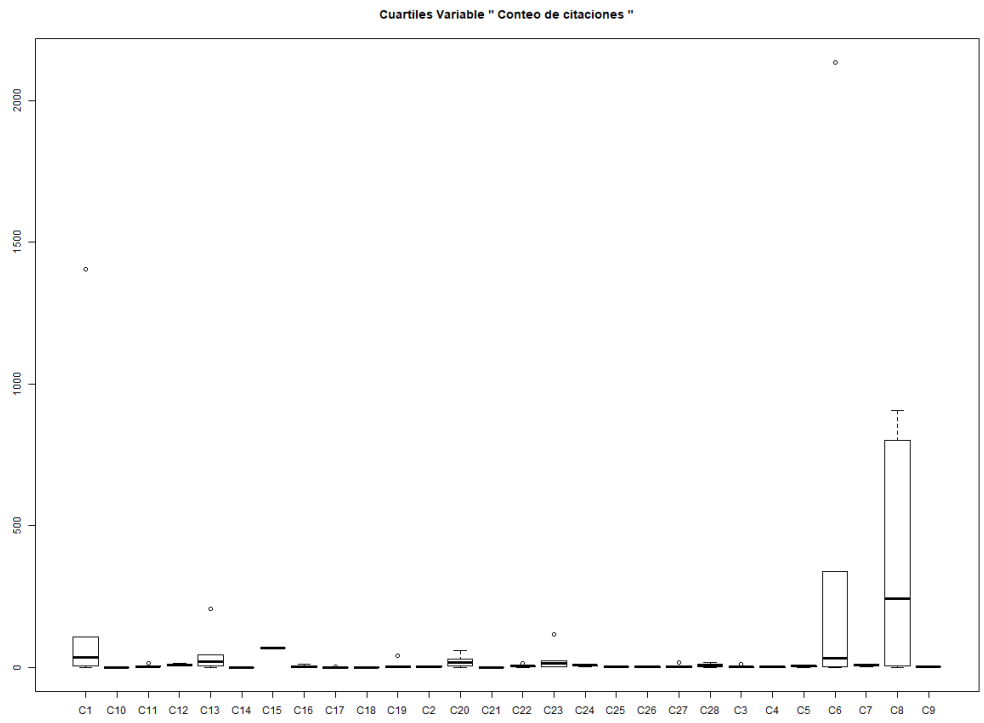


Figura 30. Cuartiles de la variable de conteo de citaciones aplicando K-Means en la técnica 2.

El clúster 17 está caracterizado por la producción de artículos científicos afines al área de Información y comunicación con un 100%, en la Figura 31 se presenta que, a pesar de tener un poco de dispersión, sigue teniendo una media y una mediana del 100% que dista mucho respecto al resto de clústeres. Otros clústeres afines a esta área son el 4, 5, 16 y 19, los cuales tienen una alta dispersión, pero una medida de tendencia central mayor que el resto de los clústeres.

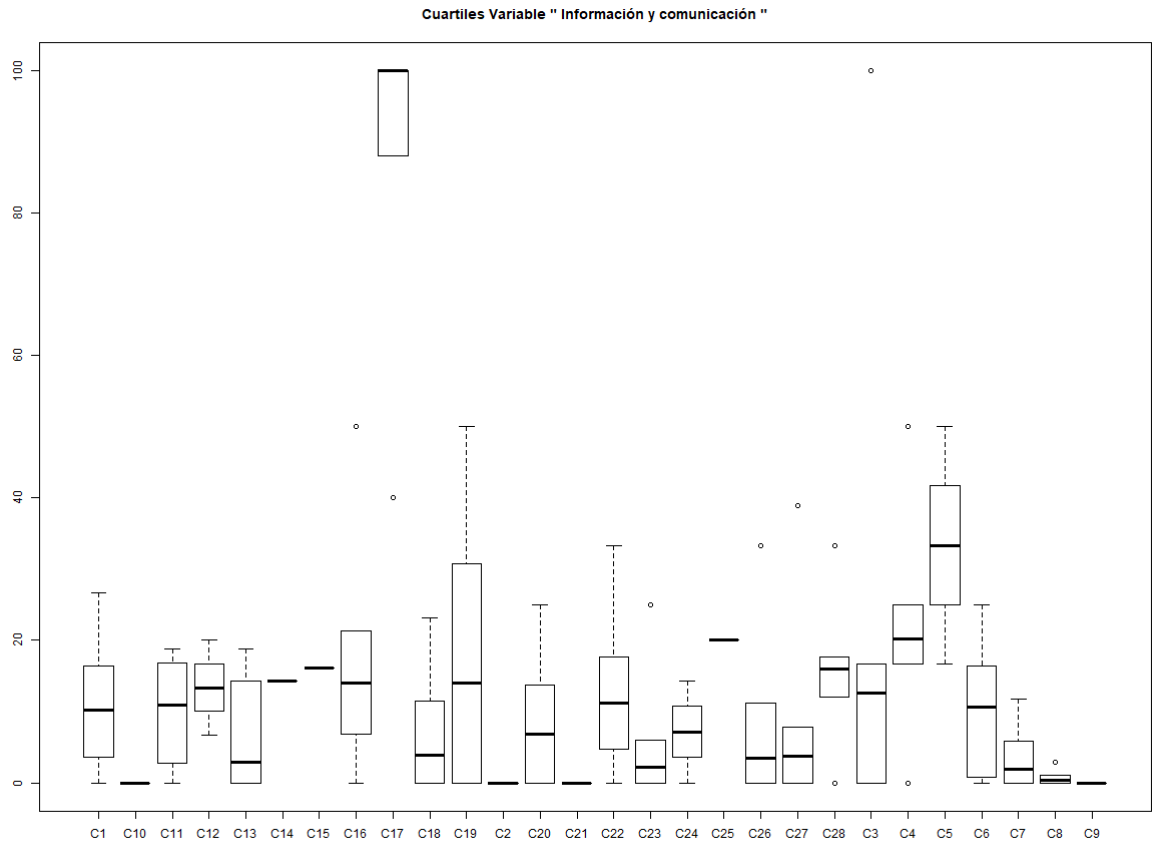


Figura 31. Cuartiles de la variable de Información y comunicación aplicando K-Means en la técnica 2.

5.5.2. Evaluación e interpretación clustering con Louvain

Al igual que con K-Means, se tomó como medida de centralidad la mediana por la dispersión de los datos y un posterior análisis aplicando cuartiles para ver la dispersión de los datos. El clúster de mayor rendimiento es el clúster 3, teniendo los mayores valores para las variables de conteo de citas, citas al autor, artículos Q2, Q3 y sin cuartil en SJR, además de tener una buena tasa de proyectos de investigación por investigador. También se caracteriza por tener afinidad con las áreas de Política, derecho y economía con 28.6%, Países y agrupaciones de países con 23.6% y Ciencias sociales y humanas con un 20%.

El clúster 1 es considerado como de buen rendimiento, debido a que tiene la mayor tasa de conteo de proyectos por investigador y buen rendimiento en las variables de conteo de citas, citas al autor, artículos en cuartiles Q2 y sin cuartiles, además de tener una afinidad a las áreas de Política, derecho y economía con 23.6%, Ciencias humanas y sociales con 17.7% y Países y agrupaciones de países con 16.7%.

El clúster 2 también es considerado de buen rendimiento con un alto valor de conteo de citas, citas al autor, coautores y artículos Q2 y sin cuartil, pero con una mediana de proyectos de

investigación de 0. El clúster 2 tiene afinidad con las áreas de Países y agrupación de países, Educación, ciencias sociales y humanas, Información y comunicación con un 16.7% y Política, derecho y economía con un 19%.

El clúster 4 tiene el más alto rendimiento en el conteo de proyectos de investigación, pero baja citación. Su producción está categorizada con la mayor tasa con respecto al resto de clústeres en las áreas de Países y agrupación de países con 45%, Política, derecho y economía con 33% y Ciencias humanas con 32.3%.

El clúster 5, a pesar de no tener altas puntuaciones en las medias de citaciones y artículos científicos, tiene afinidad con las áreas de la Educación con un 11% y Ciencia con 12%.

El clúster 6, a pesar de no tener altas puntuaciones en las medias de citaciones y artículos científicos, tiene el más alto rendimiento en el conteo de proyectos de investigación y un 100% de afinidad con el área de Información y comunicación.

El análisis completo de los diferentes clústeres formados a partir de la técnica de Louvain se presenta a continuación, utilizando como referencia la medida de centralidad de la mediana.

Clúster 1:

- Tiene 72 observaciones.
- Tienen la más alta mediana de conteo de proyectos de investigación, con 2 respecto al resto de clústeres.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de citaciones con 2, conteo de citaciones al autor con 2, conteo de coautores con 6.5, conteo de artículos sin cuartil SJR con 2, conteo de artículos Q2 en SJR con 2, porcentaje de artículos afines a Países y agrupación de países con 16.7%, porcentaje de artículos afines a Información y comunicación con 7.5%, porcentaje de artículos afines a Política, derecho y economía con 23.65% y en el porcentaje de artículos afines a Ciencias sociales y humanas con 17.7%.
- Tienen una baja mediana respecto al resto de clústeres en el porcentaje de artículos afines a Educación con 7.7% y en el porcentaje de artículos afines a Ciencia con 4.05%.
- Todos los integrantes son técnicos investigadores.

Clúster 2:

- Tiene 61 observaciones.
- Tienen la más alta mediana del porcentaje de artículos afines a Educación con 16.7% respecto al resto de clústeres.

- Tienen una buena mediana respecto al resto de clústeres en el conteo de citaciones con 2, conteo de citaciones al autor con 2, conteo de coautores con 7, conteo de artículos sin cuartil SJR con 2, conteo de artículos Q2 en SJR con 2, porcentaje de artículos afines a Países y agrupación de países con 16.7%, porcentaje de artículos afines a Información y comunicación con 16.7%, porcentaje de artículos afines a Política, derecho y economía con 19%, porcentaje de artículos afines a Ciencia con 4.3% y en el porcentaje de artículos afines a Ciencias sociales y humanas con 16.7%.
- Tienen una mediana de conteo de proyectos de investigación de 0.
- Todos los integrantes son estudiantes investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 43.

Clúster 3:

- Tiene 33 observaciones.
- Tienen la más alta mediana respecto al resto de clústeres en el conteo de citaciones con 10, conteo de citaciones al autor con 8, conteo de coautores con 11, conteo de artículos sin cuartil SJR con 4, conteo de artículos Q2 en SJR con 4 y en el conteo de artículos Q3 en SJR con 1.
- Tienen una buena mediana respecto al resto de clústeres en el conteo de proyectos de investigación con 1, porcentaje de artículos afines a Países y agrupación de países con 23.8%, porcentaje de artículos afines a Educación con 10.9%, porcentaje de artículos afines a Información y comunicación con 4.1%, porcentaje de artículos afines a Política, derecho y economía con 28.6% y en el porcentaje de artículos afines a Ciencias sociales y humanas con 20%.
- Todos los integrantes son técnicos docentes.

Clúster 4:

- Tiene 24 observaciones.
- Tienen la más alta mediana respecto al resto de clústeres en el conteo de proyectos de investigación con 2, porcentaje de artículos afines a Política, derecho y economía con 33.3%, porcentaje de artículos afines a Países y agrupación de países con 45% y en el porcentaje de artículos afines a Ciencias sociales y humanas con 32.3%.
- Tienen una normal mediana de conteo de coautores con 5 respecto al resto de clústeres.
- Tienen una baja mediana respecto al resto de clústeres en el conteo de citaciones con 1, conteo de citaciones al autor con 1 y en el conteo de artículos sin cuartil SJR con 1.
- Todos los integrantes son técnicos investigadores.

- Todos los integrantes son miembros del Grupo de Investigación 13.

Clúster 5:

- Tiene 16 observaciones.
- Tienen la más alta mediana del porcentaje de artículos afines a Ciencia con 12.5% respecto al resto de clústeres.
- Tienen buena mediana del porcentaje de artículos afines a Educación con 11.25% respecto al resto de clústeres.
- Tienen una normal mediana de conteo de coautores con 5 respecto al resto de clústeres.
- Tienen una baja mediana respecto al resto de clústeres en el conteo de citaciones con 1, conteo de citaciones al autor con 1, conteo de artículos sin cuartil SJR con 1 y en el porcentaje de artículos afines a Información y comunicación con 3.15%.
- Todos los integrantes son estudiantes investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 37.

Clúster 6:

- Tiene 5 observaciones.
- Tienen la más alta mediana respecto al resto de clústeres en el porcentaje de artículos afines a Información y comunicación con 100% y en el conteo de proyectos de investigación con 2.
- Tienen una normal mediana de conteo de coautores con 5 respecto al resto de clústeres.
- Tienen una mediana de conteo de citaciones y citaciones al autor de 0.
- Tienen mediana de conteo de artículos sin cuartil SJR de 0.
- Todos los integrantes son técnicos investigadores.
- Todos los integrantes son miembros del Grupo de Investigación 4.

El clúster 3 tiene la mayor mediana de citaciones respecto al resto de clústeres, en la Figura 32 se muestra que el clúster tiene un alto grado de dispersión comparado con el resto de los clústeres en los cuales casi no existe dispersión, pero su media es de 0. Esto expresa que el clúster 3 está compuesto por las personas con mayor número de citas y el resto de los clústeres carecen de investigadores con grandes índices de citaciones.

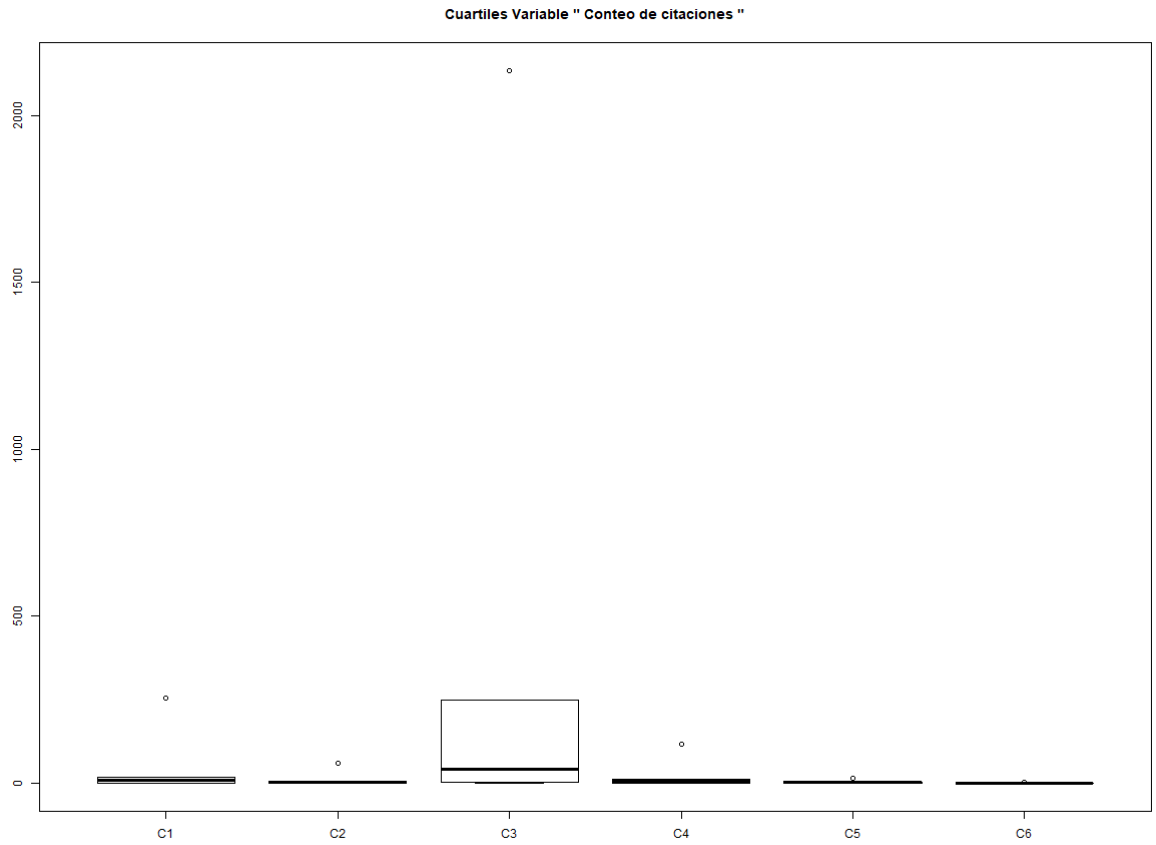


Figura 32. Cuartiles de la variable de conteo de citaciones aplicando Louvain en la técnica 2.

En la Figura 33 podemos ver cómo el diagrama de bigotes muestra que el clúster 5 está conformado por investigadores del 20% al 100% de afinidad con el área de Ciencias sociales y humanas, teniendo una dispersión de los datos muy grande. Esta dispersión en el clúster 5 hace que a pesar de que la media sea 40%, la mediana sea de 12.5 % afectando de gran manera esta característica del clúster.

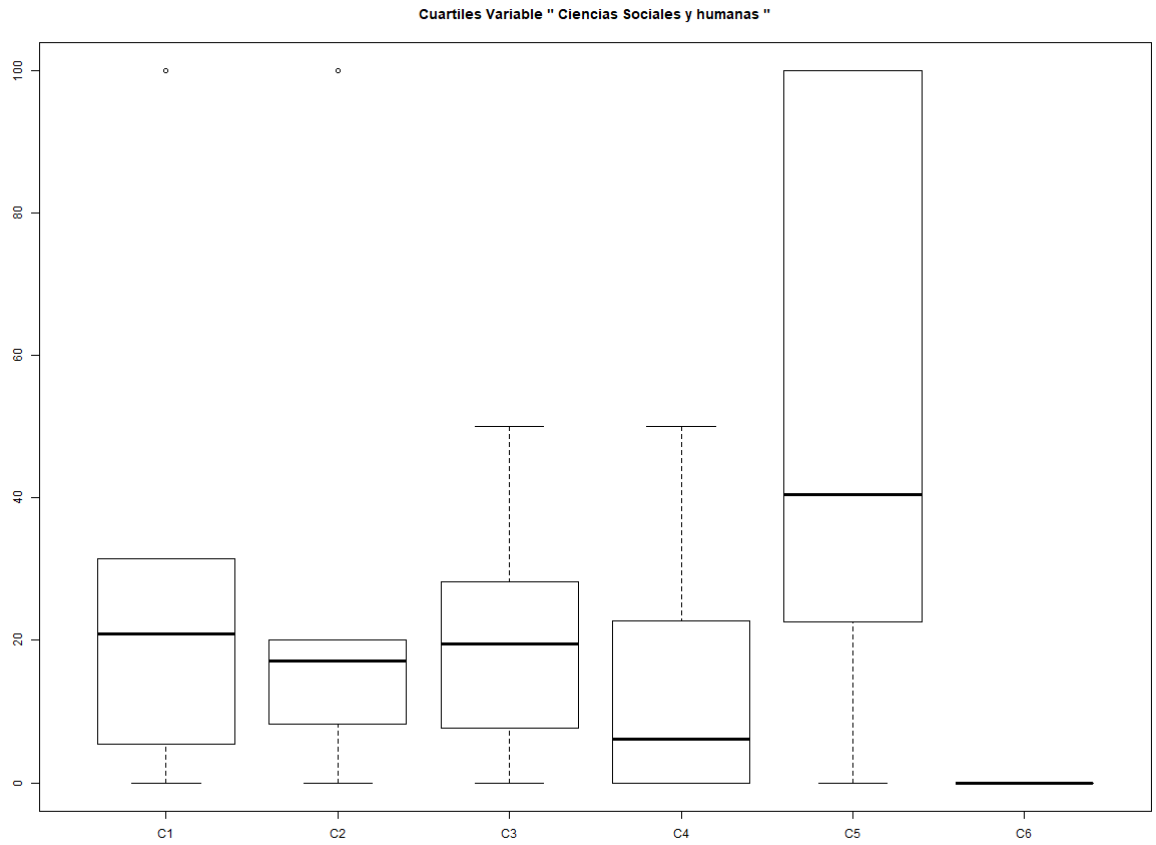


Figura 33. Cuartiles de la variable de Ciencias sociales y humanas aplicando Louvain en la técnica 2.

6. Comparativa de técnicas de descubrimiento de conocimiento de perfiles

Para la comparativa de las técnicas aplicadas, en el presente trabajo se tomó en cuenta las características internas de formación de los grupos y las características semánticas que describen a los grupos. Esta comparativa contrasta estas características en función del valor agregado que aporta esta información a la institución.

En la ejecución de las técnicas de descubrimiento de conocimiento se tomó en cuenta algunas métricas internas del proceso de minería de datos, permitiendo contraponer los resultados en función de estas métricas como se presenta en la Tabla 12. Tanto en la técnica 1 como en la técnica 2 se aplicaron los clústeres de K-Means y Louvain, tomando como medidas el número de clústeres formados, así como la medida de silueta que describe si los objetos de un clúster son similares en relación con el resto de los clústeres, la calidad del clúster aumenta entre mayor sea el valor.

Al aplicar K-Means se obtuvieron 28 clústeres en la técnica 2, mientras que en la técnica 1 se obtuvieron 29 clústeres, pero los clústeres en la técnica 1 fueron de mayor calidad debido a que su puntaje de silueta es de 0.363, mientras que el puntaje de silueta de la técnica 2 fue de 0.282. El puntaje de las siluetas de las dos técnicas indica que los clústeres formados en la técnica 1 presentan más similitud entre sí en relación con el resto de los clústeres que los clústeres de K-Means de la técnica 2.

Utilizando el clustering de Louvain se presenta algo similar que al utilizar K-Means, debido a que la técnica 1 obtuvo 7 clústeres y la técnica 2 obtuvo 6 clústeres, pero presentan una diferencia en la calidad, siendo favorable para la técnica 1 al tener una silueta de 0.113 siendo mayor a la técnica 2 con una silueta de 0.105. Se debe aclarar que el clustering de Louvain al ser un algoritmo codicioso, siempre buscara optimizar los grupos para encontrar una solución más generalizada, lo que se traduce en una mayor distancia entre los objetos de un clúster.

Tabla 12. Comparación de métricas del proceso de minería de datos.

Métrica	Técnica 1	Técnica 2
Número de dimensiones del dataset	60	67
Número de registros	221	221
Número de componentes de ACP	39	41
Clústeres con K-Means	29	28
Promedio de silueta con K-Means	0.363	0.282

Clústeres con Louvain	7	6
Promedio de silueta con Louvain	0.113	0.105

Para encontrar el K-Óptimo en el algoritmo de K-Means se utilizó la medida de silueta; en la Figura 34 se presenta el gráfico lineal de la silueta contra el número de clústeres de la técnica 1 y técnica 2. En la gráfica se puede ver cómo la técnica 1 tiene un mayor grado en su silueta que la técnica 2 cuya gráfica permanece abajo, pero la técnica 2 converge un clúster antes que la técnica 1. La gráfica refuerza el hecho que el clúster con K-Means en la técnica 1 tienen mayor calidad de su conformación que en la técnica 2.

Figura 34. Gráfica lineal del K-Óptimo de K-Means en las técnicas de KDD.

En cuanto a los clústeres formados con Louvain se puede ver en las figuras 35 y 36, la medida de silueta de cada uno de los clústeres de la técnica 1 y la técnica 2 respectivamente. La primera técnica presenta métricas de silueta muy bajas para los clústeres C2, C3, C4 y C5, mientras que el clúster 1 y 6 presentan una medida superior a los clústeres bajos; por último, el clúster 7 presenta una buena medida de silueta de 0.84 que se puede justificar debido a que el número de observaciones de ese clúster es la más baja respecto al resto de clústeres con 5.

En la técnica 2 se presentan puntajes de silueta bajos para los clústeres C1, C2, C3 y C4, pero C5 tiene un mayor puntaje que los clústeres bajos e incluso que los clústeres C1 y C6 de la técnica 1. Para el caso del clúster 6 se observa que el puntaje de silueta es bueno con 0.82, casi igual que el puntaje del clúster 7 de la técnica 1, pero esto se debe a que el clúster C7 de la técnica 1 y el clúster C6 de la técnica 2 contienen exactamente a las mismas observaciones, reforzando la singularidad de los miembros de este grupo.

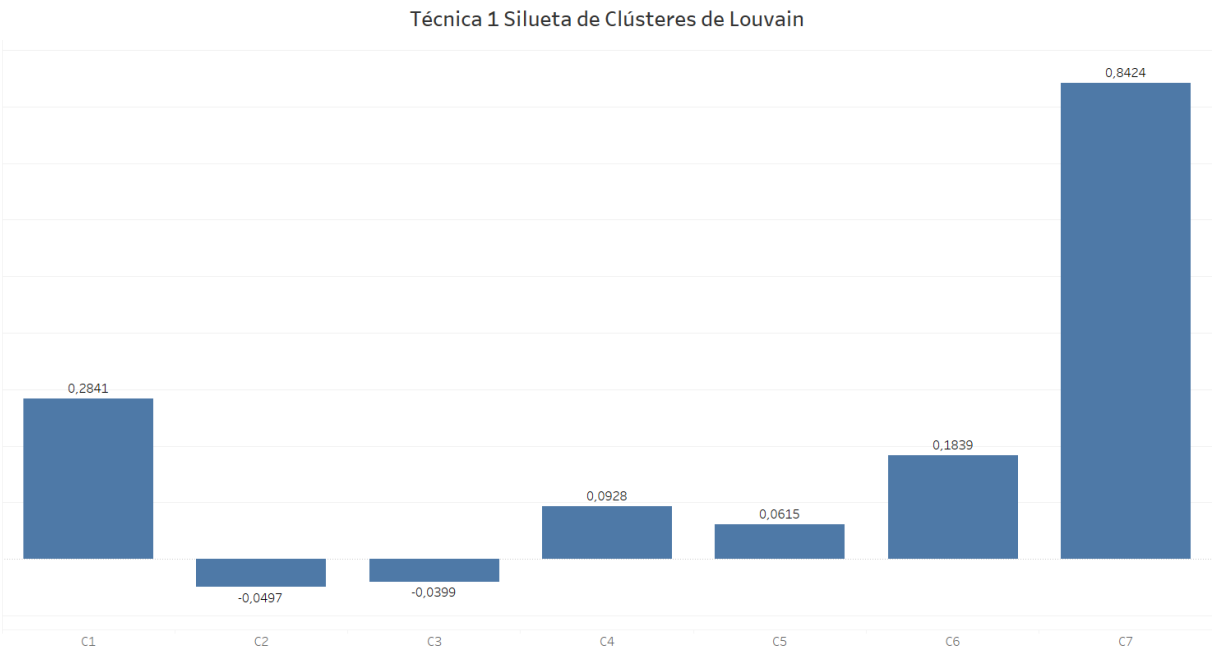


Figura 35. Gráfica de barras del puntaje de silueta de los clústeres utilizando Louvain en la técnica 1.

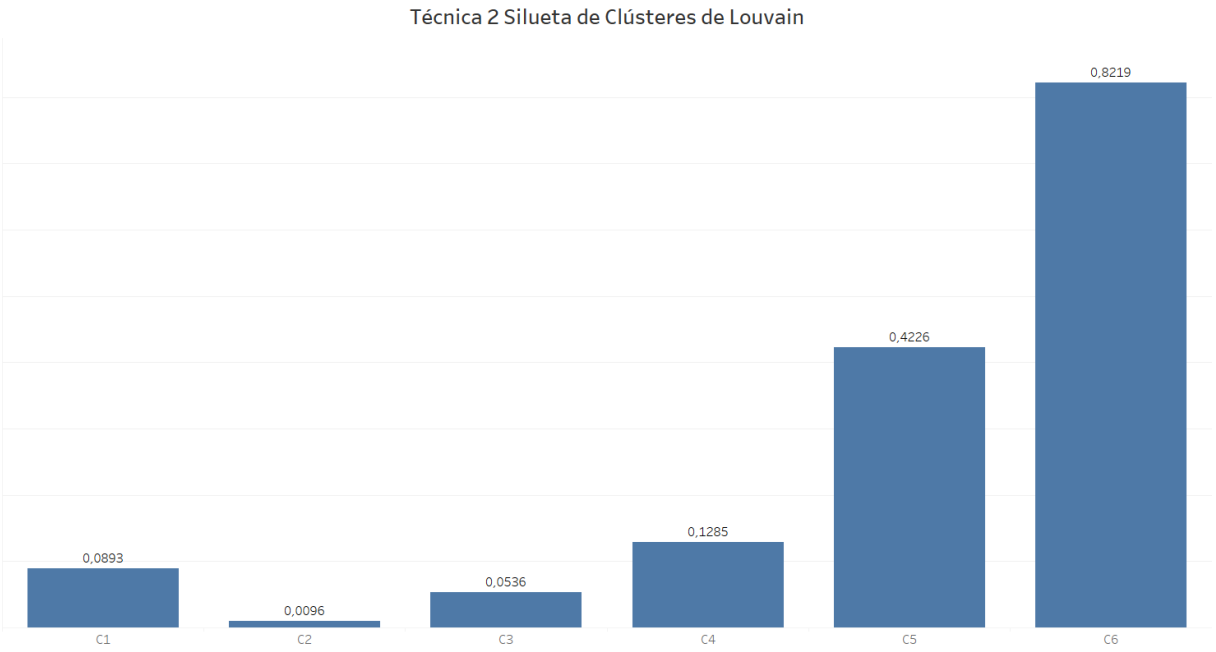


Figura 36. Gráfica de barras del puntaje de silueta de los clústeres utilizando Louvain en la técnica 2.

Después de contrastar las medidas internas de conformación de los clústeres, se presenta la comparativa de la información que se encontró a partir del proceso de descubrimiento de conocimiento en las tablas 13 y 14.

Al comparar la información encontrada en los clústeres utilizando K-Means, se categorizaron de acuerdo con el rendimiento de los integrantes de cada clúster, siendo de bajo rendimiento al solo tener descripciones de los roles, grupos pertenecientes y en el caso de la técnica 2

categorizaciones de áreas de conocimiento buenas. Los clústeres con investigadores de mediano rendimiento tienen medidas de producción científica buenas tales como conteo de coautores, citaciones, artículos por cuartil, proyectos de investigación y, en el caso de la técnica 2, categorización de áreas de conocimiento altas. La Tabla 13 contiene la descripción de los clústeres representados con la letra C, acompañada por el número del clúster y, entre paréntesis, el número de observaciones del clúster.

Se aprecia que la técnica 1 tiene una gran cantidad de clústeres de bajo rendimiento con 20 clústeres que categorizan a 160 investigadores, también presenta 6 grupos de mediano rendimiento, representando a 51 investigadores y, por último, contiene a 3 clústeres de alto rendimiento con 10 investigadores. Se tienen también 4 clústeres de bajo rendimiento y 2 de alto rendimiento con solamente una observación en sus clústeres.

La técnica 2, en cambio, presenta menos clústeres de bajo rendimiento con 12 clústeres que representan a 133 investigadores, también presenta más clústeres de mediano rendimiento con 7 clústeres que representan a 57 investigadores. Los clústeres de alto rendimiento también son mayores, con 9 clústeres que representan a 31 investigadores; pero, dentro de estos clústeres, 4 solamente tienen una observación.

El contraste con K-Means revela que la técnica 1 forma un embudo al tener menos investigadores de alto rendimiento en las dimensiones establecidas, siendo el clúster 8 el que contiene a los mejores investigadores, en cambio la técnica 2, al incluir las dimensiones de categorización de la producción científica, desarrolla un comportamiento más equilibrado al incluir más investigadores en los grupos de alto y mediano rendimiento. La técnica 2 también presenta más grupos con una sola observación en la categoría de alto rendimiento, justificando el hecho de que tienen una medida menor de silueta para los clústeres de esta técnica.

Tabla 13. Comparativa de la categorización de los clústeres usando K-Means.

Categoría de rendimiento	Técnica 1	Técnica 2
Clústeres de investigadores con bajo rendimiento	C1(4), C2(10), C6(28), C7(9), C8(4), C11(16), C12(13), C13(3), C15(4), C16(16), C18(8), C19(4), C20(2), C21(1), C22(1), C23(1), C26(16), C27(1), C28(13), C29(6)	C3(17), C4(30), C11(6), C13(13), C16(4), C19(9), C20(6), C22(6), C23(13), C24(2), C26(11), C27(16)
Clústeres de investigadores con mediano rendimiento	C4(3), C5(18), C9(4), C17(15), C24(9), C25(2)	C1(23), C2(9), C5(2), C14(1), C18(3), C21(6), C28(13)

Clústeres de investigadores con alto rendimiento	C3(1), C10(8), C14(1)	C6(10), C7(3), C8(8), C9(1), C10(1), C12(2), C15(1), C17(5), C25(1)

Al igual que en el contraste de las categorizaciones de rendimiento con K-Means, se utilizó los mismos parámetros de clasificación para los clústeres utilizando Louvain, presentados en la Tabla 14. En la técnica 1 se detectó 2 clústeres de mediano rendimiento con 94 investigadores y 5 clústeres de alto rendimiento con 117 investigadores. En la técnica 2, en cambio, se encontró 2 clústeres de mediano rendimiento con 77 investigadores y 4 clústeres de alto rendimiento con 134 investigadores.

En la comparativa se puede observar que la técnica 2 agrupa a más investigadores en clústeres de alto rendimiento que en la técnica 1, al igual que en la comparativa de K-Means. Se puede afirmar que los clústeres con Louvain tienden a ubicar más investigadores dentro de la categoría de alto rendimiento, puesto que en la técnica 1 al utilizar K-Means solamente se obtuvo 10 investigadores de alto rendimiento contra los 117 en Louvain. No se detectó clústeres con bajo rendimiento lo cual es otra muestra del comportamiento optimista de los clústeres con Louvain.

Tabla 14. Comparativa de la categorización de los clústeres usando Louvain.

Categoría de rendimiento	**Técnica 1**	**Técnica 2**
Clústeres de investigadores con mediano rendimiento	C1(62), C4(32)	C2(61), C5(16)
Clústeres de investigadores con alto rendimiento	C2(45), C3(36), C5(22), C6(9), C7(5)	C1(72), C3(33), C4(24), C6(5)

7. Conclusiones y trabajo futuro

El presente capítulo desarrolla las conclusiones más relevantes del contraste de las dos técnicas de descubrimiento de conocimiento y propone líneas base para futuras investigaciones a partir de la experiencia recopilada en el TFM.

7.1. Conclusiones

Las técnicas de descubrimiento de conocimiento aplicadas entregaron una visión de la composición de perfiles de comunidades científicas con las que cuenta la UPS, basadas en la producción científica indizada en Scopus, en donde dichos perfiles son representados como clústeres, producto de la fase de minería de datos. Los clústeres aplicados fueron K-Means y Louvain, los cuales presentaron una visión global y una más granular de los perfiles de investigadores, respectivamente.

La técnica 1, aplicando Louvain, presentó una visión global de la investigación de la UPS en Scopus, caracterizando a los perfiles de investigadores como los perfiles de C2 que son investigadores con múltiples proyectos de investigación y productos científicos de impacto Q2; el perfil C3 tiene el mayor número de coautores y, a su vez, el mayor número de citaciones; y los perfiles C5, que producen solamente artículos de impacto Q3. Por otro lado, aplicando K-Means, se presentó una visión más granular en la técnica 1 en donde se destaca a C10 como el perfil de alto rendimiento de la UPS con productos científicos de impacto Q1 y Q2 con un alto grado de citación, también se presentan perfiles de buen rendimiento que producen artículos de impacto y tienen un alto grado de citación correspondiente a 51 investigadores y, por último, presenta un conjunto de perfiles de investigadores con baja tasa de producción, que corresponde a 160 investigadores.

La técnica 2 utilizó también la caracterización de la investigación bajo áreas de conocimiento de la UNESCO, a través de modelado de conocimiento con estructuras semánticas y procesos de similitud, lo cual enriqueció la información de inicio del dataset, a diferencia de la técnica 1, provocando el descubrimiento de nuevos patrones implícitos en la información del dataset. Al utilizar el clustering de Louvain, en la técnica 2 se presentó una visión global de los perfiles de investigadores en función de su producción indizada en Scopus, como el caso del perfil C3 con alto grado de citación, coautores y producción científica de alto impacto, el perfil C4 con una alta tasa de proyectos y productos científicos en las áreas de Educación, Ciencia y Política y el perfil C6 que tienen una alta tasa de proyectos y productos científicos del área de Información y comunicación. La visión más granular, al aplicar K-Means, presentó menos perfiles de bajo nivel y más perfiles de alto rendimiento que en la técnica 1, como el perfil C6

que tienen productos científicos de alto impacto afines al área de Política y economía o el perfile C8 que también produce artículos de alto impacto en el área de Ciencias sociales.

El clustering de K-Means tiene un mayor grado de detalle de los perfiles de investigadores con un mayor puntaje de silueta que al aplicar el clustering de Louvain, pero con un mayor número de clústeres e incluso clústeres formados por una sola persona que podrían estar categorizados como valores atípicos. Este clustering presentó un espectro completo de investigadores clasificados por su rendimiento entre los niveles bajo, medio y alto.

Al aplicar el clustering de Louvain se obtuvo una medida de silueta baja, pero un número de clústeres menor a los clústeres en K-Means, esto debido a que al ser un algoritmo codicioso tiende a generalizar los grupos. Los clústeres a partir de Louvain presentan una clasificación de investigadores más optimista, debido a que tiende a clasificar a más investigadores como de alto rendimiento y no tienen investigadores de bajo rendimiento en la clasificación de sus clústeres.

La técnica 1 y la técnica 2 presentan diferencias en la conformación de las métricas internas de sus clústeres, siendo la técnica 1 la que tiene mayores investigadores similares dentro de sus clústeres en relación con el resto de los clústeres, lo que indica, mediante las medidas de silueta, que hay clústeres más sólidos con el dataset de la técnica 1 después de la fase de minería de datos.

7.2. Líneas de trabajo futuro

Dentro de la conformación de los datasets se deberían agregar SDRs de otras organizaciones que indicen los trabajos de los investigadores de las IES, esto con el fin de aumentar el rango de investigadores que se incluyen en la formación de perfiles, debido a que otro porcentaje como es el caso de las Ciencias sociales generalmente no publican en Scopus.
Se puede abordar otras aproximaciones para la agrupación de los investigadores, con diferentes técnicas de clustering y tomar en cuenta la prominencia de los temas de investigación en relación con el resto del mundo o limitar el corte de la fecha de los productos científicos a los más recientes. También se podrían medir variables cualitativas de la apreciación de expertos en el área de la dirección de investigación, realizando una prueba con respecto a los resultados del clúster, para, de esta manera, recibir y contrastar esa retroalimentación.

8. Bibliografía

[1] L. H. Marcial and B. M. Hemminger, “Scientific data repositories on the Web: An initial survey,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 10, pp. 2029–2048, 2010.

[2] A. Abbasi and J. Altmann, “On the correlation between research performance and social network analysis measures applied to research collaboration networks,” in *2011 44th Hawaii International Conference on System Sciences*, 2011, pp. 1–10.

[3] L. Colledge and R. Verlinde, “Scival metrics guidebook,” *Netherlands: Elsevier*, 2014.

[4] Elsevier B.V., “Elsevier Developers,” 2019. [Online]. Available: https://dev.elsevier.com/.

[5] D. LOSHIN, *Data cleansing*. 2007.

[6] M. Sato and L. C. Jain, “Innovations in fuzzy clustering: Theory and applications (Studies in fuzziness and soft computing),” 2006.

[7] P. Kenekayoro, K. Buckley, and M. Thelwall, “Clustering research group website homepages,” *Scientometrics*, vol. 102, no. 3, pp. 2023–2039, 2015.

[8] J. Bollen, H. de Sompel, A. Hagberg, and R. Chute, “A principal component analysis of 39 scientific impact measures,” *PLoS One*, vol. 4, no. 6, p. e6022, 2009.

[9] F. De-moya-anegón, E. Herrán-páez, A. Bustos-González, E. Corera-álvarez, G. Tibaná-herrera, and F. Rivadeneyra, “Ranking Iberoamericano de instituciones de educación superior 2019 (SIR Iber).,” pp. 1–121, 2019.

[10] L. Waltman, N. J. Van Eck, and E. C. M. Noyons, “A unified approach to mapping and clustering of bibliometric networks,” *J. Informetr.*, vol. 4, no. 4, pp. 629–635, 2010.

[11] L. Šubelj, N. J. van Eck, and L. Waltman, “Clustering scientific publications based on citation relations: A systematic comparison of different methods,” *PLoS One*, vol. 11, no. 4, p. e0154404, 2016.

[12] A. Maedche and V. Zacharias, “Clustering ontology-based metadata in the semantic web,” in *European conference on principles of data mining and knowledge discovery*, 2002, pp. 348–360.

[13] A. Hotho, S. Staab, and G. Stumme, “Ontologies improve text document clustering,” in *Third IEEE international conference on data mining*, 2003, pp. 541–544.

[14] M. Bernotas, K. Karklius, R. Laurutis, and A. Slotkien\.e, “The peculiarities of the text document representation, using ontology and tagging-based clustering technique,” *Inf. Technol. Control*, vol. 36, no. 2, 2015.

[15] A. L. Tello, “Ontologías en la Web semántica,” *España Univ. Extrem.*, 2001.

[16] I. Nonaka, H. Takeuchi, and M. H. Kocka, *La organización creadora de conocimiento: cómo las compañías japonesas crean la dinámica de la innovación*. Oxford University

Press México DF, 1999.

[17] F. Van Harmelen, V. Lifschitz, and B. Porter, *Handbook of knowledge representation*, vol. 1. Elsevier, 2008.

[18] G. Antoniou and F. Van Harmelen, *A semantic web primer*. MIT press, 2004.

[19] M. Horridge, “Protege OWL tutorial,” 2009.

[20] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.

[21] M. C. Suarez-Figueroa, A. Gomez-Perez, E. Motta, and A. Gangemi, “Introduction: Ontology engineering in a networked world,” in *Ontology Engineering in a Networked World*, Springer, 2012, pp. 1–6.

[22] B. González-Pereira, V. Guerrero-Bote, and F. Moya-Anegon, “The SJR indicator: A new indicator of journals’ scientific prestige,” *arXiv Prepr. arXiv0912.4141*, 2009.

[23] M. C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas, “How to write and use the ontology requirements specification document,” in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2009, pp. 966–982.

[24] W3C, “Simple Knowledge Organization System,” 2012. [Online]. Available: https://www.w3.org/2004/02/skos/.

[25] University of Southern California, “Karma a Data Integration Tool,” 2016. [Online]. Available: https://usc-isi-i2.github.io/karma/.

[26] Ontotext, “GraphDB,” 2019. [Online]. Available: http://graphdb.ontotext.com/.

[27] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, “Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web,” *Semant. Web*, vol. 8, no. 3, pp. 437–452, 2017.

[28] T. Bosch *et al.*, “DDI-RDF Discovery Vocabulary,” 2015. [Online]. Available: http://rdf-vocabulary.ddialliance.org/discovery.html.

[29] D. Shotton and S. Peroni, “Funding, Research Administration and Projects Ontology (FRAPO),” 2015. [Online]. Available: http://www.sparontologies.net/ontologies/frapo.

[30] VIVO, “VIVO,” 2019. [Online]. Available: https://duraspace.org/vivo/.

[31] J. A. Pastor Sánchez, “UNESKOS Vocabulary,” 2015. [Online]. Available: http://skos.um.es/TR/uneskos/.

[32] J. A. Pastor Sánchez, “Tesauro de la UNESCO,” 2013. [Online]. Available: http://skos.um.es/unescothes/.

[33] R. Arp, B. Smith, and A. D. Spear, *Building ontologies with basic formal ontology*. Mit Press, 2015.