Artículos

Aplicación de la minería de texto como técnica de análisis para la identificación de enfoques temáticos en un grupo de investigación

Synchronized telerehabilitation care in rural zones supported by teleoperated technological aids: application to a Colombian case

Romero Pérez, Ivón; Muñoz Reyes, Yessica

Ivón Romero Pérez ivonromeroperez@gmail.com
Universidad Simón Bolívar, Colombia
Yessica Muñoz Reyes
Universidad Simón Bolívar, Colombia

Investigación e Innovación en Ingenierias
Universidad Simón Bolívar, Colombia
ISSN-e: 2344-8652
Periodicidad: Frecuencia continua
vol. 10, núm. 2, 2022
revingenieria@unisimonbolivar.edu.co

Recepción: 17 Agosto 2022
Aprobación: 19 Octubre 2022

URL: http://portal.amelica.org/ameli/journal/778/7784089013/

DOI: https://doi.org/10.17081/invinno.10.2.5907

Resumen: Objetivo: El propósito de este artículo es develar las características, enfoques temáticos y dinámica de trabajo de un grupo de investigación a partir de la aplicación de la minería de texto como técnica de análisis de información.
Método: Se desarrolló un estudio bibliométrico de carácter cuantitativo basado en la aplicación de técnicas lexicométrica y de minería de texto. La muestra fue intencional y no probabilística, compuesta por 185.212 ocurrencias de 40 documentos publicados por un grupo de investigación registrado en la plataforma GrupLAC del del Sistema de Información Científica de Colombia (ScienTi).
Resultados y discusiones: Como hallazgos se extrajeron los dominios temáticos, objeto de estudio, población y contextos de la producción intelectual, se identificó la dinámica de trabajo de los investigadores y se establecieron los principales enfoques temático de investigación del grupo. Sin embargo, para validar estos resultados fue indispensable la participación de un experto en la identificación de dichas categorías de análisis.
Conclusiones: la técnica de la minería de texto es una herramienta fundamental para evaluar la producción y dinámica de trabajo no solamente un grupo de investigación sino también de instituciones o centros que desean conocer con certeza los enfoques temáticos que desarrollan los investigadores. Esta información contribuye al fortalecimiento de las políticas públicas que estén orientadas al desarrollo de la investigación científica de un país.

Palabras clave: Análisis bibliométrico, Dominios temáticos, Enfoques de investigación, Lexicométrica, Minería de texto.

Abstract: Objective: The purpose of this article is to reveal the characteristics, thematic approaches and work dynamics of a research group based on the application of text mining as an information analysis technique.
Method: A quantitative bibliometric study was developed based on the application of lexicometric and text mining techniques. The sample was intentional and non-probabilistic, composed of 185,212 occurrences of 40 documents published by a research group registered in the GrupLAC platform of the Colombian scientific information system.

Results and discussions: As findings, the thematic domains, object of study, population and contexts of intellectual production were extracted, the researchers' work dynamics were identified and the main thematic research approaches of the group were established. However, in order to validate these results, the participation of an expert in the identification of these categories of analysis was indispensable.
Conclusions: the text mining technique is a fundamental tool to evaluate the production and work dynamics not only of a research group but also of institutions or centers that wish to know with certainty the thematic approaches developed by researchers. This information contributes to the strengthening of public policies oriented to the development of scientific research in a country.

Keywords: Bibliometric analysis, Thematic domains, Research approaches, Lexicometrics, Text mining.

Introducción

La tendencia actual a evaluar la actividad científica ha tomado en los últimos años gran fuerza, tanto que se ha convertido en una herramienta fundamental para los gobiernos de los países al momento de determinar el grado de avance y desarrollo científico que tienen sus instituciones, grupos o comunidades científicas en uno o más campos de conocimiento. Por ello, la ciencia cumple una función social y su progreso se condiciona a la dinámica de producción de la comunidad científica, esto hace que el sistema científico de un país sea evaluado para determinar sus fortalezas y debilidades, lo que conlleva al análisis de la producción científica y a la planificación de acciones que conduzcan al fortalecimiento de las políticas públicas para el desarrollo de la investigación científica y tecnológica de un país [1].

En tal sentido, la evaluación de la ciencia se constituye como un importante elemento ya que permite asignar recursos para la investigación y el desarrollo tecnológico. Estos resultados se convierten en insumos necesarios para que los líderes gubernamentales puedan tomar decisiones acertadas frente a la asignación y distribución eficiente de dichos recursos [2]. Por ello, existen estudios que evalúan la actividad científica a través de procesos, métodos y técnicas especializadas, de modo que, es posible determinar los perfiles, características, dinámicas, estructuras y tendencias de investigación de un país, institución, centro, grupo o comunidad científica de manera general o específica

Estos estudios son los denominados *estudios bibliométricos* que permiten conocer las temáticas y las líneas de investigación en un determinado campo científico, así como identificar los centros y grupos de investigación que están produciendo, la relación entre géneros de los investigadores, las citaciones entre las publicaciones y las auto citas; es decir, todos los factores que permiten tener una perspectiva clara sobre la situación actual de la investigación en un campo concreto [3].

Actualmente, estos estudios generan importantes indicadores que proporcionan un análisis concreto de la actividad científica, aportando datos sobre la situación científica de un país y permitiendo evaluar su rendimiento y su impacto en la comunidad; además, suministran información de la producción,

visibilidad e impacto no solo de las publicaciones científicas sino de los investigadores que están trabajando al interior de los centros o grupos de investigación [4].

Desde este punto de vista, este artículo presenta un análisis descriptivo que tiene como finalidad develar los enfoques temáticos y las dinámicas de producción del grupo de investigación Joaquín Aarón Manjarrés del campo de las ciencias sociales y jurídicas generado durante el periodo 2011 a 2017. Para esto, se realizó un estudio bibliométrico a partir de la aplicación de técnicas de minería de texto, tomando como fuente de información la producción disponible en la plataforma GrupLAC del Sistema de Información Científica de Colombia (ScienTi), la cual almacena la información de los grupos de investigación reconocidos por el Ministerio de Ciencia Tecnología e Innovación (Minciencias).

En la primera sección de este artículo se encuentra la perspectiva teórica del estudio en donde se presentan los antecedentes teóricos con relación al objeto de análisis. Posterior a esto, se describe el método de la investigación empleado, los resultados y la discusión sobre los hallazgos encontrados. Finalmente se plantean las conclusiones que están orientadas a los centros y grupos de investigación para que estos puedan evaluar la dinámica y producción de su comunidad, pero específicamente para fortalecer los determinados frentes de investigación.

Perspectiva teórica

La minería de texto una herramienta bibliométrica

La bibliometría es una disciplina que se emplea en los procesos de evaluación de la ciencia y se aplica por medio de herramientas computacionales que sirven para el análisis de grandes volúmenes de información. Además, aporta los instrumentos necesarios para medir la actividad científica especialmente desde el punto de vista cuantitativo se basa en el análisis de las fuentes bibliográficas [1] y sus resultados son insumos para optimizar la toma de decisiones [6,7].

En este sentido, las herramientas utilizadas para realizar este tipo de estudios son los llamados *indicadores bibliométricos*. Estos indicadores son datos numéricos que son extraídos de los documentos que publican los investigadores y permiten analizar las distintas características de su actividad científica vinculada tanto a su visibilidad, producción e impacto como a su consumo de información [8].

Dentro de estos indicadores bibliométricos existen varias tipologías, sin embargo, algunos autores los agrupan en cuatro categorías [1]. La primera, corresponde a los indicadores de producción o actividad, estos se basan en la enumeración y cuantificación de las publicaciones que se generan de un autor, grupo de investigación, institución, disciplina o país durante un periodo de tiempo determinado [9]. La segunda, hace referencia a los indicadores de impacto o visibilidad, estos proporcionan información sobre la influencia y el reconocimiento que tienen los autores y los trabajos publicados dentro de la comunidad científica. Entre ellos se distinguen el índice de citaciones, el factor de impacto y el índice h.

La tercera categoría incluye los indicadores de colaboración o relacionales que determinan la actividad y la cooperación científica entre autores o instituciones [2]. En este se distinguen el índice de coautoría, utilizados para evaluar la colaboración entre autores e instituciones con el fin de determinar desde un

ámbito local, nacional o internacional el grado y el tipo de relaciones entre los autores y las instituciones [10]. Y por último se encuentran los indicadores de asociaciones temáticas, estos se dividen en tres enfoques: el análisis de citas comunes, las referencias comunes y el análisis de palabras comunes o también conocido como análisis de palabras asociadas u co-ocurrencia [1].

Este último enfoque se ha convertido en una de las herramientas bibliométricas contemporáneas más utilizada para estudiar la dinámica y estructura de las comunidades científicas, fue desarrollada por el *Centre de Sociologie de I'Innovation* inicialmente para estudiar las redes y mapas de conocimientos que subyacen al analizar la co-ocurrencia de palabras que aparecen en los documentos científicos.

De esta manera, el método de palabras asociadas es una técnica que permite la representación y visualización de las estructuras científicas por medio de las palabras claves en el contenido de dos o más documentos científicos o académicos, esto ayuda a determinar la relación que existe entre los actores de un determinado campo de investigación facilitando así la toma de decisiones [11].

Por su parte, el análisis de las palabras asociadas es el estudio de las veces que aparecen dos términos en un texto dado cuyo fin es identificar la estructura conceptual y temática de un dominio científico. Este método de palabras se utiliza en diferentes disciplinas para develar las relaciones o similitudes en un conjunto de datos, tal es el caso de la minería de texto [12].

La minería de texto es el procesamiento de la información digital que se encuentra disponible en grandes volúmenes de documentos, mediante el uso de técnicas de análisis [12]. Es decir, permite la identificación y la extracción de patrones y relaciones interesantes del contenido textual de varios documentos transformándola en un lenguaje de fácil comprensión para los usuarios [13, 14, 15, 16, 17].

En definitiva, se considera la minería de texto como un proceso automático donde a través de métodos y técnicas se extrae y procesa información que es utilizada para la identificación de perfiles, relaciones, tendencias y patrones que estaban antes ocultos.

Algunos autores en la literatura afirman que, para lograr este propósito, la minería de texto debe hacerlo a través de determinadas técnicas tales como: la identificación de términos o conceptos inherente que emergen de las relaciones y semejanzas que se evidencian, el agrupamiento de documentos similares *(clustering),*la categorización automática de textos y la visualización de colecciones de textos [18]. Estas técnicas ayudan a que la información sea útil y fácil de comprender facilitando de esta manera la toma de decisiones.

Asimismo, existen diversos estudios bibliométricos aplicados en las distintas áreas del conocimiento que utilizan la minería de texto como herramienta para el análisis y el procesamiento de grandes cantidades de información. La mayoría de estos estudios están siendo empleados para medir la producción científica de las instituciones, centros o grupos de investigación; sin embargo, este tipo de evaluación en América Latina especialmente en el campo de las Ciencias Sociales y Jurídicas son poco frecuente y bastante compleja de aplicar por la heterogeneidad que existen entre las disciplinas que la conforman.

No obstante, existen algunos estudios que han demostrado las bondades de analizar grandes cantidades de documentos a través de la minería de texto, entre

ellos, se destaca un estudio donde se determinó la dinámica y evolución del campo de la cienciometría mediante la identificación de patrones de conocimiento, cambios y tendencias de investigación asociadas a las palabras claves, registradas en los títulos y resúmenes de artículos publicados en la revista *Scientometrics* durante el periodo 2005-2010 [19]. Para ello, se basaron en la utilización del método de palabras asociadas y la utilización de la minería de texto como herramienta fundamental para el procesamiento, análisis y visualización de la información.

Por otra parte, en un trabajo de investigación el cual tuvo como objetivo organizar el material bibliográfico utilizando un clasificador automático, se aplicó la minería de texto y el análisis de los registros bibliográficos utilizando la base de datos LIBRUNAM [20], en ella, tomaron como referencia los documentos de la clasificación Z pertenecientes al área de bibliotecología, ciencias de la información y recursos de información. La aplicación de la minería de texto en este trabajo permitió la identificación de patrones y temas asignados en un conjunto de documentos facilitando el proceso de organización de los documentos digitales.

De igual manera, en otro estudio se analizaron las publicaciones de Iberoamérica sobre Ciencia, Tecnología y Sociedad correspondiente al periodo comprendido entre 1970-2013 y las cuales fueron extraídas de la base de datos de la *Web of Science (WoS)*[21]. En este estudio, utilizaron la técnica de *clúster*bibliográfico para la identificación de estructura temática de la investigación en CTS. En este sentido, el *clustering* es aquel que agrupa documentos de acuerdo con la similitud que haya entre ellos [22]. Los autores en este trabajo tomaron las publicaciones que tenían semejanzas en sus referencias, utilizaron la herramienta *Biblio Tools* para el análisis de los *clústeres*, e identificaron las diferentes líneas temáticas. Algunos autores afirman que el análisis de clústeres bibliográficos permite determinar las principales características de la producción científica, específicamente palabras claves y temas comunes [21].

Asimismo, otros autores buscaban aplicar las técnicas de categorización automática de documentos a través de la minería de texto [23]. Hallaron que la categorización de documentos de texto es una aplicación de la minería de texto en la cual se atribuye a los documentos una o más categorías que ayuda a la organización de tareas y la gestión de la información. En este estudio los autores aseguran que las técnicas aplicadas (recuperación de información y aprendizaje automático) facilitan la extracción del conocimiento y la asignación automática de categorías dentro de un conjunto de texto.

En otro trabajo que tuvo como finalidad extraer datos de los perfiles y publicaciones de 15 Universidades en Google Scholar utilizando la técnica de minería de texto no estructurada se implementó un algoritmo en el lenguaje R que permitió automatizar el proceso, estructurar los datos, mejorar el tiempo de extracción, y analizar la información de forma rápida y oportuna [24].

Finalmente, con base en los estudios encontrados en la literatura se puede afirmar que, la minería de texto se convierte en un recurso bibliométrico que permite analizar documentos científicos almacenados en múltiples plataformas de información, con la finalidad de que puedan ser utilizada por investigadores, grupos, centros e instituciones de investigación para el estudio de la dinámica

y estructura de conocimiento de un grupo de investigación a partir de su producción intelectual, empleando métodos y técnicas estadísticas que ayuden a la asociación y agrupamiento automático de los datos, definición de categorías, enfoques y líneas de investigación poco perceptibles para los analistas de contenido.

Método

El propósito fundamental de este estudio fue develar las características, enfoques temáticos y dinámica de trabajo de un grupo de investigación en el campo de las Ciencias Sociales y Jurídicas. Para esto, se desarrolló un estudio bibliométrico de carácter cuantitativo [25, 26] a partir de la aplicación de la minería de texto como técnica de análisis de información [27].

Este estudio se basa en la evaluación de indicadores bibliométricos derivado del análisis lexicométrico de los resultados generados a partir de la aplicación de técnicas de minería de texto utilizadas para extraer patrones de conocimiento en datos textuales con la cuantificación y relación automática de las ocurrencias de las unidades de análisis (palabras y expresiones frecuentes) que aparecen frecuentemente en un corpus, que luego son analizadas por medio de un tratamiento estadístico [28, 29, 30].

Población y Muestra

Para la realización de esta investigación se contó con toda la producción bibliográfica de un grupo de investigación perteneciente al campo de las ciencias sociales y jurídicas registrada durante el periodo 2011 a 2017 en la plataforma GrupLAC del Sistema Nacional de Ciencia, Tecnología e Innovación (ScienTi) administrada por Minciencias. La muestra fue intencional y no probabilística y para ello se seleccionaron 40 documentos (artículos, libros y capítulos) en el idioma español, disponibles en formatos legibles y digitales (Pdf, Txt o Word), registrados por el grupo de investigación Joaquín Aarón Manjarrés (GIJAM), tal como se describe en la Tabla 1.

Tabla 1
Distribución de la muestra

Distribución de la muestra	Cantidad
Número de artículos	31
Número de libros	4
Número de capítulos de libros	5
Total de documentos que representa la muestra	40
Número de ocurrencias	185.212

Elaboración propia.

Técnicas de recolección y análisis de la información

De manera específica para la recolección de la información se utilizaron las técnicas de escalonamiento multidimensional (MDS) y *clustering* que permitieron organizar y agrupar de forma automática los datos de acuerdo con ciertas unidades de medida. Para el análisis de la información obtenida se

utilizaron códigos de programación procesados con R y el *software IraMuteQ* . *Orange*.

Proceso metodológico

Para la evaluación de los indicadores bibliométricos de este estudio se empleó el siguiente proceso metodológico: en primer lugar, se partió de la identificación de las categorías emergentes derivada del análisis lexicométrico y del agrupamiento de información textual obtenida a través de la aplicación de técnicas de minería de texto. Esto se realizó siguiendo las pautas y los procedimientos implementados por autores que han realizado este tipo de análisis [18, 31, 30, 20, 32, 33]. Estas pautas son descritas a continuación:

a. **Recuperación de información:** es una etapa a través del cual se establecen los objetivos del estudio y se determinan las fuentes de información y los resultados esperados. Para este estudio, el objetivo se centró en identificar las características y los enfoques de investigación del grupo de investigación con el propósito de apoyar la toma de decisiones frente a redefinición de sus líneas de investigación. Para ello, se seleccionaron, se recuperaron y se extrajeron los contenidos de los artículos, libros y capítulos de libros publicados por dicho grupo.

b. **Pre-procesamiento:** en este punto se busca reducir la información, eliminando, normalizando y extrayendo los términos relevantes para el estudio. Para ello, se eliminaron las palabras vacías *(stop Word),* es decir, los símbolos, caracteres y términos poco relevantes para el estudio, tales como: preposiciones, artículos, conjunciones, signos de puntuación, entre otros. Seguidamente, se convirtieron todas las palabras a minúscula; se aplicó la técnica de lematización *(stemming)* para normalizar, agrupar y reducir el conjunto de palabras a la forma lingüística que guardan un significado semántico equivalente a un lema o raíz léxica original [34, 18]. Por último, se aplicó la técnica *tokenización* para segmentar el corpus en unidades de análisis básicas; en este caso, palabras y expresiones frecuentes, con las cuales, se infieren categorías y conceptos claves que representan el corpus [35, 7].

c. **Matriz de término-documentos:** indudablemente para analizar estas relaciones y similitudes entre las palabras y los documentos es necesario transformar los datos en un modelo de espacio vectorial [36], en donde el corpus es representado por una matriz de término y documentos compuesto de . filas que corresponde a los documentos y columnas que representan los términos [32]. En esta etapa, se almacena "la frecuencia con que cada una de las palabras (formas gráficas) de los vocabularios son utilizadas por los individuos de cada categoría de la variable" [37]. Para representar este vector se utilizó el modelo *Bag of words (BoW)* y la *función TF-IDF,* con la cual se determina la frecuencia de aparición de cada término en cada documento y se determina la relevancia de cada término a partir de la frecuencia inversa en el conjunto de documentos.

d. **Clustering o agrupamiento automático:** como técnica de reducción de los datos, se empleó el *clustering* jerárquico que sirve para agrupar y representar de forma automática los documentos en pequeños subgrupos o clúster según las características y similares que comparten [38, 18]; así como, la técnica de escalonamiento multidimensional (MDS) para visualizar los datos como puntos sobre un espacio bidimensional, con el propósito de identificar las similaridad entre los documentos por medio de una medida de distancia [39].

e. **Análisis e interpretación de los resultados:** de acuerdo con los estudios de análisis textual [14, 40], estos deben realizarse con la ayuda de un experto temático cuyo fin sea interpretar, evaluar y validar de manera acertada los patrones de conocimiento que derivan de la aplicación de las distintas técnicas de análisis, en este caso, para identificar los términos, categorías, expresiones y grupos representativos.

Resultados y discusión

Teniendo en cuenta el objetivo de este estudio el cual fue develar las características, enfoques temáticos y dinámica de trabajo de un grupo de investigación en el campo de las Ciencias Sociales y Jurídicas, en este apartado se presentan los resultados encontrados a partir de la metodología empleada.

Categorización de las unidades de análisis

Los resultados de la aplicación del análisis lexicométrico permitieron extraer un conjunto de palabras claves que fueron la base para la definición de las variables categóricas asociadas con la producción intelectual del grupo de investigación. Con la utilización del *software IraMuteQ* versión 0.7 se obtuvieron 5.926 formas activas de un total de 15.548. (Tabla 2).

Tabla 2
Resumen del análisis lexicométrico

Característica de la información	Cantidad
Número de textos	40
Número de ocurrencias	185.212
Número de formas	15.548
Número de formas activas	5.926
Media de ocurrencias por texto	4630,30

Elaboración propia.

Estos resultados han sido representados por una nube de palabras (Figura 1) que muestra la relevancia que existe en distintas unidades léxicas dentro de un conjunto de datos que fueron delimitadas sólo a las formas activas (palabras) que tiene una frecuencia mayor o igual a 100.

Figura 1
Nube de formas activas frecuentes
Elaboración propia.

Como puede observarse, las palabras *Derecho y Social* son las formas activas (palabras) más representativas dentro del conjunto de datos, esto, tiene sentido teniendo en cuenta que la producción intelectual del grupo de investigación pertenece al campo de las Ciencias Jurídicas y Sociales. La identificación de estos términos es un primer acercamiento hacia la representación del discurso que emplea el grupo de investigación.

Esta nube palabras se convierte en un recurso visual para develar los conceptos claves e ideas principales, en este caso, es evidente que el grupo investiga sobre temas jurídicos, políticos y sociales, pero también, aborda temáticas asociadas con los procesos de enseñanza y aprendizaje en el contexto jurídico, una temática que no era tan evidente antes del procesamiento y análisis de los datos.

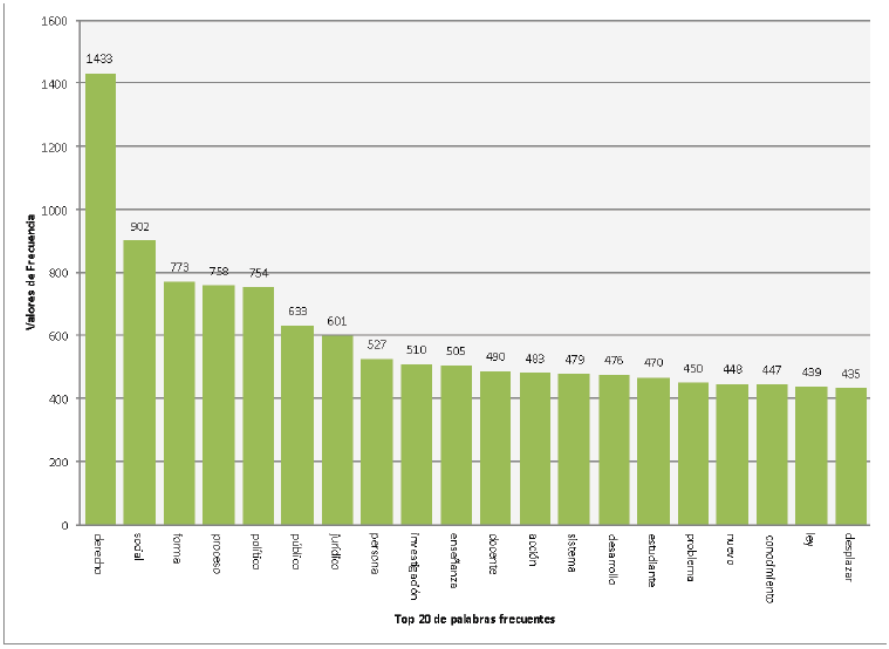

Figura 2
Top 20 de palabras frecuentes
Elaboración propia.

A diferencia de la nube de palabras, la Figura 2 es otro recurso visual que permite observar las palabras claves y su nivel de significancia de forma más detallada, porque en este se jerarquiza y organiza cada palabra según su frecuencia de uso. Para este estudio, se seleccionaron sólo las 20 palabras más representativas, destacándose nuevamente: Derecho (1433) y Social (902) como la más utilizadas, seguida de: forma (773), proceso (758), político (754), público (633), jurídico (601), persona (527), investigación (510), enseñanza (505), docente (490), acción (483), sistema (479), desarrollo (476), estudiante (470); problema (450) y ley (439). También, se identificaron palabras emergentes como: nuevo (448), conocimiento (447), y desplazar (435), con las cuales, se infieren unos enfoques temáticos asociados a la generación de nuevo conocimiento y los problemas de desplazamiento.

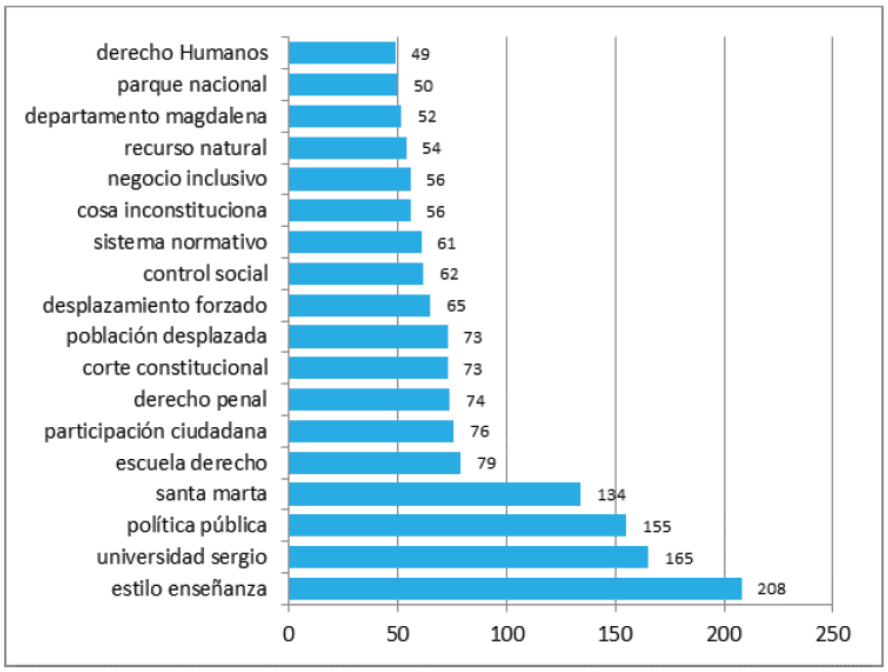

Figura 3
Principales expresiones del grupo
Elaboración propia.

Aplicando otros procedimientos con el *software R*, se logró extraer y seleccionar las 18 expresiones frecuentemente utilizadas por los investigadores en sus publicaciones (Figura 3), permitiendo de esta manera avanzar hacia la identificación y construcción de las categorías de análisis asociadas con la producción intelectual del grupo de investigación.

Con estos resultados y con la ayuda de un experto, se infirió el *dominio de conocimiento* como una categoría de análisis que emerge a través del uso de las siguientes palabras frecuentes: derecho, forma, proceso, político, público, jurídico, ley, sistema y norma; así como, de las expresiones derecho fundamental, derecho penal, sistema normativo, corte constitucional y política públicas.

Seguidamente, se identificó la categoría *Objeto de estudio*, con las palabras: social, problema, desarrollo, conocimiento, vida, investigación, desplazamiento, nuevo, y acción; incluyendo las expresiones: política pública, participación ciudadana, control social, desplazamiento forzado, negocio inclusivo, recursos naturales y estilo de enseñanza. Asimismo, se estableció la categoría Población de la expresión población desplazada. Finalmente, se infirió la categoría *Contexto*a partir del término Colombia y las expresiones departamento del Magdalena, Santa Marta, Universidad Sergio Arboleda, escuela de derecho y parques nacionales.

Teniendo en cuenta lo anterior y con el propósito de facilitar la interpretación de estos resultados, se organizó en la Tabla 3 cada una de las categorías emergentes de acuerdo con las 20 expresiones y 25 palabras frecuentemente utilizadas en el discurso de la producción científica del grupo de investigación.

Tabla 3
Categorías de análisis

Categoría	20 expresiones frecuentes	25 palabras frecuentes
Domino de conocimiento	Derecho fundamental	Derecho, forma, proceso, político, público, jurídico, ley, sistema y norma
	Derecho penal	
	Sistema normativo	
	Corte constitucional	
	Política pública	
Objeto de estudio		Social, problema, desarrollo, conocimiento, vida, investigación, desplazamiento, nuevo, y acción
	Participación ciudadana	
	Control social	
	Desplazamiento forzado	
	Negocio inclusivo	
	Recursos naturales	
	Estilo de enseñanza	Enseñanza, docente, estudiante, estilo, sistema y actividades
Población	Población desplazada	Persona y grupo
Contexto	Colombia	Colombia
	Departamento del Magdalena	
	Santa Marta	
	Universidad Sergio Arboleda	
	Escuela de derecho	
	Parque nacional	

Elaboración propia.

Análisis de Clúster

El análisis de clúster fue otro recurso utilizado para agrupar y visualizar los documentos que guardan ciertas similitudes temáticas, en esto se determinan los enfoques de investigación y quiénes son los autores que fortalecen dichas líneas, asimismo, cómo se reagrupan internamente los autores como un equipo de trabajo.

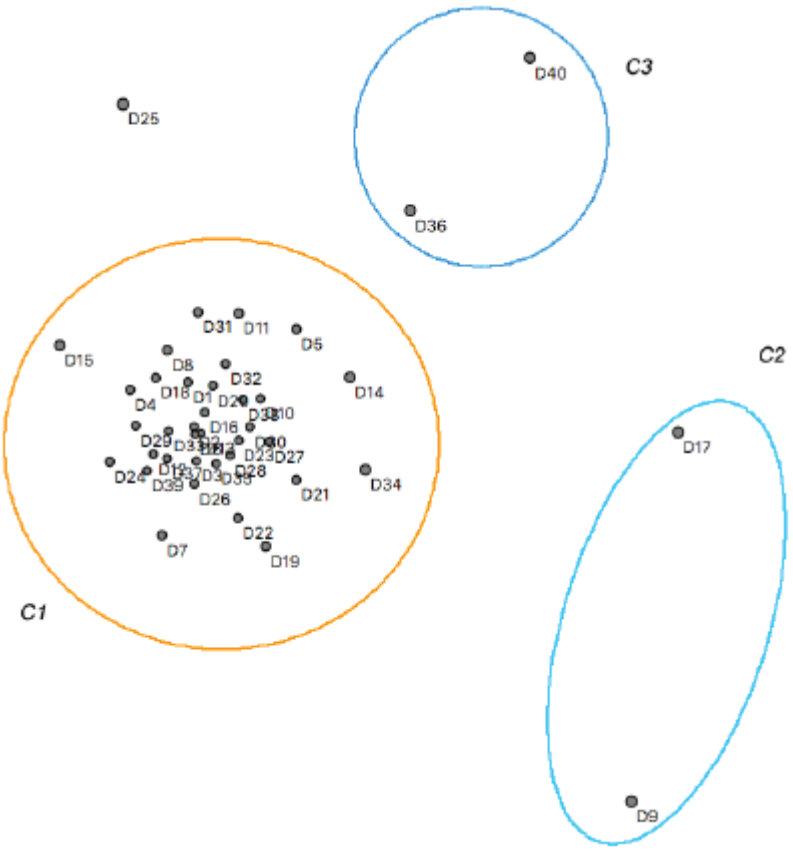

Figura 4
Clúster de documentos
Elaboración propia.

De acuerdo con la Figura anterior, el grupo de investigación se caracteriza por trabajar en tres enfoques temáticos y un tema emergente. El clúster C1, aglomera la mayor parte de la producción intelectual, lo que significa que en este se aglomera la base de conocimiento que define el enfoque temático principal del grupo de investigación, así como, los autores más representativos.

Este clúster, está conformado por 35 documentos: D1, D2, D3, D4, D5, D6, D7, D8, D10, D11, D12, D13, D14, D15, D16, D18, D19, D20, D21, D22, D23, D24, D26, D27, D28, D29, D30, D31, D32, D33, D34, D35, D37, D38 y D39 [41].

Para determinar con certeza el enfoque principal de este primer clúster fue necesario extraer las expresiones frecuentemente utilizadas (Tabla 4), estos terminan por constituirse en la base para inferir los subtemas de investigación predominantes al interior de este subgrupo. En este caso, se obtuvieron: estilos de enseñanza y política pública como las expresiones más frecuentes, ambas ancladas al contexto de la Escuela de derecho de la Universidad Sergio Arboleda de la ciudad de Santa Marta en el Departamento del Magdalena, así como, otros temas relacionados con el derecho, pero sobresaliendo temáticas emergentes como: la población desplazada, la participación ciudadana y los recursos naturales.

Tabla 4
Expresiones frecuentes del clúster 1

Expresiones Frecuentes	Frecuencia
Estilo enseñanza	208
Política pública	155
Santa Marta	134
Universidad Sergio Arboleda	165
Escuela de Derecho	79
Participación ciudadana	76
Derecho penal	74
Corte Constitucional	73
Población desplazada	73
Desplazamiento forzado	65
Control social	62
Sistema normativo	61
Negocio inclusivo	56
Cosa inconstitucional	55
Departamento Magdalena	52
Recurso natural	51
Derecho humano	47

Elaboración propia.

Por otra parte, para determinar la dinámica y estructura del *clúster* C1, se examinó la información contenida en este subconjunto de documentos. Los resultados obtenidos y presentes en la Figura 5, muestra en detalle documentos dispersos y un subgrupo de documentos altamente relacionados que han sido considerados como el Núcleo Central de Conocimiento (NCC) del clúster 1 por sus similitudes léxicas.

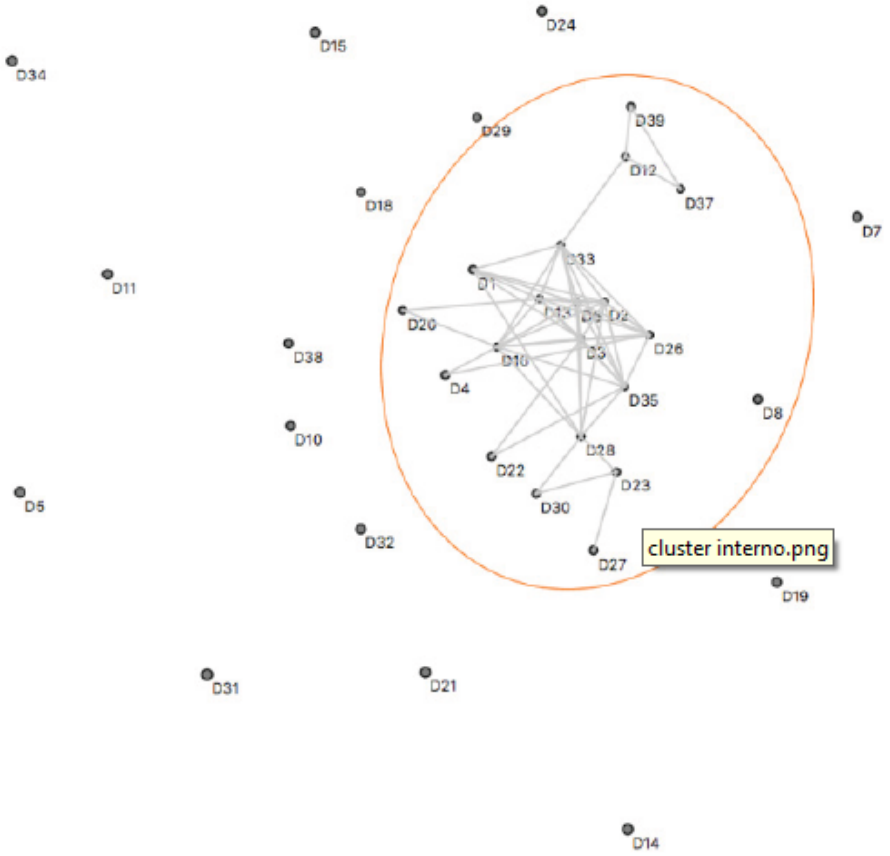

Figura 5
Núcleos de documentos del clúster 1
Elaboración propia.

Este núcleo de conocimiento permite visualizar con más detalle la estructura y la red relacional que se ha tejido alrededor de ocho (8) frentes de investigación altamente especializados, los cuales, han sido definidos a partir de las expresiones frecuentes de 15 documentos producidos por seis investigadores activos en el núcleo de conocimiento.

Tabla 5
Núcleo central de conocimiento del clúster 1

		Autores	Red de Documentos
Núcleo de enlaces del cluster 1	D1	Latorre, E.I. (2011a)	Expresiones Frecuentes Negocio inclusivo (56) Estilo de enseñanza (46) Comunidad indígena (14) Plan de Ordenamiento territorial (14) Semilla certifica (13) Recursos Naturales (13) Derecho Romano (11) Calidad de vida (11)
	D2	Latorre, E.I. (2011b)	
	D6	Latorre, E.I. (2013b)	
	D13	Latorre, E.I. (2015a)	
	D16	Latorre, E.I. (2015b)	
	D27	Latorre, E.I., Caballero, Y., Narváez, B., y Quiñones, A. (2015b)	
	D23	Latorre, E.I., Caballero, Y., Narváez, B., y Quiñones, A. (2015a)	
	D30	Latorre, E.I., Caballero, Y., Narváez, B., y Quiñones, A. (2015d)	
	D3	Polo, N. (2012)	
	D33	Polo, N. (2016a)	
	D20	Parra, J., Agudelo, C., Díaz, M., y Salazar, D. (2015)	
	D26	Díaz, M. (2015a)	
	D4	Correa & Cano (2014)	
	D22	Agualdelo, C. (2015b)	
	D35	Bustamante, A. (2016)	

Elaboración propia

Como puede observarse en la tabla 5, estos frentes de investigación especializados giran alrededor de temas asociados con: negocio inclusivo (56), estilo de enseñanza (46), comunidad indígena (14), Plan de Ordenamiento

Territorial (14), semilla certifica (13), recursos naturales (13), derecho romano (11) y calidad de vida (11). Cada una de estas expresiones ha sido tomada de los documentos: D1, D2, D13, D16, D30, D23, D27, D4, D6, D3, D33, D20, D26, D22, y D35, [41]. Entre ellos se destaca Latorre, E con el mayor número documentos relacionados dentro del núcleo, por tanto, este autor puede considerarse como el investigador principal dentro de estos frentes de investigación.

Por otra parte, el *clúster* C2 está conformado por los libros resultados de investigación correspondientes a los documentos D9 y D17. Como se observa en la Figura 4, ambos documentos se ubican distantes del clúster principal (C1), esto quiere decir que son documentos que han sido desarrollo desde un enfoque temático diferente.

Al analizar las palabras frecuentes que aglomeran este grupo, se infiere que el enfoque de investigación de estos investigadores se centra en temas del campo de las Ciencias de la Educación, principalmente porque utilizan palabras frecuentes como: estudiante (444), docente (362), educación (346), conocimiento (322), estilo (320), proceso (319), investigación (293), desarrollo (288), derecho (284), aprendizaje (221), pedagogía (206), actividad (196), práctica (181), formación (175), social (175), universidad (171), didáctica (163) y teatro (148). Además de utilizar expresiones como estilo de enseña (162), enseñanza y aprendizaje (28), estilo de aprendizaje (26), docente (24), acto pedagógico (23), modelo de enseñanza (23) y estrategia didáctica (22), todas estas descritas en la tabla 6:

Tabla 6

Expresiones frecuentes del clúster 2

Expresiones Frecuentes	Frecuencia
Estilo enseñanza	162
Escuela de derecho	67
Universidad Sergio Arboleda	53
Santa Marta	41
Rodrigo de Bastida	29
Derecho universidad	28
Enseñanza y aprendizaje	28
Estilo aprendizaje	26
Docente	24
Acto pedagógico	23
Ciencia jurídica	23
Modelo enseñanza	23
Estrategia didáctica	22
Objeto estudio	22
Abogado	21

Elaboración propia

Por otro parte, el *clúster C3* representado por los libros D36 y D40, se caracteriza por desarrollar otro enfoque temático especializado. En este caso, este grupo aborda temáticas como: experimentación jurídica (40), litigio estructural (25), veeduría ciudadana (24), América Latina (19), opinión pública (19) y gestión pública (18), tal como se muestra en la Tabla 7.

Tabla 7
Expresiones frecuentes del clúster 3

Expresiones Frecuentes	Frecuencia
Política pública	102
Participación ciudadana	73
Control social	58
Población desplazada	41
Experimentalismo jurídico	40
Desplazamiento forzado	38
Corte constitucional	31
Litigio estructural	25
Situación de desplazamiento	25
Veeduría ciudadana	24
Ciudadanía y control	23
Público privado	22
Departamento magdalena	20
Derecho fundamental	20
América latina	19
Opinión pública	19
Gestión pública	18
Promedio nacional	18
Servicio público	18

Elaboración propia

Con respecto al documento D25, existe un claro distanciamiento temático con el núcleo de conocimiento principal, sin embargo, por las expresiones descritas en la Tabla 8, se puede inferir que este documento aborda temas tanto del campo de las Ciencias Ambientales como de las Ciencias Jurídica y Sociales, es decir que es un clúster que se especializa por desarrollar temas de Derecho Ambiental.

Tabla 8
Expresiones frecuentes del clúster 4

Expresiones Frecuentes	Frecuencia
Área marina	34
Ecosistema marino-costero	32
Recurso natural	32
Parque nacional	45
Ecosistema marino	24
Área protegida	23
Medio ambiente	21
Protección del ecosistema	21
Zona costera	21
Diversidad biológica	17
Ley aprobatoria	17
Actividad económica	16
Marina costera	16
Ambiente sano	15
Fauna y flora	15
Marina protegida	13
Área marino-costera	10

Elaboración propia

También hay que destacar que es un documento que aborda temas relacionados con la protección de los recursos naturales (flora y la fauna) y de los

ecosistemas marino-costeros, especialmente, los que se encuentran en las áreas protegidas como son los parques nacionales.

Conclusiones

De acuerdo con los resultados obtenidos en este estudio se concluye que Joaquín Aarón Manjarrés (GIJAM) es un grupo de investigación que se caracteriza por indagar sobre las problemáticas sociales del contexto regional y local en el marco de los derechos humanos y dentro del contexto colombiano, especialmente de las poblaciones en condiciones vulnerables como son los desplazados y las formas de desarrollo social a través de la generación de empleo por medio de los denominados negocios inclusivos.

Asimismo, se caracteriza por desarrollar investigaciones principalmente en tres campos de conocimiento (Tabla 6), en primer lugar, en el Campo de las Ciencias Jurídicas y Sociales, el cual está anclado con la naturaleza, dinámica y aglomeración de gran parte de la producción intelectual; en segundo, en el campo de las Ciencias de las Educación, que integra publicaciones de los investigadores que trabajan sobre temas relacionados con los procesos académicos y estilos de aprendizaje de los estudiantes de derecho dentro del contexto educativo de la Universidad Sergio Arboleda, seccional Santa Marta; y en tercer lugar, en el campo de las Ciencias Ambientales a partir de las publicaciones desarrollada alrededor del derecho ambiental y la protección de los recursos naturales de las área protegidas como son los parques nacionales.

Tabla 8
Campos de ConocimientoGIJAM

Grupo	Documentos	Campo de Conocimiento	Lineas de investigación emergentes
C1	D1, D2, D3, D4, D5, D6, D7, D8, D10, D11, D12, D13, D14, D15, D16, D18, D19, D20, D21, D22, D23, D24, D26, D27, D28, D29, D30, D31, D32, D33, D34, D35, D37, D38 y D39	Ciencias Jurídicas y Sociales	Derecho Penal Derecho Constitucional Desplazamiento Forzado Negocios inclusivos
C3	D36 y D40		Política y gestión pública
C2	D9 y D17	Ciencias de la Educación	Estilos de enseñanza en el área del derecho
C4	D25	Ciencias Ambientales	Derecho Ambiental

Elaboración propia.

Con la producción de este grupo de investigación, se logró desarrollar un estudio bibliométrico que utiliza técnicas de minería de texto como herramienta de análisis para la identificación de los enfoques temáticos de una comunidad científica, llegando a la conclusión de que es una herramienta fundamental para examinar y analizar grandes volúmenes de información de naturaleza textual y que además permite develar las características y estructuras de la producción intelectual, así como, la red de conocimiento y trabajo que se teje alrededor de ella.

La técnica de clasificación no solo permitió corroborar los enfoques de investigación explícitos antes del estudio sino develó los temas emergentes y la dinámica de trabajo alrededor de un núcleo de conocimiento. Al comparar estos resultados con la información registradas en la plataforma GrupLAC también se pudo evidenciar que la dinámica de producción del grupo es coherente con la declaración de sus líneas de investigación.

Por tanto, esta herramienta puede ser útil para que los centros de investigación puedan evaluar la dinámica y producción de sus grupos de investigación, pero especialmente, para fortalecer los enfoques emergentes en coherencia con las

políticas públicas orientadas al desarrollo de la ciencia. Del mismo modo, puede ayudar a establecer los equipos de trabajo más apropiados para fortalecer determinados frentes de investigación.

En base a esta información, los gobernantes y líderes de investigación pueden tomar decisiones estratégicas frente a situaciones o problemas actuales que impacten positivamente en el campo disciplinar, definiendo o redefiniendo sus líneas o enfoques de investigación como estructuras de trabajo en equipo.

Referencias bibliográficas

[1] G. Pérez Reyes y A. Martínez Rodríguez, La Ciencia como empresa social: su evaluación desde la bibliometría. *Biblios [Revista electrónica]*. No. 55, pp. 27-39., 2014. DOI: https://doi.org/10.5195/ biblios.2014.157.

[2] R. Sancho, Indicadores bibliométricos utilizados en la evaluación de la ciencia y la tecnología. Revisión bibliográfica. *Revista española de documentación científica [Revista electrónica]*. Vol. 13. Núm. (3-4), pp. 842-865., 1990. Disponible en: ht tps://digital.csic.es/handle/10261/23694

[3] A. Castillo y M. Carretón, Investigación en comunicación: estudio bibliométrico de las revistas de comunicación en España. *Comunicación y Sociedad [Revista electrónica]*. Vol. XXIII, Núm. (2), pp. 289- 327., 2010. Disponible en: http://h dl.handle.net/10045/22678.

[4] V. Tomás Górriz y V. Tomás-Casterá, La Bibliometría en la evaluación de la actividad científica. *Hosp Domic [Revista electrónica]*. Vol. 2 Núm. (4), pp. 145-63., 2018. DOI: http://doi.org/10.22585/hospdomic. v2i4.51

[5] E. Spinak, Indicadores Cienciométricos. *Ciência da informação [Revista electrónica]*. Vol. 27 Núm. (2), pp. 141-148., 1998. Disponible en: http://revista.ibict.br/ciin f/article/view/795

[6] J. Amézquita López, D. Martínez Torres, J. Martínez Torres y F. Maza Ávila. Bibliometría, infometría y cienciometría. *Proyecto de investigación: Diseño e implementación de la catedra CTS+ I (Ciencia, Tecnología, Sociedad e Innovación) en la Universidad de Cartagena*. Cartagena: Ediciones Unicartagena, 2011.

[7] Y. Pérez Guadarramas, A. Rodríguez Blanco, A. Simón Cuevas, W. Hojas-Mazo, y J. Olivas, Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos. *Procesamiento del Lenguaje Natural [Revista electrónica]*. Núm. 59, pp. 39-46., 2017. Disponible en: http://j ournal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5491

[8] E. Sanz y C. Martin, Técnicas Bibliométricas aplicadas a los estudios de usuarios. *Revista General de Información y Documentación [Revista electrónica]*. Vol. 7 Núm. (2), pp. 42-64., 1997. Disponible en: https://revistas.ucm.es/index.php/RGID/a rticle/view/RGID9797220041A

[9] M. Peralta, M. Frías y O. Chaviano, Criterios, clasificaciones y tendencias de los indicadores bibliométricos en la evaluación de la ciencia. *Revista Cubana de Información en Ciencias de la Salud [Revista electrónica]*. Vol. 26 Núm. (3), pp. 290-309, 2015. Disponible en: http://www.redalyc.org/articulo.oa?id=3776457 62009.

[10] B. Velasco, J. Eiros, J. Pinilla y J. San Román, La utilización de los indicadores bibliométricos para evaluar la actividad investigadora, *Aula Abierta, [Revista electrónica]*. Vol. 40 Núm. (2), pp. 75-84., 2012. Disponible en: https://dialnet.u nirioja.es/servlet/articulo?codigo=3920967

[11] R. Ruiz y R. Bailón, El método de las palabras asociadas (I): La estructura de las redes científicas. *Boletín de la Asociación de la Revista Andaluza de Bibliotecarios [Revista electrónica]*. Vol. 14 Núm. (53), pp. 43-60, 1998. Disponible en: http://eprints.rclis.org/13003/

[12] C. Gálvez, Análisis de co-palabras aplicado a los artículos muy citados en Biblioteconomía y Ciencias de la Información (2007-2017). *Transinformação, v30(3),* 277-286., 2018. http://dx.doi.org/10.1590/2318- 08892018000300001

[13] D. Swanson, "Complementary structures in disjoint science literatures." Presented in Proceedings of the 14th Annual International ACM/SIGIR Conference. *University of Chicago*, September 1991. DOI: https://dl.acm.org/doi/10.1145/122860.122889

[14] M. Hearst, "Untangling text data mining." Presented in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. *Association for Computational Linguistics, University of California, June 1999*. https://aclanthology.org/P99-1001.pdf

[15] D. Sullivan, *Document warehousing and text mining*. New York: Wiley Computer Publishing, 2001.

[16] R. Feldman, J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Pres, 2007.

[17] E. Brun, y Senso, J, Minería textual. *El Profesional de la Información [Revista electrónica]*. Vol. 13 Núm. (1), pp. 11–27., 2004. Disponible en: http://eprints.rclis.org/11491/1/Artmineriapdf.pdf

[18] S. Ravikumar, A. Agrahari, y Y. Singh, Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics (2005–2010). *Scientometrics [Revista electrónica]*. Vol. 102 Núm. (1), pp. 929-955., 2015. https://doi.org/10.1007/s11192-014-1402-8

[19] B. Contreras, Minería de texto en la clasificación de material bibliográfico. *Biblios, [Revista electrónica]*. Núm. (64)., 2016. Disponible en: http://biblios.pitt.edu/ojs/index.php/biblios/article/view/309

[20] D. De Filippo, y L. Levin, Detección y análisis de "clúster bibliográfico" en las publicaciones de Iberoamérica sobre ciencia, tecnología y sociedad (1970-2013), *Investigación Bibliotecológica, [Revista electrónica]*. (número especial), pp. 123-148., 2017. Disponible en: http://rev-ib.unam.mx/ib/index.php/ib/article/view/57888/51919

[21] A. Valero. "Técnicas estadísticas en Minería de texto" (Tesis de grado). Dpto de Estadística e Investigación Operativa, *Universidad de Sevilla, Sevilla*, España. 2017. Disponible en: https://idus. us.es/handle/11441/63197

[22] M. Pérez, y C. Cardoso, Minería de texto para la categorización automática de documentos, *Cuadernos de la Facultad, [Revista electrónica]*. Núm. (5), pp. 11-45., 2010. Disponible en: https://www.ucasal. edu.ar/htm/ingenieria/cuadernos/archivos/5-p11-alicia-articulo-cuadernos-formateado.pdf

[23] D. Murillo y D. Saavedra. "Implementación de algoritmo en el Lenguaje R para extraer los datos de los Perfiles en Google Scholar utilizando la técnica web Scraping de Minería de datos". Trabajo presentado en el X Congreso de Computación para el Desarrollo, COMPDES [En línea]. 2017. https://ridda2.utp.ac.pa/bitstream/handle/123456789/3105/vipe-algoritmo-google-scholar. pdf?sequence=1&isAllowed=y

[24] M. Bordons y M. Zulueta, Evaluación de la actividad científica a través de indicadores bibliométricos. *Revista Española de Cardiología, [Revista electrónica]*.

Vol. 52. Núm. (19), pp. 790-800., 1999. https://doi.org/10.1016/S0300-8932(99)75008-6

[25] F. Ruiz, Estudio Bibliométrico sobre la Producción Científica en la UCLM. España: Departamento de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha., 2012. Disponible en: http://alarcos.inf-cr.uclm.es/per/fruiz/wos-uclm/

[26] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD-96 Proceedings, [Revista electrónica].* pp. 82-88. https://www.aaai.org/Papers/ KDD/1996/ KDD96-014.pdf

[27] L. Lebart, A. Salem y M. Bécue, *Análisis estadístico de datos y textos.* España: Milenio, 2000.

[28] M. Páramo, Significados de adolescencia y psicoterapia: análisis lexicométrico de discursos grupales. *Acta Colombiana de Psicología, [Revista electrónica].* Vol. 13 Núm. (2)., 2010. Disponible en: https:// actacolombianapsicologia.ucatolica.edu.co/article/view/377

[29] D. Hernández. "*BiblioMineR: una herramienta estadística para la revisión bibliográfica"* (Tesis de Maestría). Universidad Politécnica de Catalunya, Barcelona, España, 2012. http://hdl.handle. net/2099.1/16500

[30] B. Lee y Y. Jeong, Mapping Korea's national R&D domain of robot technology by using the co-word analysis. *Scientometrics, [Revista electrónica].* Vol. 77 Núm. (1), 3-19., 2008. https://doi.org/10.1007/ s11192-007-1819-4

[31] R. Caraguay. "*Aplicación de técnicas de minería de texto para el agrupamiento de componentes académicos en base a los contenidos de planes docentes".* (Trabajo de pregrado). Universidad Técnica Particular de Loja, Ecuador, 2016.

[32] P. Murphy, Basic Text Mining in R. [Online]. 2017. Disponible en : https://rpubs.com/pjmurphy/265713

[33] M. Bécue Bertaut. Mesurem els mots ? L ' anàlisi estadística de textos. En A. Ventura (Presidencia), I Jornades de Llengua i Estadística. Ponencia llevada a cabo en la de Lengua- Estadística de l'Idescat, Barcelona, España, 2009.

[34] M. Vallez, y R. Pedraza Jiménez, El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. *Hipertext.net, [Revista electrónica].* Núm. 5, 2007. Disponible en: https://www.upf.edu/hipertextnet/numero-5/pln.html

[35] G. Salton, "*Automatic text processing: the transformation, analysis, and retrieval of information by computer."* Boston: Addison-Wesley Longman Publishing Co, 1989.

[36] L. Oliva, M. Hernández, y C. Castro, Más que palabras nos dicen los adolescentes que desean migrar. Estudio estadístico de las respuestas a una pregunta. *Revista de Psicología Social Aplicada, [Revista electrónica],* Vol. 1 Núm. (1), pp. 55–73., 2012. Disponible en: https://revistas.innovacionumh.es/index.php/lexmercatoria/article/view/892

[37] A. Jain, M. Murty, y P. Flynn, Dataclustering: areview. *ACMcomputingsurveys(CSUR*), [Revista *electrónica], Vol.* 31 Núm. (3), pp. 264-323, 1999. Disponible en: https://dl.acm.org/doi/10.1145/331499.331504

[38] F. Wickelmaier, *"An Introduction to MDS."* Dinamarca: Aalborg University, 2003.

[39] I. Romero-Pérez, Y. Alarcón-Vásquez, y J. García-Jiménez, Lexicométrica: Enfoque aplicado a la redefinición de conceptos e identificación de unidades temáticas.

Biblios, [Revista electrónica] Núm. 71., 2018. Disponible en: http://biblios.pitt.edu/ojs/index.php/biblios/article/view/466/0

[40] I. Romero-Pérez, y E. Latorre-Iglesias, Listado de documentos seleccionados para el análisis y visualización de las tendencias y dinámicas de producción del grupo de investigación Joaquín Aaron Manjarrés (2011-2017)., 2018. Figshare. *Journal contribution. [Revista electrónica]*. https://doi.org/10.6084/m9.figshare.5900227.v1