Universidad Central “Marta Abreu” de Las Villas

Facultad de Matemática, Física y Computación

Trabajo para optar por el Título Académico

Máster en Ciencia de la Computación

AGRUPAMIENTO DE ARTÍCULOS CIENTÍFICOS BASADO EN LA EXTRACCIÓN DE FRASES RELEVANTES

Autor:

Lisvandy Amador Penichet

Tutores:

Dra. María Matilde García Lorenzo

Dr. Damny Magdaleno Guevara

2018

Hago constar que el presente trabajo fue realizado en la Universidad Central “Marta Abreu” de Las Villas como parte de la culminación de los estudios de la maestría Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del tutor

Firma del jefe del Seminario de Inteligencia Artificial

PENSAMIENTO

Son tus decisiones y no tus condiciones las que determinan tu destino.

Anthony Robbins.

AGRADECIMIENTOS

A Dios.

A mi mamá y a mi hermana por apoyarme en todas las decisiones que tomo.

A Odle por confiar ciegamente en mi capacidad de hacer lo que me proponga, porque ahora estamos lejos pero siempre vamos a ser inseparables, porque se alegra de mis éxitos como si fueran los de ella.

A mi papá, mi abuela Célida y mi tía Mary por siempre estar pendientes del avance de este trabajo.

A mis compañeras del IBP Amandita, Novisel, Yelenys, Marta y todos los demás, por su gran amistad y por contribuir como nadie a mi formación como profesional de la ciencia.

A Amanda y Ernesto porque a pesar de que los veo poco sé que siempre puedo contar con ellos.

A Mario, Gio, Elier y Heikel porque en el momento que he necesitado de ellos han estado ahí sin pensarlo dos veces.

A Neysi porque a pesar de estar lejos seguimos siendo los mejores amigos.

A Deysi, que la convertí a la fuerza en mi madrina adoptiva.

Al Dr Amed Leiva Mederos por su inmensa confianza en la pertinencia de esta investigación.

A mis tutores Marilyn y Damny por nuevamente hacerse cargo de mí.

A Ivan, por la paciencia que tiene conmigo, por su confianza y por hacer cualquier cosa por hacerme feliz.

RESUMEN

La gestión del conocimiento a partir de la información recogida en la bibliografía científica resulta imprescindible para los investigadores en función de optimizar el tiempo de que disponen. El agrupamiento automático de datos se perfila como una de las técnicas que facilitan este proceso. Este permite formar grupos de documentos afines a partir de una colección obtenida mediante un proceso de recuperación de información. Este trabajo tuvo como objetivo: desarrollar un método de agrupamiento de artículos científicos a través de la extracción de frases relevantes obtenidas de los títulos y de las referencias bibliográficas para mejorar la gestión del conocimiento a partir de la literatura científica. Para formar los grupos se tomaron como centroides las frases relevantes contenidas en la intersección de los títulos de los artículos con los títulos de las referencias. Además, se creó un grafo de conexiones de los artículos basado en las frases relevantes que comparten. Mediante este grafo se eliminó el solapamiento entre grupos y se asignaron a los grupos los documentos que no contenían las palabras centroides. Para evaluar los resultados del método propuesto se utilizaron siete medidas externas de calidad del agrupamiento. Como casos de estudios fueron usados artículos científicos provenientes de diferentes áreas del conocimiento. Los experimentos realizados demostraron la factibilidad del método propuesto en el agrupamiento de artículos científicos.

ABSTRACT

The knowledge's management from the information collected in the scientific literature is essential for researchers in order to optimize the available time. The automatic data clustering is emerging as a technique that facilitate this process. This allows forming groups of related documents from a collection obtained through a process of information retrieval. The objective of this work was, to develop a method of scientific articles clustering through the extraction of relevant phrases, obtained from titles and bibliographic references to improve knowledge management based on scientific literature. To form the groups, the relevant phrases contained in the intersection of the articles' titles with the titles of the references were taken as centroids. In addition, a connection graph of the articles was created based on the relevant phrases they share. Through this graph the overlap between groups was eliminated and the documents that did not contain the centroid words were assigned to the groups. To evaluate the results of the proposed method, seven external measures of clustering quality were used. As case studies were used scientific articles from different knowledge's areas. The experiments carried out demonstrated the feasibility of the proposed method in the clustering of scientific articles.

TABLA DE CONTENIDOS

INTRODUCCIÓN

Los volúmenes de información disponibles a nivel mundial crecen a diario y las colecciones de datos se vuelven cada vez más heterogéneas, grandes, diversas y dinámicas (Magdaleno Guevara et al., 2016); por lo que es más complejo para los usuarios identificar la información relevante (Aljaber et al., 2010).

Uno de los principales métodos usados para la gestión del conocimiento es el agrupamiento de datos (Qian and Zhang, 2003). El problema del agrupamiento consiste en encontrar grupos de objetos similares en un conjunto de datos, donde la similaridad entre un par de objetos se calcula usando una función de similitud (Aggarwal and Zhai, 2012).

Específicamente, el agrupamiento de artículos científicos se torna una tarea de suma importancia; ya que es necesario dotar a los investigadores de herramientas capaces de agilizar el proceso de identificación de la información relevante y de esta manera puedan hacer un uso más eficiente del tiempo que disponen.

Los artículos científicos tienen una estructura bien definida (título, autores, palabras claves, resumen, contenido, referencias bibliográficas). Esto facilita que en muchos casos no sea necesario hacer uso de todo el artículo para lograr un buen resultado al aplicar algún método de agrupamiento.

Explotar de manera correcta esta estructura, en función de identificar cuáles de estas partes resultan más significativas cuando se desea saber qué tan similares son dos artículos, puede contribuir a incrementar la eficiencia en el agrupamiento de artículos científicos, dado que se reduce considerablemente el tiempo computacional al no tener que procesar todo el documento. Al mismo tiempo se puede incrementar la eficacia, debido a que la extracción de términos se focaliza en partes del artículo que brindan una información más detallada y precisa del mismo.

Lo antes expuesto es una problemática que la ciencia aún no aborda de manera completa y justifica el siguiente **planteamiento de investigación**:

Muchos trabajos reportados en la literatura usan partes específicas del artículo para el agrupamiento de los mismos. Estos se enfocan, principalmente, en las referencias bibliográficas (Aljaber et al., 2010, Garfield et al., 2013, Small, 1973, West et al., 2016) específicamente en el análisis de la co-citación de los artículos haciendo uso del *ScienceCitationIndex* (SCI). Otros trabajos están direccionados a la comparación de los artículos basados en las frases relevantes que ellos comparten, pero sin hacer énfasis en ninguna parte específica del artículo (Kumar and Srinathan, 2008, Kim et al., 2013, Lebret and Collobert, 2014). Sin embargo, la extracción de frases relevantes en los títulos de las referencias bibliográficas, así como en los títulos de los artículos en cuestión, ha sido poco estudiada y no se han encontrado en la literatura métodos de agrupamiento que hagan uso específico de este tipo de información.

La **hipótesis** de investigación es la siguiente:

La extracción de frases relevantes en partes representativas del artículo científico que no son dimensionalmente complejas, aumenta la eficiencia y eficacia en el agrupamiento de este tipo de documento.

Por lo antes mencionado se propone como **objetivo general** de esta investigación:

Desarrollar un método de agrupamiento de artículos científicos a través de la extracción de frases relevantes obtenidas de los títulos y de las referencias bibliográficas para facilitar la gestión del conocimiento a partir de la literatura científica.

Para dar cumplimiento al objetivo general se plantean los siguientes **objetivos específicos**:

1- Realizar un análisis crítico de métodos de agrupamiento de artículos científicos, así como métodos para la extracción de frases relevantes.
2- Implementar un algoritmo de extracción de frases relevantes en las referencias bibliográficas y en los títulos de los artículos científicos.
3- Desarrollar un método de agrupamiento de artículos científicos basado en las frases relevantes.
4- Evaluar el método de agrupamiento propuesto a partir de conjunto de corpus de artículos científicos previamente etiquetados.

Las **preguntas de investigación** planteadas son:

1- ¿Cómo extraer de los títulos y de las referencias bibliográficas frases relevantes que sean capaces de englobar temas específicos?
2- ¿Qué método de agrupamiento utilizar para conformar grupos de artículos basados en la comparación de las frases extraídas?
3- ¿Cómo evaluar el método propuesto?

El **valor teórico** de la investigación está directamente vinculado con su novedad científica.

El **valor práctico** del trabajo está enfocado a:

Disponer de un algoritmo de agrupamiento, que permita procesar grandes volúmenes de artículos científicos y obtener conocimiento relevante a partir de la información recuperada, con el propósito de facilitar a los investigadores y docentes el inicio de una revisión del estado del arte, organizar por temáticas los artículos que han sido recopilados por el comité científico de un evento, así como tener una idea de las asociaciones que existen entre los documentos recuperados.

Para la presentación de esta investigación, esta tesis se estructuró de la forma siguiente: una Introducción, donde en lo esencial se caracteriza la situación problémica y se fundamenta el problema científico a resolver, así como la estrategia general seguida para su solución como problema científico. Un Capítulo 1, que contiene el marco teórico-referencial que sustentó esta investigación. En el siguiente capítulo se resume y explica la forma en que se realizó la extracción de términos y representación de los documentos, así como el método de agrupamiento desarrollado. En el tercer capítulo se evalúa el método propuesto y se demuestra la factibilidad del uso del mismo para el agrupamiento de artículos científicos. Las Conclusiones y Recomendaciones derivadas de la investigación realizada, la Bibliografía consultada y un grupo de Anexos como complemento de los resultados expuestos.

1

MÉTODOS DE AGRUPAMIENTO DE ARTÍCULOS CIENTÍFICOS

Capítulo 1. Métodos de agrupamiento de artículos científicos

En este capítulo se describe una panorámica general de los algoritmos de agrupamiento, haciendo énfasis en los métodos de agrupamiento de documentos y particularmente en aquello diseñados para el agrupamiento de artículos científicos. Se abordan los temas relacionados con la representación de documentos donde se hace particular hincapié en el uso de frases relevantes para la representación de artículos científicos. Por último, se hace referencia a las principales medidas de similitud y distancia para la comparación de documentos, así como las principales métricas para la evaluación de la calidad de un proceso de agrupamiento.

1.1 Algoritmos de agrupamiento

Existen en la literatura varias definiciones sobre el agrupamiento de datos (Forsati et al., 2013, Nasir et al., 2013, Monroy Medina, 2016, Tineo et al., 2016) pero todas ellas llegan a un punto de convergencia donde se define el agrupamiento como un proceso en el cual un conjunto de objetos de tamaño n es dividido en k grupos ($k \leq n$) donde se trata que los elementos que pertenecen a un mismos grupos sean altamente similares entre sí y los elementos que pertenecen a grupos diferentes sean lo más disímiles posible.

Una gran cantidad de algoritmos de agrupamiento aparecen en la literatura científica (Guha et al., 1998, Bezdek et al., 1984, Rajeshwari et al., 2015, Sert et al., 2015). Según (Magdaleno Guevara et al., 2015) estos se pueden clasificar siguiendo diversos criterios, como pueden ser: tipo de los datos de entrada, criterios para definir la similitud entre los objetos, conceptos en los cuales se basa el análisis y forma de representación de los datos.

Si la participación del usuario influye en el agrupamiento, se tienen otras dos clasificaciones: algoritmos de agrupamiento automático y algoritmos de agrupamiento semiautomático.

El agrupamiento de datos se convirtió en un área de investigación cada vez más importante y necesaria debido al desarrollo vertiginoso que experimentan los dominios de aplicación modernos, como la *Word Wide Web* (Forsati et al., 2013). Es por ello, que

en la literatura científica se pueden encontrar diversos algoritmos diseñados para este propósito. Algunos de ellos se mencionan a continuación.

Cuando se habla de agrupamiento de datos uno de los algoritmos pioneros es el QMODEL (Miesch, 1976), el cual, a pesar de presentar cierta inestabilidad, sobre todo manifestada en conjuntos de datos que presentan valores alejados de la media, sirvió de punto de partida para el desarrollo de nuevos algoritmos que trataran de corregir las principales deficiencias del mismo. Así surge, por ejemplo, el algoritmo *Fuzzy c-Means* (Bezdek et al., 1984) el cual se basa en la teoría de los conjuntos borrosos y reduce los problemas que se presentaban con el algoritmo QMODEL y los valores extremos.

En (Hartuv and Shamir, 2000) se presenta un algoritmo de agrupamiento basado en la teoría de grafos. Los autores consideran cada vértice del grafo como uno de los objetos a agrupar y las aristas que conectan a los vértices contienen la similitud que existe entre cada par de vértices conectado. Luego, los grupos que se obtienen al aplicar el algoritmo son los subgrafos que presentan un alto valor de conectividad. Una de las ventajas de este algoritmo es que no se necesita proporcionar la cantidad de grupos como parámetro de entrada, ya que el propio algoritmo en el proceso iterativo es capaz de determinar los grupos que conforman la colección. Según los autores, las pruebas aplicadas a este algoritmo, demuestran que obtiene resultados favorables manifestados en la homogeneidad de los grupos.

Otro de los algoritmos de agrupamiento que se puede encontrar en la literatura es el G-Star, propuesto por Suárez y Pagola en (Suárez and Pagola, 2007). Este es otro de los algoritmos basado en la teoría de grafos en el cual se considera cada objeto a agrupar como uno de los vértices del grafo. La idea general es seleccionar como centros potenciales aquellos vértices que tienen una mayor incidencia de otros vértices sobre ellos y a partir de ahí, construir los grupos.

Uno de los algoritmos más populares cuando se habla de agrupamiento es el *k-Means*. Este es uno de los algoritmos particionales más usado (Celebi et al., 2013). La idea general que sigue es representar cada clúster por la media de sus puntos, o lo que es lo mismo, por su centroide. Para ello selecciona inicialmente *c* centroides de manera aleatoria y asigna cada uno de los objetos a agrupar al centroide con respecto al cual

tenga mayor similaridad. Iterativamente recalcula los centroides basado en los clústeres formados y reasigna los documentos. El algoritmo se detiene cuando se satisface la condición definida como criterio de parada o se alcanza el máximo número de iteraciones (Nasir et al., 2013). La mayor desventaja de este algoritmo es que se necesita proporcionar el número de grupos que se espera obtener.

La calidad de las soluciones que ofrece el método *k-Means* depende en gran medida de la posición en que se encuentran los centroides seleccionados inicialmente. Se han desarrollado algunos métodos que intentan disminuir la dependencia de este método a la selección de los centroides iniciales (Likas et al., 2003, Banerjee and Ghosh, 2004, Arthur and Vassilvitskii, 2007, Wang et al., 2010, Tzortzis and Likas, 2014).

Una de las primeras modificaciones al algoritmo *k-Means*, con el objetivo de disminuir las deficiencias del mismo asociadas a la selección de los centroides, fue el método *global k-Means* propuesto por Likas y colaboradores en (Likas et al., 2003). La idea del algoritmo es que la solución para un problema de agrupamiento con M grupos se puede obtener mediante una serie de búsquedas locales. Por lo cual el algoritmo parte de resolver los problemas para 1, 2, …, M-1 clústeres, coloca en cada una de estas iteraciones los centroides para cada grupo en la posición óptima y llega por tanto a la iteración M con la posición óptima de los centroides iniciales, ya conocida. La principal desventaja de este algoritmo radica en que, al ser un método determinístico, la complejidad computacional aumenta considerablemente. Es por esta razón, que los autores introdujeron modificaciones al mismo con el objetivo de lograr soluciones más eficientes sin afectar la eficacia del método.

El método *MinMax k-Means* propuesto por Tzortzis y Likas en (Tzortzis and Likas, 2014) se presenta como uno de los más eficientes a la hora de disminuir los problemas asociados a la selección de los centroides iniciales del método *k-Means*. Los autores asignan un peso a cada uno de los clústeres, el cual es proporcional a la varianza de los mismos. Además, se optimiza una versión pesada del objetivo del *k-Means* para restringir en la solución, la aparición de clústeres que tengan altos valores de varianza. Este método logra mejorar los resultados del agrupamiento en donde la selección inicial de los centroides por el método *k-Means* no fue buena.

Cada una de las variantes implementadas al método *k-Means* ofrecen versiones del mismo que reducen la complejidad computacional en el proceso de agrupamiento, y han traído consigo que se obtenga lo que pudiera llamarse la familia de algoritmos *k-Means*. Esta familia está compuesta por una amplia gama de algoritmos donde se pueden mencionar los algoritmos Fuzzy *k-Means* con sus diferentes variantes, los algoritmos *k-Means* basados en metaheurísticas y otros.

En (Pinto et al., 2010) se presenta una variante del algoritmo de agrupamiento *k-Star*. Este, como la mayoría de los algoritmos de agrupamiento, requiere como parámetro una matriz de similitud que recoja el grado de semejanza entre cada par de objetos de la colección. La principal ventaja de este algoritmo radica en su capacidad, durante el proceso iterativo, de descubrir automáticamente la cantidad de grupos que se deben formar.

Los algoritmos anteriormente referidos son algoritmos de agrupamiento usados para obtener grupos en diferentes contextos de aplicación, como pueden ser: manejo de información geográfica (Castro et al., 2014), gestión de personal (Malbernat et al., 2015), clasificación de datos económicos (López et al., 2016), agrupamiento de textos (Casillas García, 2015) entre otros. Se han desarrollado algunos métodos de agrupamiento que son especialmente útiles para contextos específicos. Para el caso particular de la minería de texto, son varios los trabajos que se pueden encontrar, algunos de ellos se refieren en el epígrafe siguiente.

1.2 Agrupamiento de documentos

Cuando se habla específicamente de agrupamiento de documentos se encuentran en la literatura varios métodos diseñados con este propósito. Algunos de ellos fueron concebidos para trabajar con tipos de documentos específicos como pueden ser: artículos periodísticos, textos contenidos en correos electrónicos, documentos que tienen naturaleza política, artículos científicos y otros. Sin embargo, muchos de ellos fueron pensados para trabajar con colecciones de documentos de naturaleza no específica, por lo cual pueden ser usados para cualquier tipo de documentos.

Entre los algoritmos diseñados para el agrupamiento de documentos se tiene el algoritmo Autoclass (Neto et al., 2000). Este es un algoritmo determinista, no requiere especificar el número de grupos a obtener y además de obtener los grupos de los documentos, obtiene las palabras claves de cada grupo.

En (Gil-García and Pons-Porrata, 2010), los autores desarrollan un algoritmo jerárquico dinámico para el agrupamiento de documentos. Este algoritmo representa la colección a agrupar mediante un grafo de $ß_0$-semejanza, donde cada vértice representa un grupo, por lo cual se parte de un grafo de *n* vértices, donde *n* es la cantidad de objetos a agrupar. Dos vértices estarán conectados únicamente si su semejanza supera un umbral definido. Luego se aplican sucesivas transformaciones al grafo a través de un algoritmo de cubrimiento hasta que se obtiene un grafo $ß_0$-semejante completamente inconexo. Según (Domínguez et al., 2014), este algoritmo obtiene buenos resultados, pero en colecciones con un elevado número de objetos consume gran cantidad de memoria, lo que reduce la cantidad máxima de documentos a agrupar.

Un enfoque muy interesante en el agrupamiento de texto es el que se presenta en (Iezzi, 2012). Los autores desarrollan un nuevo método de agrupamiento de texto considerando el enlace entre los términos y los documentos. Se propone el uso de algunas medidas de importancia de los documentos, entre ellas: *degree centrality*, *closeness centrality*, *eigenvector index*, *betweenness centrality*. Para eliminar la correlación entre las medidas aplicadas los autores recurren al análisis de componentes principales. Luego de este paso se aplica una medida de distancia que permite obtener una matriz de dimensión (*n*x*n*) la cual recoge el grado de disimilitud entre cada par de documentos. Con el enfoque propuesto se puede identificar tanto el rol de los términos como de los documentos en un corpus determinado. Esto permite identificar documentos relevantes basado en la interrelación entre ellos.

El uso de metaheurísticas para incrementar la eficiencia en el agrupamiento de grandes colecciones de documentos también ha sido trabajado. Así se pueden encontrar trabajos como (Forsati et al., 2013) donde se presenta un nuevo enfoque para el agrupamiento de documentos basado en la meta heurística búsqueda armónica. Los autores proponen primeramente un método de agrupamiento puro, basado únicamente en la búsqueda

armónica, el cual, encuentra clústeres casi óptimos en un tiempo computacional razonable. Luego el agrupamiento basado en la búsqueda armónica es integrado de tres formas diferentes con el algoritmo *k-Means*. Esto tiene como objetivo obtener clústeres más compactos, aprovechando la capacidad de exploración de la búsqueda armónica y la potencialidad en el refinamiento del algoritmo *k-Means*. Los experimentos realizados demuestran que los métodos híbridos propuestos obtienen mejores resultados que: el método *k-Means* puro, el agrupamiento basado en algoritmo genético y el agrupamiento basado únicamente en la búsqueda armónica.

También algunos autores han trabajado en la optimización de algoritmos de agrupamiento a través de la paralelización de los mismos, de manera tal que métodos que funcionan eficazmente pero que reducen considerablemente su eficiencia con el aumento del tamaño de las colecciones a agrupar, puedan mejorar y utilizarse en este tipo de contextos. Un trabajo con este enfoque se presenta por González-Soler y colaboradores en (González-Soler et al., 2015) en el cual los autores presentan una modificación del algoritmo DClustR desarrollado por (Pérez-Suárez et al., 2013). En esta versión, denominada CUDA-DClust, se paraleliza el trabajo del algoritmo DClustR, lo que permite reducir las deficiencias del mismo asociadas al alto consumo de memoria en grandes colecciones de documentos y al tiempo computacional asociado al procesamiento de las mismas.

El agrupamiento de textos es un método no supervisado por lo cual no existen clases o etiquetas. De este modo el método de agrupamiento debe descubrir la relación entre los documentos y basado en esta relación construir los grupos (Bernotas et al., 2015). Se hace necesario disponer de una representación de los documentos para establecer una comparación entre los mismos, basado en un criterio determinado, así como también disponer de una función capaz de discernir si dos documentos pertenecen a un mismo grupo o no.

1.2.1 Modelo espacio vectorial para la representación de documentos

Una de las representaciones de preferencia por los autores de la minería de texto es el Modelo Espacio Vectorial (VSM, por sus siglas en inglés). En (Lanquillon, 2001) se define el Modelo Espacio Vectorial como:

Sea $d \in D$ un documento textual. La representación de d es el vector documento $d=\rho(d)=(w_1,\ldots,w_m)^T \in R=R^m_+$, donde cada dimensión corresponde a un término en la colección de documentos y w_i denota el peso del *i*-ésimo término. El conjunto de esos *m* términos indexados, $V=(t_1, \ldots, t_m)$ es referido como el vocabulario.

La representación de un documento mediante el VSM lleva consigo cuatro pasos fundamentales, los cuales según (Arco García, 2005) son:

1. Transformación del corpus
2. Extracción de términos
3. Reducción de dimensionalidad
4. Normalización y pesado de la representación

La aplicación de estos cuatro pasos conlleva a la representación de un documento mediante los términos que más se repiten en el mismo siempre que estos pertenezcan al vocabulario que se construyó. De este modo, se parte del conjunto de términos del documento y mediante la transformación del corpus se obtiene una representación homogénea de los términos, para ello se aplican varias reglas, entre ellas: convertir todas las letras a minúscula o mayúscula, eliminar los signos de puntuación, sustituir las contracciones por sus palabras originales, eliminar los caracteres no alfa-numéricos y otras.

En el siguiente paso, para cada documento, se construye un vector de frecuencia donde a cada término se le hace corresponder su frecuencia de aparición, de manera tal que en el paso subsecuente se puede reducir la cantidad de términos a través de la selección de aquellos que resultan ser los más representativos. Este es precisamente el objetivo de la reducción de la dimensionalidad, lograr una representación más precisa de los documentos, donde los términos que no aportan información relevante como son los artículos, conjunciones, preposiciones y pronombres, conocidos como palabras de parada, se eliminan. También se elimina aquellos términos cuya frecuencia de aparición es baja y por tanto no se consideran representativos del documento. Es válido aclarar que reducir la dimensionalidad tiene mayor peso en el caso que no esté construido el vocabulario representativo de la colección. De otro modo basta con seleccionar para cada documento los términos que pertenecen al vocabulario y desechar aquellos que no

pertenecen. Por último, se tiene el paso de normalización y pesado de la matriz el cual consiste en generar para cada documento *d* un vector de peso basado en el vector de frecuencia, donde a cada término *t* de *d* se le va a asignar un peso que representa la importancia que tiene *t* en dicho documento con respecto a su frecuencia en todos los documentos. Para la normalización y pesado de la matriz se pueden encontrar en la literatura disimiles variantes (Manning and Schütze, 1999, Sebastiani, 1999, Berry and Castellanos, 2004). Una de las más usadas es la TF-IDF (Huang, 2008, Yan et al., 2013, Costa and Ortale, 2018). Esta se basa en la idea de que los mejores términos que representan un documento d_i son aquellos que tienen una alta frecuencia de aparición en el mismo, y no se encuentran en el resto de los documentos o se encuentran en una pequeña parte de los mismos, por lo cual estos deben tener un mayor peso, ver ecuación 1.1. Matemáticamente se define como:

$$TFIDF(d,t) = tf(d,t) * \log\left(\frac{|D|}{df(t)}\right) \tag{1.1}$$

Donde $tf(d,t)$ es la frecuencia absoluta del término *t* en el documento *d*, $df(t)$ es la cantidad de documentos en los que aparece el término *t* y $|D|$ es la cantidad total de documentos de la colección.

El Modelo Espacio Vectorial (VSM) representa el significado de los elementos léxicos como vectores en un espacio semántico (Kiela and Clark, 2014). La gran ventaja de esta representación es que se reduce la dimensionalidad del documento ya que este se representa solo por sus términos relevantes. En algunos trabajos como (Hotho et al., 2002, Hotho et al., 2003, Bloehdorn et al., 2005) se critica el VSM haciendo hincapié en que esta representación, utilizada para los métodos de agrupamiento, resulta a menudo insatisfactoria, ya que ignora las relaciones entre los términos importantes que no co-ocurren literalmente. Esta afirmación es correcta debido a que el VSM considera la frecuencia de aparición de los términos, pero no tiene en cuenta la sinonimia y los términos homónimos lo cual puede provocar ruido cuando se desea determinar cuán similares son dos documentos. En (Nasir et al., 2013) se propone una extensión del modelo espacio vectorial para solucionar el problema anteriormente mencionado, para ello los autores aplican el suavizado semántico basado en la sabiduría lingüística,

sabiduría de las multitudes, sabiduría de corpora. Los resultados obtenidos al aplicar esta propuesta demuestran que el agrupamiento mejora significativamente comparado con la utilización del VSM tradicional.También existen trabajos como el de (Cobo et al., 2006) donde se analizan variantes de técnicas de representación de los datos tomando como base el Modelo Espacio Vectorial y donde se hace la extracción de términos usando diferentes técnicas como el Análisis de Semántica Latente, la Factorización en Matrices No Negativas y el Análisis en Componentes Independientes.

La introducción de modificaciones al VSM favorece el aumento de la eficacia en los resultados del proceso de gestión del conocimiento que se desea aplicar. Las principales modificaciones a este modelo se concentran en la extracción de términos, ya que son precisamente estos los que representan al documento.

1.2.1.1 Extracción de frases relevantes para la representación de documentos

Entre las modificaciones aplicables para aumentar la eficacia en la representación de los documentos se encuentra la extracción de frases relevantes (se usa indistintamente el término frases relevantes o palabras claves). Esta consiste en extraer no solo los términos independientes que tienen mayor frecuencia de aparición, si no extraer aquellas combinaciones de términos que representan de manera más precisa el tema general que trata el documento. En la literatura se reportan varios trabajos que se enfocan en esta línea de investigación (Srikant and Agrawal, 1996, Zaki, 2001, Pei et al., 2001, Yang et al., 2007, Kumar and Srinathan, 2008, Kim et al., 2013, Lebret and Collobert, 2014), algunos de ellos se describen a continuación.

El algoritmo GSP (Generalized Sequential Patterns), creado por Agrawal y Srikant en (Srikant and Agrawal, 1996), se encuentra entre los algoritmos más representativos dentro de la minería de secuencias frecuentes (Hernández, 2016). En este algoritmo se seleccionan primeramente las palabras de tamaño uno que superan el umbral definido para considerar una palabra como frecuente. A partir de estas palabras se conforman las frases de tamaño dos y se analiza cuáles de ellas superan el umbral. Esto conlleva a recorrer nuevamente el texto, para determinar la frecuencia de cada una de las frases formadas. De esta manera, se desechan las frases que no superen el umbral, lo que permite reducir la cantidad de nuevas frases que se van a formar a partir de las frases que

se tienen como frecuentes. Este método garantiza que se obtengan las frases más frecuentes, pero la eficiencia del mismo se reduce considerablemente cuando aumenta el tamaño del texto, debido a que es necesario recorrer el texto cada vez que se forman nuevas frases a partir de las que se tiene.

Otro de los trabajos basado en la técnica de extracción de frases relevantes es el desarrollado por Kumar y Srinathan en (Kumar and Srinathan, 2008). Este tuvo como objetivo fundamental la extracción de frases representativas en artículos científicos mediante el uso de técnicas de selección y filtrado de n-gramas.

Los autores se enfocaron en la extracción de n-gramas de tamaño *n* con $1 \leq n \leq 4$. Para ello inicialmente crean, para cada documento, un diccionario que contiene todos los términos independientes distintos presentes en el documento, los que representan los n-gramas de tamaño uno. Luego se procesa cada oración del documento para formar los n-gramas de tamaño mayor igual a dos. Estos n-gramas son separados en listas dependiendo de su tamaño. De este modo los n-gramas de tamaño uno, estarán en una lista, los de tamaño dos en otra y así sucesivamente. Luego se analiza la frecuencia total de ocurrencia de cada uno de los n-gramas, pero no de todos, solo la frecuencia de aquellos que cumple que:

$$\left(0 \leq P_0 < \frac{N_i}{2}\right) \; o \; \left(P_0 > \frac{3 * N_i}{4}\right)$$

Donde N_i es la cantidad de palabras de la oración *i* y P_0 es el índice donde comienza el n-grama P en la oración *i*, tomando como incremento de índice el cambio de una palabra a otra y no de un carácter a otro.

Este criterio de discriminación, según los autores, se basó en varias observaciones realizadas y en algunos trabajos que demuestran que las frases sustantivas mayoritariamente se presentan al inicio y final de las oraciones.

Luego se calcula el peso de cada uno de los n-gramas seleccionados en el paso anterior, el cual está dado por la posición de la primera ocurrencia del n-grama y por la cantidad de veces que él ocurre en el documento. Mediante el uso de este peso es que los autores seleccionan las frases representativas del documento. Para ello ordenan

descendentemente de acuerdo al peso de los n-gramas cada una de las cuatro listas. Seleccionan el 10% del total de frases que se desean encontrar de la lista de n-gramas de tamaño cuatro, el 20% de la lista de n-gramas de tamaño tres y el 70% restante de las listas de tamaño dos y tamaño uno. Estos elementos seleccionados constituyen las frases representativas del documento.

La mayor limitante de este método de selección de frases relevantes está dada por la alta complejidad computacional que presenta. Debido a que se tienen en cuenta todos los n-gramas de tamaños entre uno y cuatro gramas presentes en todo el documento, es necesario recorrer el documento completo dos veces, primero para realizar la selección de términos y luego para la conformación de los n-gramas y el cálculo de su peso.

En (Kim et al., 2013) se analiza la eficacia de 19 métodos diseñados para la extracción de palabras claves en artículos científicos, los cuales fueron propuestos para la *Workshop on Semantic Evaluation 2010*. Los autores realizan una comparación entre: las palabras extraídas por los sistemas sometidos a evaluación, las palabras propuestas por los autores de los artículos y las palabras seleccionadas por un conjunto de estudiantes a los cuales le fue asignada la lectura de los artículos tomados como casos de estudio. Para esto, se determina primeramente el cubrimiento máximo que pueden tener los métodos de extracción de términos que se evalúan, dado que algunas de las palabras claves propuestas por los autores de los artículos y también algunas de las seleccionadas por los estudiantes no ocurren literalmente en el texto de los artículos. Se llega a la conclusión que el cubrimiento de las palabras propuestas por los autores es de un 85% mientras que el de las propuestas por los estudiantes es de un 81%. Sobre esta conclusión se analizó cuál de los métodos propuestos logró los mejores resultados, basado en la medida *F-Measure* obtenida para cada uno de ellos. Se tomaron para el análisis los 15 primeros candidatos a palabras claves determinados por cada método.

El método propuesto por el equipo HUMB (Romary, 2010) obtuvo los mejores resultados. Este método se basa en la selección de n-gramas ($1 \leq n \leq 5$). Para la selección de candidatos y rankeo de los mismos, utilizan un árbol de decisión basado en varios rasgos que incluye: la estructura del documento, el contenido y un puntaje léxico-semántico a partir de términos base. Finalmente, los candidatos son sometidos a un

nuevo proceso de ranking mediante el uso de un modelo probabilístico basado en las palabras claves asignadas por autores en una colección independiente.

Otro trabajo interesante donde se considera la representación de documentos a través de n-gramas es el desarrollado por Lebret y Collobert en (Lebret and Collobert, 2014). Los autores plantean que en la representación de documentos mediante términos independientes se pierde información semántica, y es muy probable que los términos con las mismas frecuencias en los mismos documentos estén semánticamente relacionados. Sobre la base de estas consideraciones proponen un modelo para la representación de documentos mediante conceptos semánticos. Para ello, a partir de una colección de documentos usada como conjunto de entrenamiento, aplican tres pasos fundamentales: obtener para cada documento un vector de n-gramas, agrupar los n-gramas en *k* grupos, representar cada nuevo documento como un vector de *k* conceptos semánticos, donde un documento d_i contiene el concepto *k* si alguno de los n-gramas seleccionado en este documento está presente en el grupo *i*. Esta forma de representación de documentos reduce, tanto la pérdida de información asociada a la representación de documentos mediante términos independientes, como la dimensionalidad de la representación.

También se pueden mencionar otros trabajos donde la extracción de frases relevantes no es usada para representar documentos sometidos a procesos de clasificación o agrupamiento, si no para la recomendación de documentos. Tal es el caso del trabajo desarrollado por Habibi y Popescu-Belis en (Habibi and Popescu-Belis, 2015). En este los autores extraen palabras claves a partir de conversaciones reales, estas palabras son usadas para conformar consultas que permitan la búsqueda en internet de documentos relacionados con el tema de conversación analizado. De esta manera, se pueden recomendar a los científicos, bibliografías de interés a partir de conversaciones que estos hayan tenido, las conversaciones pueden ser, por ejemplo, presentaciones en conferencias e intercambios con otros científicos. Los autores plantean que es preferible el uso de palabras claves que la extracción de términos independientes, dado que en las conversaciones se pueden encontrar muchos términos que no tiene relación alguna con la idea general de lo que se está hablando lo que hace que aumente el ruido en la recuperación de información.

Es válido aclarar que la representación de documentos mediante la extracción de frases relevantes no es útil para cualquier tipo de documentos. Su uso es aconsejable en documentos donde su contenido está relacionado, de manera general, con un tema determinado, como por ejemplo los artículos científicos, tesis de pregrado o posgrado, textos periodísticos, etc.

Independientemente del modelo que se use para representar los documentos que se desean agrupar, se hace necesario disponer de una medida que permita establecer una comparación entre dichos documentos de forma tal que se pueda discernir si dos documentos cualesquiera de la colección deben pertenecer a un mismo grupo o no.

1.2.2 Medidas de similitud para el agrupamiento de documentos

Una medida de similitud (o de distancia) refleja el grado de cercanía de dos objetos dados (Nalawade et al., 2016), para el caso específico del agrupamiento de texto, permite calcular qué tan similares son dos documentos. Toda medida de similitud debe cumplir las dos condiciones siguientes:

- $S_{x,y} \leq S_{x,x}\ \forall x, y$ con igualdad solo cuando x=y
- $S_{x,y} \in [0,1]$

Donde $S_{x,y}$ indica la similitud que existe entre los objetos *x* e *y*.

La elección de una medida de similitud adecuada no es trivial. Para encontrar los agrupamientos naturales, la noción de similitud debe adaptarse al problema particular; es por ello que actualmente en el mundo se trabaja en la obtención de medidas que trabajen sobre tipos de datos específicos. Varias son las medidas que pueden aplicarse para determinar el grado de semejanza entre un par de documentos. Entre las más usadas están: *Dice* (Vargas Flores et al., 2016), *Jaccard* y *Coseno* (Lin et al., 2014).

En la literatura se encuentran diversos trabajos relacionados con este tema (Huang, 2008, Iezzi, 2012, Lin et al., 2014, Rajeshwari et al., 2015, Magdaleno, 2015). En algunos de ellos se desarrollan nuevas funciones de similitud (Lin et al., 2014, Magdaleno, 2015) mientras que otros se dedican a la comparación de la eficacia y eficiencia de aplicar algunas de las funciones existentes en diferentes contextos (Penney et al., 1998, Huang, 2008).

Uno de los trabajos más recientes relacionado con el desarrollo de una función de similitud para la clasificación y el agrupamiento de texto es (Lin et al., 2014). Este se enfoca en calcular la similitud entre dos documentos con respecto a un rasgo dado. Para ello la medida propuesta toma en cuenta los siguientes tres casos: (1) el mismo rasgo aparece en ambos documentos, (2) el mismo rasgo aparece en un solo documento, (3) el mismo rasgo no aparece en ninguno de los documentos. Los autores refieren que esta función obtiene mejores resultados que otras medidas de similitud, basado en los resultados de los experimentos aplicados.

En (Huang, 2008) la autora analiza y compara la efectividad de aplicar la similitud *Coseno*, la distancia *Euclidiana*, el coeficiente de correlación de *Pearson*, el coeficiente *Jaccard* y la divergencia *Kullback-Leibler* (KLD) en el agrupamiento particional de conjuntos de datos textuales. Los resultados de este trabajo demuestran que, excepto la distancia *Euclidiana*, el resto de las medidas muestran resultados similares en cuanto a la efectividad para la tarea del agrupamiento particional de documentos textuales. El coeficiente de correlación de *Pearson* y la *KLD* muestran ligeramente mejores resultados, dado que los grupos obtenidos son más balanceados y más similares a los propuestos en la clasificación de referencia manual. También se demuestra que los grupos obtenidos usando el coeficiente de *Pearson* y el coeficiente *Jaccard* son más coherentes.

En (Amador Penichet et al., 2018) se presenta la función de similitud SimRefBib, la cual permite obtener el grado de semejanza que existe entre dos artículos científicos, basado en la información que brindan las referencias bibliográficas. Esta función tiene una característica que la hace singular con respecto a otras funciones de similitud, y es que en la mayoría de los casos los valores de similitud obtenidos entre artículos que se consideran semejantes no son altos, lo cual es esperado cuando se aplica una función para la comparación de documentos. Pero al mismo tiempo se tiene que para artículos que no se consideran semejantes, el valor de similitud que se obtiene en la gran mayoría de los casos es cero. Esta última peculiaridad permite obtener una matriz de similitud donde se pueda discernir con facilidad si dos documentos deben pertenecer a un mismo grupo o no. Esta función, aunque fue diseñada específicamente para la comparación de

artículos científicos, también puede ser usada en colecciones de documentos que posean una estructura similar a la de las referencias bibliográficas de los artículos. Un ejemplo es el agrupamiento de curriculum vitaes, el cual permite dado un conjunto de autores conocer cuáles trabajan líneas de investigación relacionadas.

1.3 Agrupamiento de artículos científicos

El aumento considerable del volumen de información científica, junto a la necesidad de dar respuestas más rápidas a los problemas científicos actuales, ha traído consigo que investigadores de las ciencias de la información y de ramas afines como la computación, direccionen sus investigaciones a obtener métodos que faciliten el proceso de recuperación de información científica.

De este modo han surgido trabajos que van desde, sistemas diseñados para recomendar bibliografía científica de manera online, como es el desarrollado por (Wang and Blei, 2011), hasta otros más particulares como el desarrollado en (Kuna et al., 2015) que se centran en áreas específicas de la ciencia para poder explotar de manera más eficiente la búsqueda de información, auxiliándose de las facilidades que ofrece trabajar sobre contextos específicos. En el primero de los trabajos mencionados anteriormente, los autores combinan las bondades del filtrado colaborativo y el modelo probabilístico para generar un modelo capaz de sugerir a los autores bibliografías que deban consultar, las cuales se relacionen con las líneas de investigación que ellos llevan a cabo. En el segundo de los trabajos, los autores se enfocan en un contexto más específico y presentan un modelo para la recuperación de información específicamente en el área de la Ciencia de la Computación.

Dentro de las formas de gestión de la información, y en particular de gestión de la información científica, el agrupamiento ha ganado popularidad. Así se pueden encontrar varios trabajos en la literatura que usan técnicas de agrupamiento para lograr este propósito.

La mayoría de los trabajos que se encuentran en la literatura relacionados con el agrupamiento de artículos científicos se dirigen a usar el índice de co-citación de los artículos para determinar qué tan similares son. El número de documentos que citan a la

vez dos artículos científicos define la fuerza de co-citación entre estos últimos. Por lo tanto, la co-citación se puede definir como la frecuencia con la que dos artículos de la literatura son citados por un nuevo artículo (Small, 1973, Wang et al., 2013). La hipótesis que ha llevado a los autores a desarrollar investigaciones en este tema es que dos artículos que sean referenciados a la vez por otros artículos, un número considerable de veces, deben tratar temas similares (Hu et al., 2010). Varios son los trabajos que se pueden encontrar donde se hace uso de la co-citación para el agrupamiento de artículos científicos (Garfield et al., 2013, Boyack et al., 2013, Small, 1993), algunos de ellos se menciona a continuación.

Uno de los primeros trabajos reportados en la literatura que se enfoca en la clasificación específicamente de artículos científicos es (Garfield et al., 2013). Para el desarrollo de este sistema los autores utilizan artículos pertenecientes a la base de datos *ISI*. El método desarrollado consiste en determinar la relación que existe entre los diferentes pares de artículos teniendo en cuenta el número de veces que son co-citados, es decir, la cantidad de veces que ellos son referenciados por un mismo artículo. Luego de obtener para cada par de artículos el número de co-citas, realizan un proceso de agrupamiento, donde un par de artículos pertenece a un mismo grupo si su número de co-citas supera un umbral determinado. Una vez obtenidos los grupos es necesario etiquetar los mismos, que sería: seleccionar aquellas palabras que son representativas para cada grupo. Los autores refieren que esta es la única parte del proceso de clasificación en que es requerida la participación de un humano, lo cual ya es suficiente para no considerarlo como un proceso automático, sino semiautomático. Para clasificar un nuevo documento se buscan las coincidencias entre las referencias de este y las referencias de los artículos de cada clúster. El nuevo documento va a ser etiquetado con las etiquetas cabeceras de aquellos clústeres con los cuales sus citas tuvieron coincidencia.

En (Small and Sweeney, 1985) los autores definen una nueva forma de determinación del umbral de citas requeridas en el proceso de agrupamiento de los artículos científicos registrados en el *ScienceCitationIndex* (SCI). Este nuevo enfoque se denominó conteo fraccionario de citas. A diferencia del enfoque anterior que se basa solamente en determinar para un artículo *i* cuántos artículos lo referencian, este otorga al artículo *i* el

valor $1/r_j$, donde r_j es la cantidad de referencias del artículo *j*, el cual referencia al artículo *i*. Por tanto, el valor de co-citas del artículo *i* está dado por la suma de los valores fraccionarios de todos los artículos en los cuales es referenciado. También los autores proponen un nuevo enfoque para realizar el agrupamiento de los artículos denominado agrupamiento de nivel variable. Para la aplicación de este enfoque se normalizan los valores de co-citas usando alguna función (los autores recomiendan el coeficiente *Jaccard* o la función *Coseno*). Luego se define la cantidad máxima de elementos de un grupo, así como el valor del nivel inicial y el valor para incrementar el nivel. De esta forma todos los artículos que pertenezcan a un grupo serán aquellos que superaron el valor del nivel inicial. En caso de obtener grupos con más artículos que la cantidad máxima definida, este grupo pasa al próximo nivel, que sería el valor del nivel actual más el incremento. Por tanto, se van a formar a partir de este grupo nuevos grupos en los cuales los elementos que pertenezcan a ellos superan el nuevo nivel establecido. Según los autores el uso de ambos enfoques mejora los resultados del agrupamiento de bases de datos interdisciplinarias como SCI. Algunos factores que pueden afectar los resultados de este método son la existencia de artículos con muy pocas citas, o las peculiaridades que presentan algunos artículos como los de corte biomédico que contienen un número elevado de referencias bibliográficas.

Otro de los trabajos que hace uso del índice de co-citación para agrupar artículos científicos es (Hu et al., 2010). En este trabajo se usa la frecuencia de co-citación de 24 revistas chinas de bibliotecología y ciencias de la información para descubrir la relación que existe entre las mismas. Los resultados obtenidos permiten agrupar las revistas en cuatro grupos fundamentales, logrando de esta manera relacionar aquellas revistas que tratan temas más afines.

En (Boyack et al., 2013) los autores proponen diferentes métodos para mejorar la calidad del agrupamiento de artículos científicos basado en las co-citas de los mismos. Estos métodos, usando diferentes enfoques, analizan la posición en donde aparecen las co-citas en el texto. Los resultados de los experimentos aplicados demuestran que dos referencias que se encuentren bastante cercanas en un artículo son más parecidas que dos que se encuentren más alejadas. También basado en los resultados experimentales los autores

demostraron que a medida que aumentan los valores de similaridad disminuye el tamaño de los grupos obtenidos. Al mismo tiempo la coherencia de los grupos obtenidos aumenta cuando disminuye el tamaño de los grupos.

También se han desarrollado algunos trabajos en el agrupamiento de documentos científicos que han abordado otros enfoques diferentes al análisis de la co-citación (Aljaber et al., 2010, Magdaleno, 2015).

En (Aljaber et al., 2010) los autores proponen un nuevo enfoque para el agrupamiento de artículos científicos basado en la utilización del contexto de citación. Se considera el contexto de citación como el fragmento de texto que se encuentra alrededor de alguna marca de referencia usada para citar otro artículo científico. Los autores consideran que el contexto de citación provee sinónimos y vocabulario relativo al tema específico que trata el artículo científico, lo cual contribuye a incrementar la efectividad de la representación bolsa de palabras. Los resultados obtenidos, a través de los experimentos realizados, demuestran que el uso del contexto de citación combinado con el vocabulario del texto completo es una alternativa prometedora a la hora de determinar temas críticos en los artículos científicos.

En (Magdaleno, 2015) se presenta una nueva metodología de agrupamiento de artículos científicos en formato semiestructurado. Esta metodología hace uso tanto de la estructura como del contenido del documento para lograr mejores resultados en el agrupamiento. Para ello el autor desarrolla una función de similitud que permite mezclar los resultados del agrupamiento de los artículos viendo cada unidad estructural de los mismos de manera independiente y considerando el artículo completamente sin tener en cuenta las unidades por las cuales está compuesto. Los resultados obtenidos a través de los experimentos realizados demuestran que explotar de manera conjunta la estructura y el contenido de los artículos científicos mejora considerablemente los resultados del agrupamiento de artículos científicos.

En (Amador et al., 2017) se presenta el algoritmo de agrupamiento SemClustDML. Este algoritmo solo se basa en la información que brindan las referencias bibliográficas de los artículos científicos para obtener los grupos. La gran ventaja de este método es que no es necesario procesar todo el documento para realizar la extracción de los términos

representativos, por lo cual se reduce la complejidad del método de agrupamiento de manera general. Además, viendo el método solo desde el punto de vista de algoritmo de agrupamiento, posee la ventaja de ser capaz de obtener la cantidad de grupos de la colección sin tener que ser este un parámetro proporcionado por el usuario.

Los algoritmos de agrupamiento, de manera general, permiten obtener grupos de objetos basado en la similitud que existe entre los mismos. Sin embargo, no siempre los resultados del proceso de agrupamiento son lo suficientemente buenos. Es por esto que en muchos casos se pueden obtener, al aplicar algún método de agrupamiento, grupos donde los elementos que lo conforman no son todos similares entre sí, o lo que es lo mismo, pueden obtenerse elementos que fueron agrupados erróneamente. Esto atenta directamente contra la calidad del agrupamiento y, en consecuencia, contra la facilidad en la gestión de la información del usuario final. Es por este motivo que se necesita evaluar el desempeño de un algoritmo de agrupamiento antes de brindarlo como método de gestión de la información a los usuarios. Es válido aclarar que no necesariamente los resultados desfavorables en un proceso de agrupamientos se relacionan directamente con la eficacia del algoritmo que se usó en el mismo. En muchos casos esto se debe a configuraciones erróneas de los parámetros de entrada que necesita el método utilizado. En otros casos, se debe a que el algoritmo no es bueno en el contexto particular que se usó, lo cual no indica que no vaya a ser bueno en otros contextos.

1.4 Medidas de evaluación de la calidad del agrupamiento

Para evaluar la calidad de los métodos de agrupamiento se han desarrollado disímiles medidas, estas se pueden clasificar siguiendo diversos criterios (Mederos and Ruiz, 2012). De este modo pueden ser locales o globales en dependencia de si se usan para evaluar la calidad de un grupo específico o la calidad del resultado de un agrupamiento completo. Otro criterio está relacionado con la objetividad de la medida, así, si la medida evalúa la calidad de los grupos obtenidos tomando como referencia parámetros como la densidad de los grupos, qué tan compactos son o el grado de separación entre los mismos se considera una medida objetiva. En cambio, si la medida está más orientada a evaluar la calidad de los resultados tomando como referencia la usabilidad de los grupos obtenidos, se está en presencia de una medida subjetiva. Sin lugar a dudas, una de las

clasificaciones más usadas es la que separa las métricas para medir la calidad del agrupamiento en métricas internas y métricas externas. En los subepígrafes siguientes se profundiza en cada una de ellas.

1.4.1 Medidas internas para evaluar la calidad del agrupamiento

Las medidas internas son aquellas que permiten evaluar la calidad de los grupos obtenidos sin tener en cuenta el conocimiento externo que se tiene de ellos. De este modo se tienen medidas que hacen énfasis en la cohesión de los grupos, como puede ser la *Overall Similarity*, la cual calcula la calidad del agrupamiento basado en la similitud que tienen los pares de objetos en cada grupo (Mederos and Ruiz, 2012).

Otras medidas se basan en la dispersión que existe dentro de los grupos formados y la separación que existe entre ellos como es el caso del índice de Calinski y Harabasz (Caliński and Harabasz, 1974). Este índice determina la cohesión de un grupo basado en la distancia que existe desde todos los elementos del grupo hacia el centroide del mismo. La separación entre grupos se define como la distancia que existe desde cada centroide al centroide global.

Según (Arbelaitz et al., 2013) el índice Davies–Bouldin (ecuación 1.2) es uno de los más usados como medida interna de calidad del agrupamiento. Este usa el mismo principio del índice Calinski–Harabasz para obtener la cohesión intra grupos. Para buscar la separación entre los grupos calcula la distancia que existe entre los centroide de los mismos. En (Pal and Biswas, 1997) se proponen dos variantes de este índice basado en la teoría de grafos, donde se modifica la forma en que se calcula la cohesión intra grupos. Otros trabajos donde usan este índice son (Xiao et al., 2017, Karo et al., 2017, Bala et al., 2015).

$$DB(C) = \frac{1}{K}\sum_{C_k \in C} \max_{C_l \in C \setminus C_k} \left\{ \frac{dP(C_k) + dP(C_l)}{d(\overline{C_k}, \overline{C_l})} \right\} \tag{1.2}$$

En la ecuación anterior *C* representa el conjunto de grupos que se obtuvieron para la colección, *K* representa la cantidad de grupos y $dP(C_i)$ representa la distancia promedio al centroide dentro del clúster *i*, la cual se define como la suma de las distancia de todos

los elementos del clúster al centroide del mismo, dividida entre la cantidad de elementos del clúster.

Existen contextos de agrupamiento en los que no se dispone de centroides en los grupos obtenidos. En algunos casos porque es complicada la obtención de los mismos y en otros porque carece de sentidos hallar los centroides debido a la forma particular de agrupamiento que se aplica. Sobre esta base se han desarrollado algunos índices que permiten evaluar la calidad del agrupamiento en estos contextos donde no se dispone de centroides para los grupos.

Como medida interna para evaluar la calidad del agrupamiento destaca el índice Dunn (ecuación 1.3). Este se basa en la teoría que los grupos obtenidos mientras más compactos sean y más separados se encuentren entre sí mejor es el resultado del agrupamiento. De este modo se calcula la separación entre cada par de grupos tomando la distancia que existe entre los dos elementos más cercanos que pertenecen a cada uno de estos grupos. Para medir la cohesión del grupo se toma la mayor distancia que existe entre todos los pares de elementos que pertenecen al grupo. Con estas dos medidas se establece una proporción, donde es deseable que la distancia entre los grupos sea lo más grande posible y la distancia intragrupos sea lo más pequeña posible. Algunos autores como James Bezdek y Nikhil Pals plantean que el índice Dunn es muy sensible al ruido lo que sugiere que en grupos con elementos cercanos los resultados no son favorables. Con el objetivo de reducir las limitantes asociadas a este índice se han propuesto algunas variantes del mismo, tal es el caso del trabajo desarrollado por los autores referidos anteriormente en (Bezdek and Pal, 1995) donde propusieron cinco modificaciones del índice Dunn para disminuir su sensibilidad al ruido y de esta manera poder validar agrupamientos donde los grupos obtenidos presentan distintas formas hiperesféricas. También fueron propuestas las modificaciones D^{MST},D^{RNG}, D^{GG} para este índice en (Pal and Biswas, 1997). Estas últimas con el objetivo de variar la forma en que es determinada la cohesión intragrupo.

$$Dunn(C) = \frac{min_{C_k \in C}\left\{min_{C_l \in C \setminus C_k}\{dist(C_k, C_l)\}\right\}}{max_{C_k \in C}\{DistIn(C_k)\}} \quad (1.3)$$

Donde $dist(C_k, C_l)$, $DistIn(C_k)$ se definen como sigue:

$$dist(C_k, C_l) = \min_{x_i \in C_k} \min_{x_j \in C_l} \{d(x_i, x_j)\}$$

$$DistIn(C_k) = \max_{x_i, x_j \in C_k} \{d(x_i, x_j)\}$$

Donde *d* representa la distancia que existe entre los objetos x_i y x_j.

El índice Silhouette (Rousseeuw, 1987) no necesita disponer de los centroides de cada grupo para determinar la calidad del agrupamiento basado en la estructura de los grupos obtenidos. Este índice es del tipo suma-normalizada. Para medir la cohesión de un grupo calcula la suma de las distancias entre todos los pares de elementos pertenecientes al grupo. Para medir la separación entre un par de grupos calcula la distancia que existe entre los dos elementos más cercanos pertenecientes a cada uno de estos grupos. Este se define matemáticamente en la ecuación siguiente:

$$Sil(C) = \frac{1}{N} \sum_{C_k \in C} \sum_{x_i \in C_k} \frac{dpeOC(x_i, C_k) - dpeC(x_i, C_k)}{\max\{dpiC(x_i, C_k), b(x_i, C_k)\}} \quad (1.4)$$

Donde $dpeC(x_i, C_k)$ representa la distancia promedio del elemento x_i al resto de los elementos del clúster al cual pertenece, $dpeOC(x_i, C_k)$ representa la distancia promedia mínima de cada elemento x_i perteneciente a un clúster con respecto a los demás clústeres. Cada una de ella se define matemáticamente como sigue:

$$dpeC(x_i, C_k) = \frac{1}{|C_k|} \sum_{x_j \in C_k} d(x_i, x_j)$$

$$dpeOC(x_i, C_k) = \min_{C_l \in C \setminus C_k} \left\{ \frac{1}{|C_l|} \sum_{x_j \in C_l} d(x_i, x_j) \right\}$$

Se pueden encontrar en la literatura otros índices que evalúan la calidad del agrupamiento basado en la estructura de los grupos obtenidos. Entre ellos están los índices C (Rousseeuw, 1987), CS (Chou et al., 2004), el *Score function* (Saitta et al., 2007), el *Sym-index* (Bandyopadhyay and Saha, 2008), entre otros.

En (Arbelaitz et al., 2013) se hace un estudio detallado de 30 métricas internas para medir la calidad del agrupamiento. Los autores aplican estos índices en colecciones de datos con características variadas, de esta manera analizan la usabilidad de cada uno de ellos en diferentes contextos. Según los autores, los experimentos realizados no tuvieron como objetivo demostrar cuáles índices son mejores, dado que este comportamiento, en muchos casos, depende del contexto específico de aplicación. Sin embargo, las pruebas realizadas arrojaron que de manera general los índices Silhouette, Davies–Bouldin, Calinski–Harabasz, *generalized Dunn*, COP and SDbw destacan con respecto a los demás, dado que obtienen los mejores resultados en la mayoría de las colecciones probadas. Existen otros trabajos precedentes al de Arbelaitz y colaboradores donde se establecen comparaciones entre varios índices internos para la validación de agrupamientos (Milligan and Cooper, 1985, Dubes, 1987, Bezdek et al., 1997, Brun et al., 2007). Sin embargo, ninguno de ellos establece una comparación tan amplia en cuanto a variedad de índices y diversidad y tamaño de las colecciones utilizadas como la realizada en (Arbelaitz et al., 2013).

Las medidas internas reflejan propiedades estructurales de los grupos obtenidos. Sin embargo, no garantizan que estos grupos estén acordes con las necesidades de los usuarios. Es por ello que existen otro tipo de medidas que reflejan hasta qué punto los grupos obtenidos se asemejan a los grupos que se hubiesen obtenidos por la clasificación de un humano. Estas son las medidas externas.

1.4.2 Medidas externas para evaluar la calidad de agrupamiento

Las medidas externas evalúan qué tan bueno es el agrupamiento obtenido, mediante la comparación de los grupos resultantes con las clases de referencias de la colección que se ha agrupado. Es por este motivo que, para el uso de las mismas, se necesita que los documentos a agrupar estén previamente etiquetados. De este modo la métrica utilizada puede determinar si el documento ha sido colocado en el grupo correcto o no.

Dentro de las medidas externas se encuentra la Entropía de Shannon (Shannon, 2001). Esta medida calcula la distribución de las clases de referencia en cada uno de los grupos obtenidos. La entropía total es calculada como la sumatoria de las entropías de cada grupo (E_j), ponderada con el tamaño del grupo (n_j) y dividida entre la cantidad de

documentos de la colección (*n*) como se muestra en la Figura 1-1c. Es válido aclarar que los mejores resultados al aplicar esta medida se obtienen cuando cada grupo contiene exactamente un elemento. Si se analizan las ecuaciones a y b de la Figura 1-1 se puede llegar a la conclusión de que la probabilidad de pertenencia de los elementos de la clase *i* al grupo *j* representada como p_{ij},, que se calcula como se muestra en la Figura 1-1a donde n^i_j representa la cantidad de elementos de la clase *i* que fueron asignados al cluster *j*, va a ser uno para la clase que contiene al único elemento que forma parte del grupo y cero en el resto de los casos, por lo que la entropía del grupo *j* va a ser cero, ya que el logaritmo de p_{ij} va a tomar valor cero. Dado que esto va a suceder para todos los grupos, la entropía general va a ser la sumatoria de las entropías de *n* grupos, donde la entropía de cada grupo vale cero, dividido entre la cantidad de grupos. Por lo que se obtiene que la entropía general es cero, que es el valor óptimo para esta medida. Es necesario tener en cuenta esta particularidad referente a la entropía ya que, en la evaluación del agrupamiento, al aplicar esta medida, se puede obtener el valor óptimo de entropía sin que necesariamente se hayan obtenido los grupos esperados, sino que se haya obtenido el conjunto de documentos inicial.

a $p_{ij} = \frac{n_j^i}{n_j}$

b $E_j = -\sum_i p_{ij} \log(p_{ij})$

c $E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n}$

Figura 1-1 Ecuaciones para calcular la Entropía de Shannon.

En (Karypis et al., 1999) presentan otra variante para el cálculo de la entropía del grupo. En esta normalizan el valor que se obtiene al aplicar la ecuación de la Figura 1-1b dividiéndolo entre el logaritmo de la cantidad de grupos obtenidos (Ecuación 1.5).

$$E_j = -\frac{1}{\log(m)} \sum p_{ij} \log(p_{ij}) \quad (1.5)$$

Otra medida externa que permite evaluar la calidad del agrupamiento es el índice Folkes-Mallows (Desgraupes, 2013). Este mide la eficacia del agrupamiento basado en la proporción de los pares de objetos correctamente agrupados entre la suma de estos últimos con los incorrectamente agrupados. Son considerados pares de objetos incorrectamente agrupados tanto los que son asignados a un mismo grupo, pero pertenecen a clases de referencia diferentes, como aquellos que son asignados a grupos

diferentes y pertenecen a una misma clase de referencia. Formalmente se define este índice como se muestra en la ecuación 1.6.

$$FM = \left(\frac{a}{a+b} * \frac{a}{a+c}\right)^{1/2} \tag{1.6}$$

En esta ecuación *a* indica los pares de objetos que pertenecen a la misma clase de referencia y fueron asignados al mismo grupo, estos son los pares de objetos correctamente agrupados, *b* representa los pares de objetos que fueron asignados al mismo grupo, pero pertenecen a clases diferentes y *c* los pares de objetos que fueron asignados a grupos diferentes, pero pertenecen a la misma clase de referencia, por lo cual *b+c* representa la cantidad de objetos incorrectamente agrupados. Este índice a diferencia de la entropía de Shannon obtiene su peor valor cuando cada grupo contiene exactamente un elemento, ya que, para este caso extremo *a* y *b* toman valor cero, por lo que se indefine el cálculo de esta métrica y se asume por tanto que el índice Folkes-Mallows en este caso vale cero. También es lógico pensar que se indefine el cálculo de esta métrica para cuando *a* y *c* valen cero. Sin embargo, esto nunca ocurre dado que si *c* vale cero es porque no existe ningún par de documentos que pertenezca según la clasificación de referencia a una misma clase y que haya sido asignado a grupos diferentes en el agrupamiento, con lo cual *a* va a ser obligatoriamente distinto de cero. Resulta interesante analizar cuánto vale el índice Folkes-Mallows cuando se obtiene un solo grupo como resultado del proceso de agrupamiento. Ya que es presumible que si para *n* grupos (con *n* cantidad de objetos a agrupar) el índice toma valor cero, entonces para un solo grupo puede que el índice tome valor uno, lo cual no es deseado. Dado que $a+b+c=n(n-1)/2$ y en este caso $c=0$, se tiene que $a+b= n(n-1)/2$. En este caso *b* es obligatoriamente distinto de cero porque todos los objetos de la colección fueron asignados al mismo grupo, lo que implica que en este grupo existen objetos de diferentes clases. Esto indica que el índice Folkes-Mallows es distinto de uno en este caso particular. El análisis de estos casos extremos permite concluir que el índice Folkes-Mallows resulta una métrica consistente a la hora de determinar la calidad del agrupamiento.

Las medidas Micro-Purity y Macro-Purity, propuestas por Pinto y colaboradores en (Pinto et al., 2010), también permiten evaluar la calidad del agrupamiento basado en una

clasificación previa que se tenga de la colección que se sometió al proceso de agrupamiento. La idea general de estas medidas es que la mayor cantidad de documentos dentro de cada grupo pertenezcan a una sola clase. De aquí que se obtienen valores elevados de Purity para agrupamientos donde los grupos obtenidos presentan una alta homogeneidad, o lo que es lo mismo: los altos valores para esta medida no se obtienen solamente cuando los grupos obtenidos son muy similares a las clases de referencia, sino que la obtención, en el proceso de agrupamiento, de grupos que constituyan subgrupos de las clases de referencia también reportan altos valores de Purity. Las medidas Micro-Purity y Macro-Purity se calculan como se muestra en la Figura 1-2.

a

$$Micro - Purity = \frac{\sum_{i=1}^{k} Purity(i) * CCC_i}{\sum_{i=0}^{k} CCC_i}$$

b

$$Macro - Purity = \frac{\sum_{i=1}^{k} Purity(i)}{TC}$$

Figura 1-2 Ecuaciones para calcular (a) Micro-Purity y (b) Macro-Purity.

En las ecuaciones anteriores *k* indica la cantidad de clústeres obtenidos, CCC_i la cantidad de clases presentes en el clúster *i*, *TC* es la cantidad de grupos encontrados en el agrupamiento y *Purity(i)* se calcula como se muestra en la ecuación 1.7.

$$Purity(i) = \frac{CMRC_i}{|C_i|} \tag{1.7}$$

Donde $CMRC_i$ indica la cantidad de elementos en el grupo *i* que pertenecen a la clase más representada en dicho grupo y $|C_i|$ representa la cantidad total de elementos del grupo *i*.

Una de las medidas externas ampliamente utilizadas en la evaluación de agrupamientos es la *Overall F-measure* (OFM) (Magdaleno Guevara et al., 2015, Amador Penichet et al., 2018, Arco et al., 2006, Kumar et al., 2016). Esta combina los conceptos de precisión y cubrimiento a la hora de calcular qué tan bueno ha sido el resultado obtenido por un método de agrupamiento, basado en una clasificación de referencia. La precisión está referida, en el caso del agrupamiento, a que los documentos que sean ubicados en un grupo, pertenezcan en efecto a ese grupo según la clasificación de referencia. El cubrimiento busca que para cada grupo se logren asignar la mayor cantidad de documentos que según la clasificación de referencia debieran pertenecer al grupo. De esta manera al obtener valores cercanos a uno para la medida OFM se garantiza que los

resultados del agrupamiento sean más eficaces. En la Figura 1-3a se muestra la forma de calcular la precisión (Pr) y el cubrimiento (Re) para cada clúster *j* con respecto a la clase *i*, n_{ij} representa la cantidad de elementos de la clase *i* que fueron asignados al clúster *j;* n_i y n_j representan la cantidad de elementos de la clase *i* y del grupo *j* respectivamente. En la Figura 1-3b se muestra la ecuación para calcular la *F-measure* para un clúster con respecto a una clase, es aquí donde se establece un balance entre los conceptos de precisión y cubrimiento mediante el parámetro α con $0 \leq \alpha \leq 1$. Si se desean mayores valores de precisión se le asignan a α valores cercanos a uno. Si en cambio se desea que el cubrimiento sea mayor se le asigna a α valores cercanos a cero. De aquí se deduce que, si lo deseado es que el algoritmo de agrupamiento obtenga grupos precisos y al mismo tiempo esos grupos aglomeren la mayor cantidad de documentos posibles pertenecientes a la colección, se debe establece α=0.5, de esta forma se le otorga la misma importancia a la precisión y al cubrimiento. Por último, la Figura 1-3c muestra la forma de calcular la OFM para un conjunto de grupos obtenidos mediante un proceso de agrupamiento.

a

$$\Pr(i,j) = {n_{ij}}/{n_j}$$

$$\mathrm{Re}(i,j) = {n_{ij}}/{n_i}$$

b

$$F_\alpha(i,j) = \frac{1}{\alpha \frac{1}{\Pr(i,j)} + (1-\alpha)\frac{1}{\mathrm{Re}(i,j)}}$$

c

$$F = \sum_i \frac{n_i}{n} \max_j \{F_\alpha(i,j)\}$$

Figura 1-3 Ecuaciones para el cálculo de la Overall F-Measure.

Es evidente que las medidas externas ofrecen un criterio más sólido para determinar qué tan bueno es el resultado de un agrupamiento, dado que están basadas en la clasificación previa que un humano hizo sobre una colección dada. Sin embargo, esta misma potencialidad constituye una limitante para el uso de este tipo de medidas, porque en la mayoría de los casos no se tiene esta clasificación previa y se desea conocer qué tan bueno ha sido el resultado obtenido en el agrupamiento. Una posible solución a este problema es evaluar el método de agrupamiento con diferentes colecciones de las cuales se tenga clasificación de referencia. De este modo si el resultado obtenido para varios índices externos resulta favorable, se puede asumir que el método de agrupamiento es eficaz y por tanto aplicarlo a colecciones donde no se tiene una clasificación previa de los objetos.

Otra de las variantes que se puede aplicar es tratar de relacionar mediante modelos de regresión los valores obtenidos para las medidas internas y externas aplicadas a un agrupamiento. Para ello se utilizan colecciones de documentos previamente etiquetadas y se aplican varios métodos de agrupamiento, dando lugar a un conjunto de resultados de agrupamiento que son validados usando medidas externas e internas. Mediante la regresión lineal se analiza entre cuáles de estas medidas se puede establecer una correlación. Esto permite luego aplicar solo índices de validación internos a las colecciones de las cuales no se tiene clasificación previa. Mediante estos índices se puede analizar si los resultados obtenidos son buenos o no, calculando a través del modelo de regresión los valores que toman las medidas externas. El trabajo (Ingaramo et al., 2007) es uno de los que se basa en esta idea. En este los autores intentan establecer una correlación entre la medida de *densidad esperada* (interna) y la media *F-measure* (externa) en el agrupamiento de resúmenes científicos en dominios reducidos. Los experimentos aplicados sobre varias colecciones permiten llegar a la conclusión que en este tipo de contexto no existe una correlación entre las medidas sometidas a evaluación. Sin embargo, estas mismas medidas fueron analizadas por Stein y Niggemann en (Stein and Niggemann, 1999) en un corpus estándar etiquetado y los resultados obtenidos sí fueron favorables. Es de aquí que se puede deducir que el establecimiento de una correlación entre índices internos y externos de validación depende en gran medida del contexto de aplicación.

1.5 Conclusiones parciales

- La gestión del conocimiento mediante el agrupamiento de artículos científicos es un tema que ha sido estudiado por numerosos autores. Sin embargo, la mayoría de los trabajos desarrollados se enfocan en la co-citación de los artículos científicos para obtener los grupos de documentos similares. Este método tiene como desventaja principal, que se necesita tener acceso a índices internacionales que recogen la información referente a las citas de los artículos científicos. El acceso a estos índices no está disponible libremente para todos los usuarios, en la mayoría de los casos hay que pagar por ellos.

- El texto completo de los artículos científicos tampoco está disponible en todos los casos. Existe una gran cantidad de revistas de acceso abierto, pero no es menos cierto que la mayoría de las revistas de punta siguen cobrando por el acceso a los artículos que en ellas se publican. Sin embargo, el acceso al título, nombre de los autores, resumen, palabras claves y referencias bibliográficas es libre, incluso en las revistas que hay que pagar para obtener el artículo. De ahí la utilidad que pueden tener estos metadatos para los métodos de agrupamiento que se enfoquen en la información brindada en estas partes del artículo.
- La representación de documentos mediante términos independiente ignora en muchos casos la relación semántica que existe entre los mismo. Es por ello que se ha propuesto la representación de artículos científicos mediante frases relevantes, las cuales representan el tema general del cual trata el artículo, siendo esta forma de representación más eficaz que la representación por un modelo bolsa de palabras convencional a la vez que reduce la dimensionalidad de la representación.
- Las medidas internas permiten evaluar la calidad del agrupamiento teniendo en cuenta la estructura de los grupos obtenidos, sin necesidad de factores externos, ni clasificaciones previas. Sin embargo, en muchos casos los resultados obtenidos por estas medidas no reflejan las necesidades reales de los usuarios.
- Las medidas externas permiten evaluar la calidad del agrupamiento basado en una clasificación previa que se tenga de la colección de documentos sometida a agrupamiento. En la práctica no se dispone, para toda colección que se desee agrupar de una clasificación previa, por lo que resulta conveniente establecer una relación entre índices de evaluación internos y externos con colecciones previamente etiquetadas, para luego poder conocer que tan buenos son los resultados del agrupamiento de un método dado, basado en la evaluación previa que se tiene del mismo.

2

MÉTODO DE AGRUPAMIENTO A PARTIR DE LA EXTRACCIÓN DE FRASES RELEVANTES

Capítulo 2. Método de agrupamiento a partir de la extracción de frases relevantes.

Se propone un método para el agrupamiento de artículos científicos basado en la extracción de frases relevantes. Se analiza cuáles de las partes de los artículos científicos contienen términos más representativos del tema que trata el mismo. A partir de este análisis, se centra la extracción de frases relevantes en aquellas partes más representativas y se propone una nueva forma de agrupamiento, basada en la conexión que presentan los documentos teniendo en cuenta las frases relevantes que tiene en común.

2.1 Partes del artículo a tener en cuenta para la extracción de frases relevantes

El agrupamiento basado en palabras relevantes se fundamenta en la selección de términos que sean representativos de los temas que tratan los artículos de la colección que se desea agrupar. No todas las partes del artículo brindan términos relevantes que son representativos del tema que trata. Es por ello que primeramente se explica cuáles de estas partes son utilizadas para extraer los términos que serán usados para la representación específicamente de este tipo de documentos.

Un artículo científico puede ser divido en siete partes fundamentales o unidades estructurales: título, autores, afiliación, resumen, palabras claves, contenido y referencias bibliográficas.

Un autor no puede ser asociado unívocamente a un tema determinado ya que en un periodo de tiempo puede estar vinculado a una línea de investigación y luego vincularse a otra, por lo cual sus publicaciones no tributarían a un único tema, es por ello que no se tienen en cuenta los autores en el método de agrupamiento que se propone.

La afiliación de los autores queda totalmente descartada para la comparación de dos artículos, ya que dos artículos donde sus autores pertenezcan a una misma institución no tienen por qué tratar un mismo tema. Al mismo tiempo se tiene que dos artículos donde sus autores pertenezcan a instituciones diferentes pueden estar vinculados a una misma temática.

El contenido del artículo sí contiene términos que pueden determinar temas específicos. Sin embargo, la gran cantidad de términos que posee esta unidad estructural provoca que se obtengan tanto términos representativos como términos que no lo son. Estos últimos evidentemente generan ruido a la hora de la comparación de los artículos. Además, la gran cantidad de palabras que contiene esta unidad estructural hace que la extracción de términos sea computacionalmente compleja. Por lo cual, al comparar los beneficios de extraer términos de la unidad estructural "contenido" con las desventajas que esta extracción presenta se considera que la complejidad que supone la extracción es mayor que los beneficios que aporta.

En el caso del resumen pasa lo contrario a la sección "contenido". El resumen es un texto que generalmente no debe superar las 250 palabras y debe al mismo tiempo dar una panorámica general de todas las partes del artículo, es decir, debe dar una breve introducción, enunciar de manera concisa el trabajo realizado y mencionar los resultados obtenidos. Debido a esto, la extracción de términos relevantes en esta unidad se ve limitada por la gran variabilidad presentada.

Después de este análisis previo quedan como unidades potenciales para la extracción de términos relevantes: el título, las palabras claves y las referencias bibliográficas.

Las palabras claves constituyen términos que están relacionados con el tema que trata el artículo científico. No obstante, estas tienen la restricción, en muchos casos, que no deben aparecer en el título. Mayoritariamente las palabras que definen los autores como palabras claves, están relacionadas con la temática general que aborda el artículo, pero no son las más representativas del tema, ya que las palabras más representativas normalmente forman parte del título. Además, no tiene sentido hacer extracción de términos relevantes en el texto que forman las palabras claves debido a que es muy reducido. Por tanto, sería necesario tomar las palabras claves como aparecen y asignarle un peso directamente, ya que no es posible calcularlo dependiendo de su frecuencia de aparición, esta frecuencia es uno en todos los casos. Se prefiere, por tanto, no considerar las palabras claves propuestas por los autores, en busca de seleccionar aquellos términos más representativos del documento que no estén sujetos a restricciones, y en aras de reducir la cantidad de términos usados para representar el documento.

El título es una única oración, por lo que es lógico pensar que la extracción de términos en esta unidad estructural presenta el mismo problema que el resumen, o incluso peor, ya que cuenta con una cantidad de términos mucho menor. Esto es totalmente correcto, sin embargo, si se analizan todos los títulos de la colección a la cual se le desea realizar el proceso de agrupamiento como un conjunto, o lo que es lo mismo viendo todos los títulos como un único párrafo y en este párrafo se hace extracción de términos, entonces se obtendrán los términos más relevantes para la colección de manera general. Es por ello que el título del artículo es considerado en la extracción de términos.

Las referencias bibliográficas son consideradas muy importantes a la hora de determinar semejanza entre artículos científicos. Está claro que si dos artículos *i*, *j* referencian a varios artículos en común, estos dos artículos deben tratar temas similares. Sin embargo, no es necesario hacer uso de toda la información recogida en las referencias cuando se quiere obtener grupos de artículos similares basado en la extracción de términos relevantes. Por ejemplo, el año del artículo citado, las páginas, el volumen y otras informaciones recogidas en las referencias no se consideran relevantes, por esta razón, el método propuesto solo se centra en los títulos de los artículos referenciados. Para la selección de los términos relevantes se toman todos los títulos de las referencias de un artículo como si formaran un único párrafo y se extraen los términos de este párrafo. Generalmente los títulos de los trabajos contienen palabras que son representativas del tema que el artículo trata, es por ello que la extracción de términos en los títulos de las referencias bibliográficas garantiza que se obtengan palabras representativas del tema en cuestión que trata el artículo.

Con el análisis previo realizado se resume la extracción de términos a: los títulos de los artículos y a los títulos que forman parte de las referencias bibliográficas. Considerar solo estas dos partes del artículo en la extracción de términos presenta la ventaja que reduce el espacio de búsqueda, lo que tributa directamente a la reducción de la complejidad computacional. Además, los títulos y las referencias bibliográficas son partes del artículo que siempre están accesibles libremente, incluso en las revistas que hay que pagar para obtener el artículo. Esto permite, por tanto, poder realizar procesos de recuperación de información, agrupar los documentos obtenidos solo teniendo en cuenta estas partes de las cuales se dispone y luego centrarse en los documentos que

realmente son de interés para el usuario, sin haber tenido que pagar por una serie de documentos que probablemente no estuvieran relacionados de manera directa con las necesidades de información del mismo.

En el capítulo siguiente se demuestra estadísticamente que las referencias bibliográficas y los títulos de los artículos aportan los términos más representativos de los mismos.

2.2 Representación y transformación del corpus textual obtenido

Para la representación de un documento se va a tener en cuenta los títulos que forman parte de las referencias bibliográficas del mismo. Para ello se convierten estos títulos en una secuencia de tokens de palabras, en el paso subsecuente a la extracción de términos, estos tokens se usan para generar rasgos significativos (índices de términos). En la secuencia resultante de tokens se convierten todas las letras a minúsculas, se eliminan las marcas de puntuación al final de los tokens, se omiten los tokens que contienen caracteres alfa-numéricos, se sustituyen las contracciones por sus expresiones completas y se eliminan las palabras de parada.

Además, los títulos de cada artículo de la colección van a ser tomados en un único párrafo al cual también se le va a aplicar las transformaciones anteriormente mencionadas. En base a estos términos es que se van a seleccionar las palabras relevantes de la colección.

2.2.1 Extracción de frases relevantes

La extracción de frases relevantes se puede separar en dos fases. En la primera se extraen las frases relevantes para cada documento. Estas se extraen a partir de los títulos de las referencias bibliográficas de cada artículo. En la segunda fase se extraen las palabras relevantes de toda la colección. Estas son las que se obtienen al aplicar el proceso de extracción de términos a los títulos de todos los artículos pertenecientes a la colección que se desea agrupar. En ambas fases, la extracción de las palabras claves se realiza siguiendo el procedimiento siguiente.

1. Indexar cada uno de los tokens con sus respectivas frecuencias de aparición.
2. Seleccionar los tokens cuya frecuencia supera el umbral.

3. A partir de los tokens seleccionados en el paso anterior conformar las frases de tamaño dos.
4. Analizar la frecuencia de aparición de cada una de las frases de tamaño dos y seleccionar aquellas que superan el umbral.
5. Actualizar las frecuencias de aparición de manera independiente de los tokens/frases que conformaron la nueva frase.
6. Conformar cada frase *q* de tamaño *n* a partir de dos frases *v* (v_1,$v_{2,..}$, vn_{-1}) y *w* (w_1,$w_{2,..}$, wn_{-1}) de tamaño n-1, donde $v_{2,..}$, vn_{-1}= v_1,$v_{2,..,}$.v_{n-2}. Seleccionar las frases que superen el umbral. Aplicar paso 5.
7. Parar cuando no se pueda formar más frases o ninguna de las frases formadas en la última iteración supere el umbral.

Este procedimiento está basado en la estrategia *apriori* y sigue la idea del algoritmo GSP (Generalized Sequential Patterns) propuesto por Srikant y Agrawal (1996). La modificación fundamental realizada a este método consiste en la actualización de los pesos de las frases, cuando estas conformar una nueva frase que supera el umbral establecido para ser considerada relevante. Esta actualización permite disminuir la cantidad de frases de tamaño *n* que se consideran para general las frases de tamaño *n+1*.

Por ejemplo, supóngase que el umbral para considerar una frase como relevante en el documento d_i es 5 apariciones, además se tienen para este documento las frases de tamaño 2 con sus respectivas frecuencias (ab, 9), (bc, 5), (bd, 6), (xy, 8) y (xz,8). Se busca en el texto la frase relevante *abc* y se obtiene que su frecuencia de aparición es cinco, la cual supera el umbral. Por tanto, se actualizan las frecuencias de aparición de manera independiente de las frases que le dieron origen a esta, queda entonces (ab,4), (bc, 0), (bd, 6), (xy, 8) y (xz,8). De aquí se deduce que no tiene sentido analizar las restantes frases de tamaño tres que comiencen con *ab*, dado que la frecuencia máxima de aparición de las nuevas frases será como máximo cuatro, y este valor no supera el umbral. Además, *ab* y *bc* como frases independientes no se consideran representativas del documento en el resultado final, ya que sus frecuencias de aparición de manera independiente no superan el umbral.

A continuación, se explican detalladamente los pasos del método de extracción de frases relevantes, así como el tratamiento de las particularidades que se pueden presentar.

Inicialmente se parte de una secuencia de tokens y se produce una secuencia de términos indexados basados en estos, donde además se tiene la frecuencia de aparición de cada uno de ellos. El siguiente paso consiste en seleccionar solo aquellos tokens que constituyen palabras relevantes. En este paso se considera una palabra como candidata a relevante cuando su frecuencia de aparición supera el umbral *fap* (ecuación 2.1); *fap* es variable y depende de la cantidad de referencias bibliográficas (*CRB(i)*) del documento analizado, para el caso de la extracción de términos relevantes en cada documento. Para la extracción de frases relevantes de la colección, *CRB* es igual a la cantidad de documentos de la colección. Además *fap* es igual a *CRB/*10 cuando la colección presenta más de 100 documentos.

$$fap(i) = \begin{cases} 2 & si \quad CRB(i) \leq 10 \\ 3 & si \quad 10 < CRB(i) \leq 20 \\ 4 & si \quad 20 < CRB(i) \leq 30 \\ 5 & e.o.c \end{cases} \tag{2.1}$$

Es válido aclarar la posibilidad que ninguno de los tokens extraído supere el umbral *fap;* para estos casos extremos se seleccionarán los 4 términos con mayor frecuencia de aparición, siempre que el cuarto término tenga mayor frecuencia de aparición que el quinto, si esto no ocurre se seleccionarán todos los próximos términos con la misma frecuencia de aparición que el cuarto término.

Una vez obtenidos los tokens relevantes se pasa al proceso de unión de tokens. Este proceso es de suma importancia ya que no es de interés considerar los tokens obtenidos como simples términos aislados.

El proceso de unión de tokens, en su primera fase, consiste en buscar la frecuencia de aparición de los tokens relevantes (tomados dos a dos). Luego de obtenidos los pares de tokens relevantes, los cuales serían aquellos que superen el umbral *fap*, se analiza si algunos de los pares formados pueden unirse, esto se hace solo para los pares que la subcadena inicial del primer par coincida con la subcadena final del segundo, o viceversa, tomando como subcadena inicial la primera palabra del par y como subcadena final la última palabra. Este mismo proceso se aplica para los n-gramas de tamaño mayor que dos que se van formando. Para estos casos se toma como subcadena inicial las n-1 palabras iniciales y como subcadena final las últimas n-1 palabras. Siempre que se forme una nueva frase que supere el umbral, se actualiza la frecuencia de aparición de las dos

frases que le dieron origen, restándole a la frecuencia de cada una de ellas la frecuencia de la nueva frase formada.

Al terminar el proceso de unión de tokens se obtiene en la primera fase: las frases relevantes para el documento, así como la importancia de cada una de ellas. En la segunda fase se obtienen las frases relevantes para la colección.

El proceso de unión de tokens no será aplicado en aquellos documentos que ningún término supere el umbral *fap*, debido a que, si como términos independientes no superan el umbral, evidentemente no lo harán tampoco como términos unidos.

2.3 Método de agrupamiento

Luego del proceso de extracción de términos se tiene:

- el conjunto de términos que se consideran relevantes para la colección, estos son los que se extrajeron de los títulos de los artículos.
- Los términos relevantes para cada artículo. Estos son los que se extrajeron de las referencias bibliográficas de cada artículo.

A partir de este punto es lógico pensar en la aplicación de una medida de similitud o distancia para comparar los artículos de la colección, basado en las frases relevantes que fueron extraídas en el paso anterior. Sin embargo, con el método propuesto lo que se desea es obtener grupos de artículos donde cada grupo sea representativo de una temática. Las frases relevantes seleccionadas son representativas de los temas que tratan los artículos. Por tanto, la idea que se sigue es relacionar los artículos a partir de la conexión que ellos tienen, según las frases representativas que fueron seleccionadas para cada uno de ellos. Esta relación se representa mediante un grafo, donde cada vértice v_i representa un documento y si existe una arista entre un par de vértices (v_i, v_j) es porque estos contienen alguna palabra relevante en común. El peso de la arista está determinado por la cantidad de palabras relevantes en común que tienen los vértices que ella conecta.

La idea general del método propuesto consiste en seleccionar aquellos términos de la colección que son representativos. Esto permite obtener una partición del conjunto de documentos donde dos artículos que contengan uno de estos términos pertenecen a un mismo grupo. Los grupos por tanto son determinados por los términos representativos y en primera instancia se tendrán tantos grupos como términos representativos se tengan.

En la Figura 2-1 se formaliza el método de agrupamiento de artículos científicos basado en la representación de los mismos mediante frases relevantes. Es válido aclarar que los pasos del cero al dos están referidos a la extracción de los términos representativos de los títulos y de las referencias. Los pasos del tres al nueve son los referidos al agrupamiento.

Entrada: Conjunto de n artículos científicos $(D = \{d_1, d_2, \dots, d_n\})$, longitud mínima de cada cluster lmc.

Salida: Lista de clústeres formados (C)

Paso_0. Transformación del corpus.

Paso_1. Extracción de palabras claves de las referencias de cada artículo $PCRA$, conformar grafo de conexiones y seleccionar el conjunto de palabras relevantes de las referencias PRR.

Paso_2. Extracción de las palabras claves de los títulos de los artículos PCT.

Paso_3. Determinar lista de palabras centroides de la colección $PRCol \leftarrow PRR \cap PCT$

Paso_4. Para cada artículo i y cada palabra representativa j, si $PCRA_i \subset PRCol_j$ asignar el documento i al grupo C_j.

Paso_5. Para cada par de clúster (C_i, C_j) si $ElemSolp \geq 0.5 * |C_j|$, $C_i \leftarrow C_i \cup C_j$ donde $|C_i| \geq |C_j|$ y $ElemSolp \leftarrow C_j \cap C_i$.

Paso_6. $\forall d_i \in C_j \cap C_k$, si $PertProm(d_i, C_j) > PertProm(d_i, C_k)$, $C_k \leftarrow C_k \backslash d_i$ sino $C_j \leftarrow C_j \backslash d_i$

Paso_7. $\forall d_i \in C_j$ y $|C_j| < lmc$; $CT(d_i, C_k) \leftarrow \sum_{l=0}^{|C_k|} isConected(d_i, C_{kl})$ $C_k \leftarrow (C_k \cup d_i)$ donde $CT(d_i, C_k)$ es máxima y $C_k \neq C_j$,

$$isConected(d_i, d_l) = \begin{cases} 1 & si\ M(d_i, d_l) > 0 \\ 0 & e.o.c \end{cases}$$

Paso_8. $\forall d_i \in C_k, C_k \notin C$, si $CT(d_i, C_l) \geq CT(d_i, C_k)$ entonces $C_l \leftarrow (C_l \cup d_i)$ donde $CT(d_i, C_l)$ es máxima, $C_k \leftarrow C_k \backslash d_i$.

Paso_9. Si $|C_k| < lmc$ aplicar paso 7 a C_k, sino $C \leftarrow C \cup C_k$.

Figura 2-1 Seudocódigo del método de agrupamiento de artículos científicos basado en la extracción de frases relevantes.

2.3.1 Determinación de palabras centroides y asignación de documentos.

Para obtener las frases que son usadas como términos representativos para la conformación de los grupos, se intersectan los términos que se obtuvieron a partir de los títulos de los artículos pertenecientes a la colección (Figura 2-2b), con los términos relevantes que se obtienen a partir de los títulos de las referencias bibliográficas de todos los artículos (Figura 2-2a). Los términos relevantes de las referencias bibliográficas son aquellos que más se repiten entre los términos seleccionados como relevantes para cada artículo.

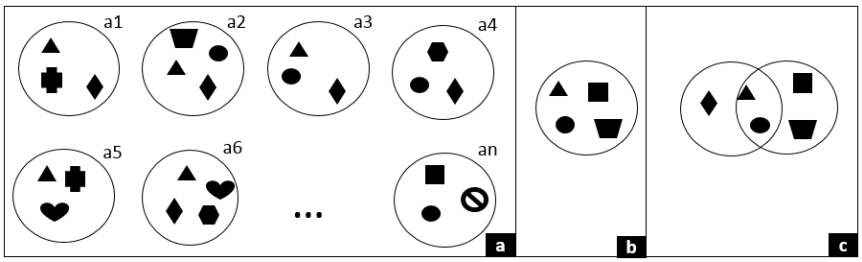


Figura 2-2 Selección de palabras consideradas términos representativos. a) palabras relevantes para cada documento seccionadas a partir de los títulos de las referencias. b) palabras relevantes de la colección seleccionadas a partir de los títulos de los artículos. c) intersección de las palabras relevantes de los títulos con las palabras relevantes de las referencias.

En la Figura 2-2c se aprecia que la intersección de las palabras relevantes de las referencias con las palabras relevantes de los títulos permite obtener finalmente los términos que serán los términos principales de los grupos a formar. Para formar los grupos preliminares se une a cada uno de estas palabras relevantes los documentos que las contengan. Esto indica que en este paso puede haber documentos que se agreguen a más de un grupo (Figura 2-3).

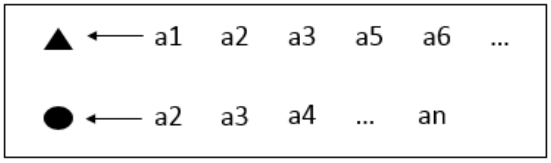


Figura 2-3 Ubicación inicial de los artículos en los grupos correspondientes basada en las palabras relevantes que estos contienen.

2.3.2 Eliminación del solapamiento.

A continuación, se analiza cuáles de estos grupos pueden ser unidos ya que se pueden tener varios términos representativos para un tema dado, esto significa que para cada uno de estos términos se obtiene un grupo de documentos, cuando en realidad si dos

términos pertenecen a un mismo tema, los documentos asociados a estos términos deben pertenecer a un mismo grupo.

Dos grupos vas a ser unidos si más del 50% de los documentos que pertenecen al grupo de menor tamaño entre ellos dos, están contenidos en el grupo de mayor tamaño. Luego, todos los documentos del grupo de menor tamaño que no pertenecían al grupo de mayor tamaño son asignados a este último y el menor de los grupos es eliminado. Dado que los centroides de los grupos preliminares que se conforman, son las palabras más representativas de la colección, y se asigna un documento a un grupo si contiene la palabra centroide del grupo, se garantiza que en la asignación inicial todos los documentos que pertenezcan a un mismo grupo estén conectados entre sí, ya que tiene en común al menos la palabra representativa del grupo. Está claro que el aumento de la cantidad de elementos solapados para un par de grupos aumenta la probabilidad de que los documentos pertenecientes a estos grupos traten un mismo tema. En este caso particular, donde se garantiza que todos los elementos del grupo están conectados entre sí, esta probabilidad se hace mayor, ya que se tiene asegurado que los elementos que no pertenecen al solapamiento están conectados con estos. Es por ello que se considera que, si el 50% de los elementos de un grupo solapa a otro grupo, estos dos grupos tratan el mismo tema. Este tema fue representado en primera instancia por dos grupos ya que dos de las palabras representativas de la colección pertenecían a este tema.

Luego del proceso de unión de grupos pueden persistir documentos que pertenezcan a más de un grupo. Para eliminar el solapamiento de los grupos se recurre al grafo de conexiones de los documentos. Se extraen de los grupos los elementos que provocan solapamiento y se analiza la pertenencia promedio de estos elementos a cada uno de los grupos que solapan. La pertenencia promedio de un elemento *i* a un grupo C_j se calcula como sigue:

$$PertProm(i, C_j) = \frac{\sum_{k=1}^{|C_j|} M(i, C_{jk})}{|C_j|} \tag{2.1}$$

Donde $|C_j|$ representan la cantidad de elementos del grupo C_j y *M* es la matriz de adyacencia que representa el grafo de conexiones entre los artículos.

2.3.3 Eliminación de grupos de tamaño reducido.

El siguiente paso en el método propuesto consiste en eliminar los grupos de tamaño reducido. El tamaño adecuado del clúster evidentemente depende de diversos factores como pueden ser: el tamaño de la colección, la cantidad de clases que esta presenta e incluso las necesidades de información del usuario específico que use el método de agrupamiento. Es por ello que el valor tomado como umbral para determinar si un grupo no tiene el tamaño adecuado se considera debe ser proporcionado por el usuario, y por ende este constituye un parámetro de entrada del algoritmo. La eliminación de un grupo pequeño consiste en reasignar cada elemento perteneciente a este grupo, al grupo que mayor pertenencia tenga, considerando como valor de pertenencia la cantidad de documentos del grupo destino con los cuales está conectado el documento que será reasignado.

2.3.4 Asignación de documentos aislados.

En este punto del algoritmo solo faltan por agrupar aquellos documentos que no contienen ninguna de las palabras que fueron seleccionadas como términos representativos de la colección, y que, en la primera etapa del algoritmo, sirvieron para formar los clústeres preliminares. Para ello se considera este conjunto de documentos como un grupo y se analiza la conexión total que presenta cada uno de estos documentos con respecto a este grupo y con respecto al resto de los grupos. El documento permanece en el grupo si la conexión total con respecto a este grupo es mayor que la conexión total con respecto a cada uno de los restantes grupos. En caso contrario el documento es reasignado al grupo con respecto al cual tiene mayor conexión total. Si al terminar el proceso de asignación de documentos aislados, la cantidad de documentos presentes en este grupo es mayor que la longitud mínima de clúster definida por el usuario este conjunto de documentos pasa a formar parte del conjunto de grupos.

2.4 Complejidad del agrupamiento

El método de agrupamiento propuesto en este capítulo consta de dos partes fundamentales: (I) la extracción de términos y representación de los documentos y (II) el agrupamiento de los artículos. Para el análisis de la complejidad computacional del método propuesto se analizan los pasos del cuatro al nueve del algoritmo, que son los

referidos al proceso de agrupamiento. Si para el proceso de agrupamiento se obtiene una complejidad que no supere la de otros algoritmos de agrupamiento, se considera entonces que la aplicación del método propuesto en esta investigación es factible. Esto se puede afirmar, dado que la extracción de términos y la representación de los documentos es más eficiente en el método propuesto por el hecho que solo se procesa una pequeña parte del artículo. Además, no se calcula el peso asociado a cada uno de los términos extraído, lo cual también complejiza la representación.

Para el cálculo de la complejidad computacional se establece *n* como la cantidad de artículos a agrupar y *k* la cantidad de frases representativas de los temas de la colección, estas son en primera instancia los centroides de los grupos preliminares. Además, es válido aclarar que en el proceso de asignación de artículos a los grupos se tiene una matriz de *k*x*k*, donde se va a llevar para cada par de grupos la cantidad de elementos solapados que presentan. Esta se utiliza como estructura de datos auxiliar para no tener que contar en los pasos de eliminación del solapamiento, la cantidad de elementos solapados. De este modo se obvian en este paso el análisis de los grupos que no presenta solapamiento.

Se tiene entonces que la complejidad de la asignación de los elementos a los grupos es $n\log(n)\,k^2\log(k)$. La eliminación del solapamiento se realiza en los pasos cinco y seis, la complejidad para el peor de los casos de estos pasos es $n^2k\log(k)$. La eliminación de clústeres con tamaño menor igual que el definido por el usuario como tamaño mínimo, supone una complejidad para el peor de los casos de $\frac{n^2}{k}\log k$. Para el paso de asignación de los elementos aislados se tiene una complejidad para el peor de los casos de n^2k. De aquí se asume que la complejidad para el agrupamiento para el peor de los casos es $n^2k\log(k)$. Esta complejidad se puede aproximar a n^2, dado que, la cantidad de grupos preliminares normalmente es un valor cercano a la cantidad de clases de la colección, y esta a su vez, es mucho menor que la cantidad de documentos. Por lo cual se puede asumir la complejidad del agrupamiento para el método propuesto como n^2.

2.5 Conclusiones parciales

- Se desarrolló un algoritmo para la extracción de frases relevantes en los títulos y las referencias bibliográficas de los artículos. Esto permite la representación del artículo mediante frases que engloban el tema general que trata el mismo.
- Se implementó un método de agrupamiento de artículos científicos basado en la representación de los mismos mediante frases relevantes. Este método tiene la peculiaridad de formar los grupos teniendo en cuenta: la conexión que tienen los artículos con las frases más representativas de la colección y las conexiones que presentan estos entre sí de acuerdo a las frases relevantes que comparten.
- La complejidad computacional del proceso de agrupamiento en el método propuesto, se comporta a la altura de otros algoritmos de agrupamientos reportados en la literatura. Esto garantiza que el método propuesto sea más eficiente que otros métodos de agrupamiento de artículos científicos, ya que gana en eficiencia en la extracción de términos y en la representación de los documentos al no tener que procesar todo el texto.

3

EVALUACIÓN DEL MÉTODO PROPUESTO

Capítulo 3. Evaluación del método propuesto

Este capítulo se dedica a la validación y verificación del método de agrupamiento propuesto. Primeramente, se hace un análisis para validar los valores seleccionados para los parámetros de configuración del algoritmo, así como para demostrar la eficacia de las unidades estructurales seleccionadas para la extracción de frases relevantes. Luego, se aplican un conjunto de métricas que permiten evaluar la calidad del agrupamiento resultante al aplicar el método propuesto.

3.1 Descripción de los casos de estudio

Como casos de estudio se usaron un total de 243 artículos científicos. Estos artículos provienen principalmente de tres fuentes, estas son: el sitio ICT[1] perteneciente al Centro de Investigaciones de la Informática, artículos recuperados de diversas revistas pertenecientes a la colección *Springer* y artículos de la revista Biotecnología Vegetal[2], que se edita en el Instituto de Biotecnología de las Plantas. De todos los artículos usados como casos de estudio se cuenta con una clasificación de referencia. Con la finalidad de disponer de diversidad en cuanto a las áreas de las ciencias de las que provienen los documentos tomados se seleccionaron artículos pertenecientes a tres áreas generales del conocimiento, estas son las ciencias técnicas, las ciencias sociales, y las ciencias biológicas. Dentro de cada una de estas áreas se cuenta con diferentes clases de referencias cada una representada por un conjunto variable de documentos. En total se cuenta con 16 clases de referencia. Además, fueron seleccionados documentos escritos tanto en español como en inglés con la finalidad de demostrar que el método de agrupamiento que se propone en este trabajo es aplicable también a artículos escritos en idioma español. En el Anexo 1 se muestra una descripción detallada de la clasificación de cada uno de los documentos.

[1] ftp:// ict.cei.uclv.edu.cu

[2] www.revista.ibp.co.cu

3.2 Validación de los valores seleccionados para el umbral fap

En la ecuación 2.1 se definió la forma para calcular el umbral *fap*. Para arribar a estos valores de umbral se realizó el siguiente análisis.

Se tomaron 100 artículos distribuidos entre las tres ramas del conocimiento con que se cuenta en los casos de estudio. Para cada uno de estos artículos se contabilizó la cantidad de veces que aparecía, en las referencias bibliográficas, la frase relevante fundamental (FRF) que representa el tema general del cual trata el artículo científico. Por ejemplo, para los artículos que tratan sobre estrés oxidativo, la frase relevante *oxidative stress* es la que mayoritariamente se debe encontrar y por tanto la que se considera como FRF. Luego se procedió a comparar la relación existente entre la cantidad de veces que aparece la frase relevante fundamental para cada artículo con la cantidad de referencias bibliográficas del artículo. Para ello, se calculó la correlación lineal entre estas dos variables, con la finalidad de verificar si es posible predecir a partir de la cantidad de referencias bibliográficas del artículo la cantidad de veces que debe aparecer la frase relevante fundamental. Los resultados de este experimento se muestran en la Figura 3-1.

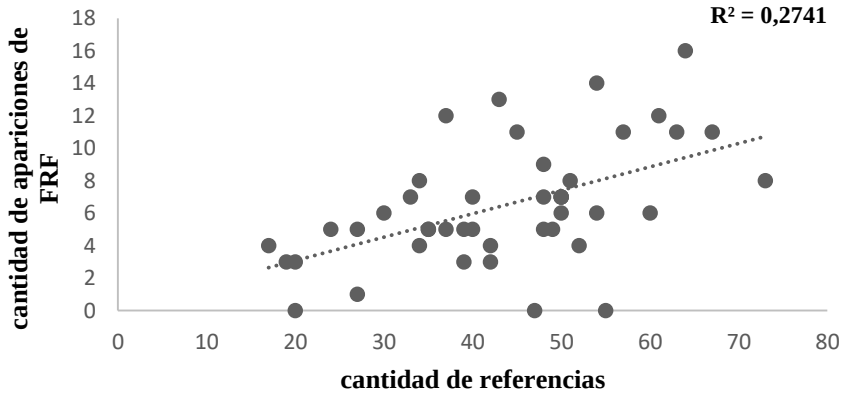


Figura 3-1 Análisis de correlación entre la cantidad de referencias bibliográficas y la cantidad de apariciones de la palabra relevante fundamental.

Como se puede observar en el gráfico anterior, no es posible establecer una correlación entre las variables cantidad de referencias bibliográficas y cantidad de apariciones de la frase relevante fundamental, esto se evidencia en el valor de R^2 el cual es bastante cercano a cero.

Es válido aclarar también, que la cantidad de referencias bibliográficas es considerablemente variable en dependencia del área del conocimiento. Si bien las referencias de los artículos científicos del área de ciencias técnicas se mantienen en un rango entre 15 y 30 referencias, e incluso se puede afirmar que la mayoría no sobrepasa las 25 referencias; los artículos pertenecientes a las ciencias biológicas cuentan con un mayor número de referencias, sobre todo si son específicamente artículos relacionados con la medicina, los cuales presentan en muchos casos más de 100 referencias bibliográficas. Esto mismo pasa con la mayoría de los artículos del área de las ciencias sociales. Esta particularidad permite suponer que, la correlación que pudiera existir entre las dos variables sometidas a evaluación puede afectarse por este rango tan amplio en que se mueve la cantidad de referencias bibliográficas. Por lo cual se procedió a analizar la correlación de estas dos variables en cada una de las áreas de conocimiento. Como resultado se obtuvo que el R^2 para los documentos del área de ciencias técnicas fue de 0.31, para las ciencias biológicas fue de 0.27 y para las ciencias sociales fue de 0.28. Esto demuestra que incluso no existe correlación entre la cantidad de referencias y la cantidad de veces que aparece la frase relevante fundamental del tema que trata el artículo, tomando los artículos por áreas de conocimiento.

Es evidente que no se puede establecer un valor único como umbral para la selección de las frases relevantes en los documentos de manera independiente a la cantidad de referencias. De esta forma, si se establece un valor pequeño como umbral los documentos que presenten un elevado número de referencias van a aportar una serie de términos que realmente no son frases relevantes del tema que trata el documento, pero, al aumentar la cantidad de referencias aumenta su probabilidad de ocurrencia. Por otra parte, si se establece un valor relativamente grande como umbral, los documentos que presentan pocas referencias se afectarán dado que resulta poco probable que en un número reducido de referencias puede ocurrir varias veces un término dado.

Es a partir de este análisis previo, que se decide dividir en cuatro intervalos la cantidad de referencias y asociar la cantidad de veces que aparece la frase relevante fundamental de cada artículo, a cada uno de estos intervalos. Por lo cual se dispone de una matriz, donde el valor *k* en la fila *i* columna *j* indica que existen *k* documentos que tienen una

cantidad de referencias que se encuentra en el intervalo representado por la fila *i* y donde la cantidad de veces que aparece la frase relevante fundamental es igual al valor asignado a la columna *j*.

Se probaron seis combinaciones diferentes de umbrales (Tabla 3-1), para ver con cuál de ellas se obtenía un mayor cubrimiento de los artículos analizados y por tanto ser seleccionada como umbral *fap*. Cada valor está relacionado con la cantidad de referencias, así en la primera combinación el valor dos se aplica para los artículos que contiene entre 1 y 10 referencias, el valor 3, para los que contienen entre 11 y 20, así sucesivamente. El cubrimiento significa que, con esta combinación de valores para el umbral, la palabra relevante fundamental para el documento se selecciona como frase representativa del documento.

Como se puede observar en la Tabla 3-1 para la primera combinación se obtiene un cubrimiento de 81.11% (de los 100 documentos analizados se obtuvo la palabra relevante fundamental para 81 de ellos) que resulta ser el mayor, lo cual es lógico debido a que presenta valores menos restrictivos. Aun así, la selección de esta combinación para el umbral *fap* se justifica ya que el cubrimiento difiere significativamente con respecto a las demás propuestas. Esto significa que si el valor de cubrimiento, por ejemplo, para la segunda combinación hubiese estado cercano al de la primera combinación, podía haber sido seleccionada la segunda, debido a que es más restrictiva y por tanto discrimina mejor a la hora de seleccionar las frases relevantes acorde a la cantidad de referencias.

Tabla 3-1 Cantidad de artículos para los cuales se obtiene la frase relevante fundamental en dependencia del umbral seleccionado.

umbral / cantidad referencias	2-3-4-5	3-4-5-6	4-5-6-7	2-4-6-8	3-5-7-9	4-6-8-10
1 a 10	5	4	3	5	4	3
11 a 20	21	17	13	17	13	10
21-a 30	14	12	8	8	5	1
31 o +	44	39	34	28	22	20
Cubrimiento	81.11%	68.25%	52.05%	59.42%	44.41%	30.66%

3.3 Factibilidad de la aplicación del método propuesto en artículos escritos en idioma español

Es conocido que la mayor parte de la literatura científica que se encuentra en internet está escrita en idioma inglés. Sin embargo, no es menos cierto que en los últimos años, con el surgimiento y desarrollo de bases de datos internacionales que indexan revistas de acceso abierto, han surgido muchas revistas que publican sus artículos en el idioma oficial del país en que se editan, gran cantidad de ellas son revistas iberoamericanas. Es por ello que resulta útil demostrar que el método propuesto puede usarse en revistas publicadas en idioma español.

Está claro que los métodos convencionales de agrupamiento de documentos, basados en la representación de los mismos mediante el modelo espacio vectorial, no son aplicables, en la mayoría de los casos, a textos que no sean escritos en idioma inglés. Esto está dado porque los métodos de selección y filtrado de términos disponibles están diseñados para trabajar con el idioma inglés, por poseer este una gramática más simple y más fácilmente adaptable al procesamiento automático. En el caso del método que se propone, el idioma no constituye una barrera para la aplicación del mismo. La mayoría de la bibliografía que consultan los autores generalmente se encuentra en idioma inglés. Además, la mayoría de las revistas exigen que los artículos cuenten tanto con el título en el idioma que se edita la revista como con el título en inglés. Por tanto, la selección de términos para la representación del artículo es factible incluso en aquellos que no son escritos en idioma inglés.

Para demostrar la afirmación anterior se seleccionaron 56 artículos de la colección tomada como caso de estudio, los cuales están escritos en idioma español. De estos se calculó la cantidad de referencias que cada uno de ellos tiene en inglés y la cantidad que tiene en español. Se aplicó el test no paramétrico de Wilcoxon para ver si existen diferencias significativas entre la cantidad de referencias en cada uno de los idiomas. Se obtuvo una significación asintótica menor que 0.05 lo cual indica que existen diferencias entre las muestras comparadas como se observa en la Figura 3-2. En esta figura se puede observar que para un promedio de 26 referencias por artículo 19 de ellas están en idioma inglés y solo 7 están en idioma español.

Si se hace un análisis más enfocado a cada artículo en particular, de los 56 artículos tomados como muestra, solo ocho de ellos presentaron más referencias en español que en inglés. De estos ocho, solo cuatro presentaron un desbalance desproporcionado que favoreció a las referencias en español, o lo que es lo mismo, la cantidad de referencias en español fue considerablemente mayor que la cantidad de referencias en inglés solo en cuatro de los 56 artículos sometidos a prueba.

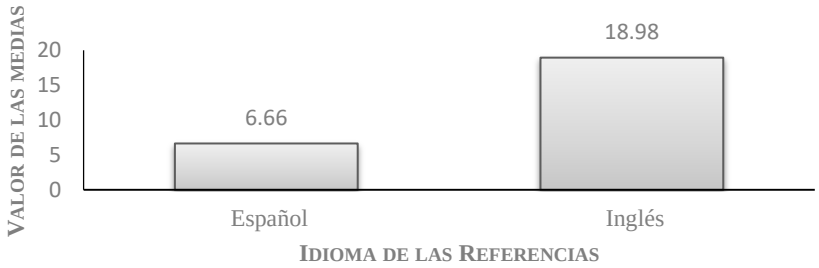


Figura 3-2 Comparación de las medias de las referencias en español y las referencias en inglés.

3.4 Justificación de las unidades estructurales tomadas para la selección de las frases relevantes

Para validar la selección de las unidades estructurales título y referencias bibliográficas, se procedió a la conformación de 10 corpus con artículos de los cuales se dispone del título, autores, resumen, contenido y referencias bibliográficas. Se tomaron los corpus 1, 2, 3, 4, 5, 7, 8, 22, 23 y 24 cuya descripción se muestra en el Anexo 2. El experimento realizado consistió en la aplicación, a cada uno de los corpus, de los algoritmos de agrupamiento k-Means y la modificación del algoritmo k-Star propuesta por Pinto y colaboradores en (Pinto et al., 2010) a la cual se hace referencia a partir de este punto como k-XStar.

Para cada corpus se aplicó la selección de términos en cada una de las unidades estructurales y a cada una de estas representaciones se le aplicaron ambos algoritmos de agrupamiento. Además, se aplicó la función OverallSimSux propuesta en (Magdaleno, 2015) a cada uno de los corpus. Esta función en una primera etapa agrupa los artículos teniendo en cuenta cada unidad estructural de manera independiente, luego fusiona los resultados obtenidos para cada agrupamiento para obtener un agrupamiento final. Los autores de esta función demostraron que el uso de la misma garantiza mejores resultados

en el agrupamiento de artículos científicos que considerar el artículo como una bolsa de palabras. Por tanto, se toman los resultados obtenidos por esta función como resultados de referencia y se considera que los agrupamientos que se asemejen a los mismos son resultados favorables. En el caso del algoritmo de agrupamiento k-Means se fijó para cada corpus la cantidad de grupos como la cantidad de clases que presenta el corpus.

Para medir la calidad del agrupamiento se aplicaron métricas de validación externa, dado que se cuenta con la clasificación de referencia de cada artículo sometido al proceso de agrupamiento. Los índices de validación seleccionado fueron: la Entropía, el índice Folkes-Mallows, Micro Purity, Macro Purity, Precisión, Cubrimiento y la medida F basada en la precisión y el cubrimiento.

Al aplicar la prueba no paramétrica de Friedman a los valores obtenidos para la entropía con el algoritmo k-Means, se obtuvo que existen diferencias significativas entre las muestras comparadas (Anexo 3, Figura A-1). Luego se aplicó el test de Nemenyi para determinar entre qué pares de unidades estructurales existen diferencias significativas. Como se puede ver en la Figura 3-3 existen diferencias para la función OverallSimSux con respecto al resumen y a los autores, pero no existe diferencia de esta con respecto a las referencias, el título y el cuerpo. Esto sugiere, que para las unidades estructurales título, referencias bibliográficas y cuerpo se obtienen resultados similares en el agrupamiento que si se considera todo el documento mediante la función OverallSimSux. El test de Nemenyi no arrojó diferencias significativas entre las referencias bibliográficas con respecto al resumen. Sin embargo, al aplicar el test no paramétrico de Wilcoxon para comparar la unidad estructural referencias bibliográficas con el resto de las unidades y con los resultados que se obtuvieron para la función OverallSimSux se evidencia que, en efecto, no existen diferencias entre los resultados obtenidos para esta unidad y para la función OverallSimSux, para el título y para el cuerpo, pero sí hay diferencias de esta unidad estructural con respecto al resumen. En el caso de la comparación del título con el resto de las unidades estructurales no se evidencian diferencias con respecto al cuerpo del documento, pero sí hay diferencias con respecto al resto de las unidades estructurales y a la función OverallSimSux (Anexo 4, Figura A-5).

Un comportamiento similar al anterior se presenta para la medida F y para el índice Folkes-Mallows (Anexo 4, Figura A-6 y Figura A-7). En el caso de los índices Micro Purity, Macro Purity, Precisión y Cubrimiento en general no se obtuvieron diferencias significativas entre las muestras comparadas. Solo se presentaron diferencias entre la unidad estructural autores y el resto de las unidades (Anexo 4, Figura A-8).

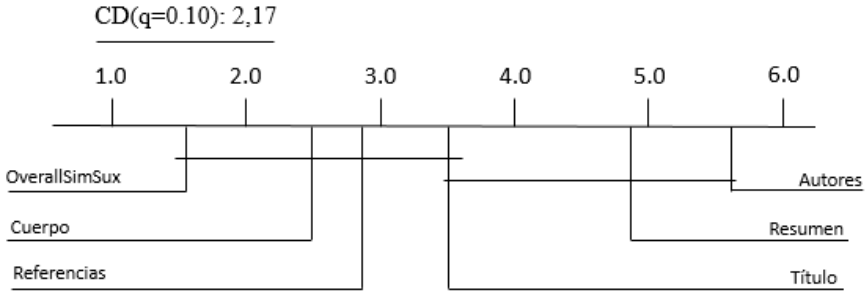


Figura 3-3 Test de Nemenyi para los valores de Entropía obtenidos al aplicar el algoritmo k-Means a cada una de las unidades estructurales.

Los resultados obtenidos en este experimento demuestran que, es posible obtener resultados estadísticamente similares en el agrupamiento de artículos científicos mediante la extracción de términos en las referencias bibliográficas que aplicando la función OverallSimSux. También es válido aclarar que estos resultados no difieren estadísticamente de los resultados obtenidos para el cuerpo del documento. Sin embargo, dada la cantidad de términos que presenta el cuerpo del documento es computacionalmente más factible extraer solo términos de las referencias bibliográficas. Tampoco existen diferencias entre las referencias y el título y entre este último y el cuerpo, aunque sí se presentan diferencias del título con respecto a la función OverallSimSux.

Los resultados descritos anteriormente se obtuvieron con el algoritmo k-Means. Para cada una de las corridas se proporcionó como cantidad de grupos a obtener, la cantidad de clases con las que cuenta el corpus. Esto provoca que los resultados del agrupamiento sean más precisos que los resultados que se obtienen, si se aplica un algoritmo de agrupamiento donde la cantidad de grupos esperados se desconozca y sea este mismo el encargado de descubrirla. Esto es lo que se presenta en la práctica, por lo cual se realizó el mismo proceso de validación, pero con el uso del algoritmo k-XStar.

Al aplicar el test de Friedman se obtuvieron diferencias significativas entre las muestras comparadas para cada una de las métricas sometidas a evaluación (Anexo 5, Figura A-9, Figura A-10 y Figura A-11). Al aplicar el test de Nemenyi se obtiene que de manera general no se presentan diferencias significativas entre las unidades estructurales para las medidas Micro Purity, Macro Purity y Precisión. Solo se presentan diferencias de la unidad autores con el resto de las unidades estructurales, siendo estas diferencias siempre a favor de las demás unidades estructurales. En el caso de las métricas Cubrimiento, medida F e índice Folkes-Mallows se presentan diferencias significativas de las unidades título, referencias y cuerpo con respecto a la unidad autores (Figura 3-4, Figura 3-5 y Figura 3-6).

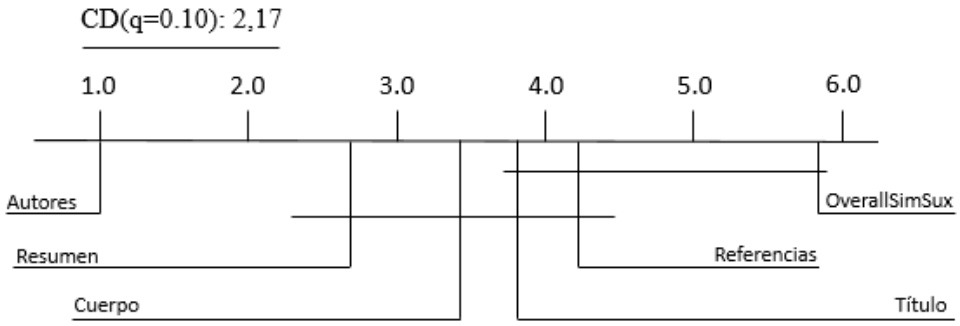


Figura 3-4 Test de Nemenyi para los valores de Cubrimiento obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

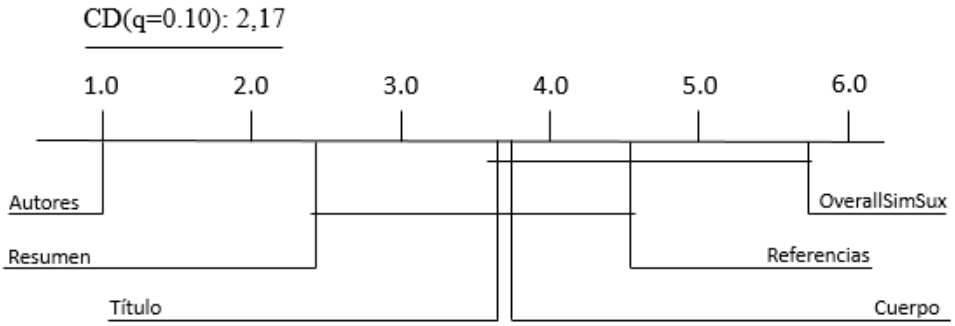


Figura 3-5 Test de Nemenyi para los valores del índice Folkes-Mallows obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

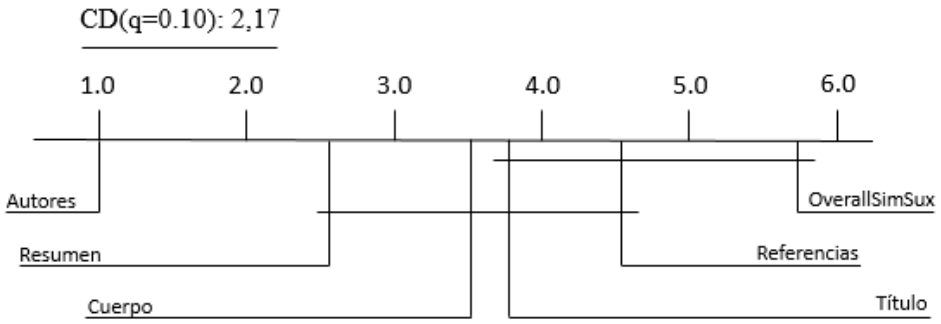


Figura 3-6 Test de Nemenyi para los valores de la medida F obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

Para la métrica Entropía se presentan diferencias de las unidades referencias, cuerpo y la función OverallSimSux con respecto al resumen (Figura 3-7). Para ninguna de las métricas se obtuvo diferencias significativas entre la función OverallSimSux y el título, la función OverallSimSux y las referencias y solo se obtuvo entre la función OverallSimSux y el cuerpo para la medida F. Sin embargo, al aplicar el test de Wilcoxon para comparar las unidades título y referencias bibliográficas con el resto de las unidades se obtiene que sí existen diferencias entre la función OverallSimSux y el título para las métricas Cubrimiento, Entropía, medida F e índice Folkes-Mallows. Además, se presentan diferencias entre las referencias y la función OverallSimSux para las métricas Cubrimiento y medida F (Anexo 6Anexo 6. Resultados del test no paramétrico de Wilcoxon en la aplica

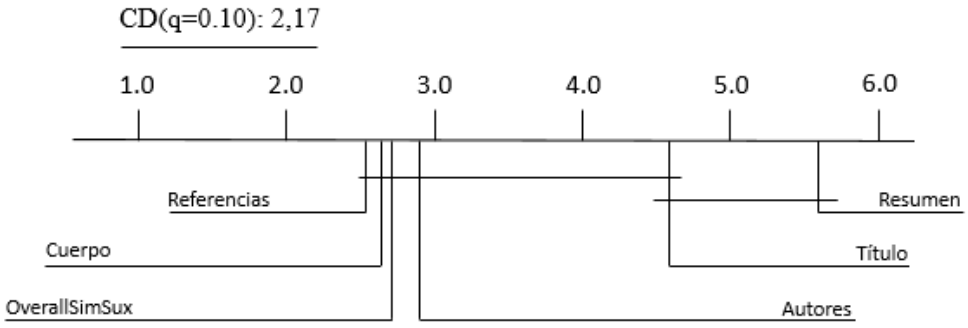


Figura 3-7 Test de Nemenyi para los valores de Entropía obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

Esta segunda parte del experimento demuestra que, para el título, el cuerpo y las referencias bibliográficas también se obtienen mejores resultados en el agrupamiento que para el resto de las unidades, al aplicar un algoritmo donde la cantidad de grupos debe ser determinada en el proceso de agrupamiento. Sin embargo, es válido aclarar que las diferencias obtenidas para algunas de las métricas sugieren que, a pesar que las

referencias bibliográficas y el título proveen mejores resultados que el resto de las unidades estructurales, no es totalmente factible usar estas unidades de manera aislada en la selección de términos para la representación de artículos científicos.

3.4.1 Comparación del uso de todas las partes de las referencias bibliográficas con el uso de solo el título

En este subepígrafe se compara la factibilidad de usar solo el título de las referencias bibliográficas en la extracción de términos y no todas las partes de la referencia. Esto reduce aún más el espacio de búsqueda en la selección de términos.

En este experimento, al igual que en el anterior, se aplicaron los algoritmos de agrupamiento k-Means y k-XStar a los 10 corpus tomados como casos de estudio. La extracción de términos se realizó con todo el texto contenido en las referencias y con solo el texto conformado por los títulos de las mismas. Se compararon los resultados obtenidos en cuanto a los siete índices de validación externos tomados como referencia. Las Figura 3-8 y Figura 3-9 muestran los resultados para k-Means y k-XStar respectivamente.

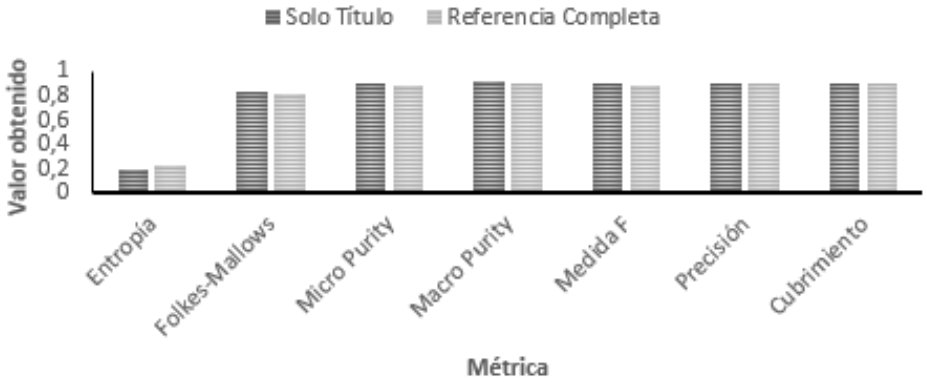


Figura 3-8 Comparación de las medias para cada una de las métricas evaluadas al aplicar el algoritmo k-Means tomando todo el texto de las referencias y tomando solo el título.

Al comparar los resultados obtenidos para cada una de las métricas mediante el test de Wilcoxon, se evidenció que no existen diferencias significativas entre las muestras comparadas (Anexo 8). En las Figura 3-8 y Figura 3-9 se puede observar la comparación de las medias para cada una de las métricas sometidas a evaluación con ambos algoritmos. No existen diferencias significativas entre la selección de términos en todo el texto de las referencias con la selección solo teniendo en cuenta el título de las misma.

Es por ello que no es necesario considerar todo el texto de la referencia para la extracción de términos, sino solamente el título de la misma.

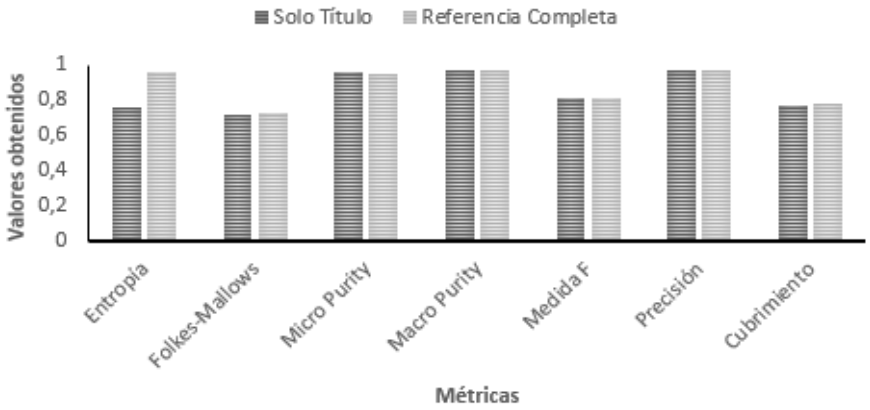


Figura 3-9 Comparación de las medias para cada una de las métricas evaluadas al aplicar el algoritmo k-XStar tomando todo el texto de las referencias y tomando solo el título.

3.5 Comparación del método de agrupamiento propuesto con otros métodos para el agrupamiento de artículos científicos reportados en la literatura

Como se mencionó anteriormente la función OverallSimSux reporta resultados favorables en el agrupamiento de artículos científicos. Esta función hace uso de todo el contenido del documento para obtener los términos que representan al mismo. Es por ello que, si el método de agrupamiento basado en la extracción de frases relevantes no difiere estadísticamente de los resultados obtenidos por la función OverallSimSux, entonces es más eficaz la aplicación del primero, dado que este solamente usa el título y los títulos de las referencias bibliográficas del artículo.

Para comprobar la hipótesis anterior, se compararon los resultados obtenidos para la función OverallSimSux con los algoritmos de agrupamiento k-Means y k-XStar, con los resultados obtenidos por el método de agrupamiento propuesto en esta investigación. Además, se comparó el método basado en frases relevantes con los resultados obtenidos por la función SimRefBib (Amador Penichet et al., 2018) diseñada para el agrupamiento de artículos científicos, basado en la información brindada por las referencias. Esta función además de los títulos de las referencias, hace uso de los autores y de los años de las mismas, lo que conlleva a un aumento de la complejidad computacional en la extracción de los términos y la representación de los documentos comparado con el método propuesto en esta investigación.

Al aplicar la prueba no paramétrica de Friedman a los valores obtenidos para las siete métricas externas que se evalúan se obtuvo que no existen diferencias significativas entre las muestras comparadas para las medidas Micro Purity, Macro Purity, Precisión y Entropía (Anexo 9, Figura A-23 y Figura A-24). Esto es un resultado favorable ya que para estos cuatro índices podemos garantizar que los resultados obtenidos por el método propuesto no difieren estadísticamente de los resultados obtenidos por la función OverallSimSux y la función SimRefBib.

En el caso del índice Folkes-Mallows, la medida F y el Cubrimiento se encontraron diferencias significativas entre las muestras comparadas (Anexo 9, Figura A-25). Al aplicar el test de Nemenyi se obtuvo que existen diferencias significativas entre los resultados obtenidos por el método propuesto con los resultados obtenidos por la función OverallSimSux cuando se usa el algoritmo de agrupamiento k-XStar y no se obtuvieron diferencias del método propuesto con la misma función cuando se aplica el algoritmo k-Means ni con respecto a la función SimRefBib (Figura 3-10, Figura 3-11 y Figura 3-12). Estos resultados son deseados, dado que para cada una de las corridas del algoritmo k-Means se fijó la cantidad de grupos como la cantidad de clases presentes en el corpus. Esto puede interpretarse como que el algoritmo k-Means, posee una ligera ventaja con respecto al método propuesto, aun así, los resultados del agrupamiento que logra la función no son mejores que los del método propuesto. Una comparación más balanceada es la que se estableció entre el método basado en frases relevantes con la función OverallSimSux usando el algoritmo de agrupamiento k-XStar y con la función SimRefBib, ya que en ambos casos la cantidad de grupos es determinada por el algoritmo de agrupamiento en el proceso iterativo. Como se observa en este caso, el método propuesto supera a la función OverallSimSux y se comporta sin diferencias con respecto a la función SimRefBib. Ambos casos constituyen resultados deseados. De hecho, se considera un buen resultado si el método propuesto se comporta sin diferencias con respecto a la función OverallSimSux dado que esta últimas hace uso de todo el documento para la extracción de los términos representativos mientras que el método propuesto solo usa el título de artículo y los títulos de las referencias. Esta misma condición es la que garantiza que los resultados obtenidos se consideren favorables también en a la comparación con la función SimRefBib.

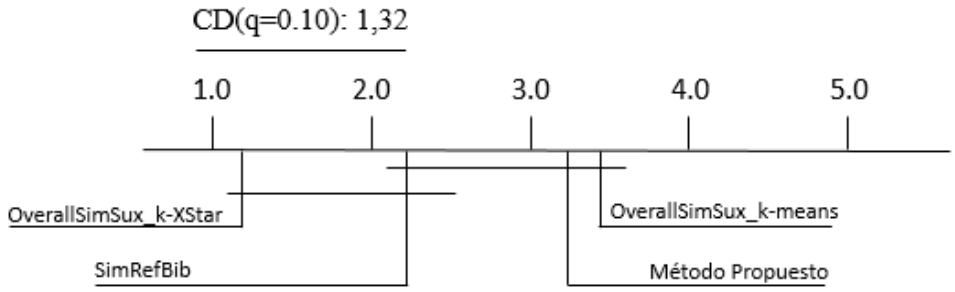


Figura 3-10 Test de Nemenyi para comparar el Método Propuesto con la función OverallSimSux y la función SimRefBib según los valores de Cubrimiento obtenidos.

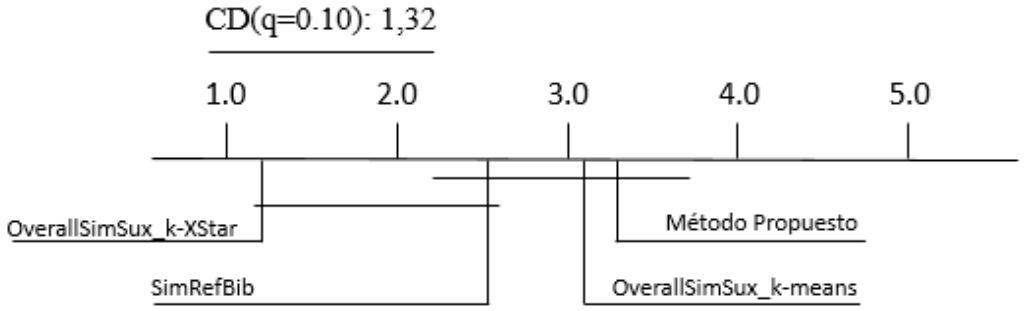


Figura 3-11 Test de Nemenyi para comparar el Método Propuesto con la función OverallSimSux y la función SimRefBib según los valores obtenidos para el índice Folkes-Mallows.

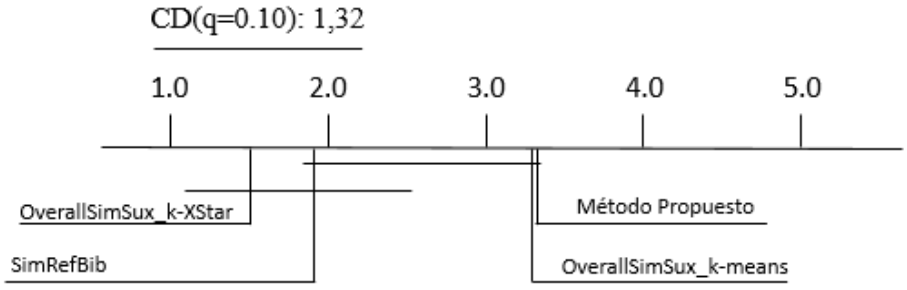


Figura 3-12 Test de Nemenyi para comparar el Método Propuesto con la función OverallSimSux y la función SimRefBib según los valores obtenidos para la medida F.

3.6 Validación de la eficacia del método propuesto en un conjunto de corpus con número variable de clases

Para medir la eficacia del método propuesto se conformaron 35 corpus con los artículos de la colección que se tomó como caso de estudio. La descripción de cada uno de estos corpus se muestra en el Anexo 2.

En la Figura 3-13 se muestra los valores alcanzados para la medida Entropía en los 36 corpus sometidos a agrupamiento.

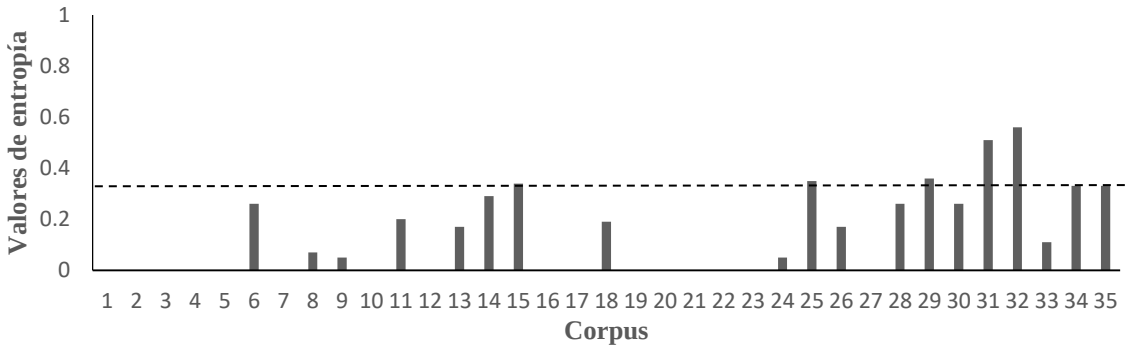


Figura 3-13 Valores de Entropía obtenidos al aplicar el método de agrupamiento basado en frases relevantes.

Como se puede observar los valores de entropía se mantienen por debajo de 0.35 para 33 de los 35 corpus sometidos a evaluación. Incluso para la mayoría de los corpus los valores se mantienen por debajo de 0.35. También se observa que para los últimos corpus es para los cuales se ven más afectados los valores de entropía. Estos corpus son los que presentan cuatro clases, lo cual puede ser un indicio de que con el aumento del número de clases aumente la entropía del agrupamiento.

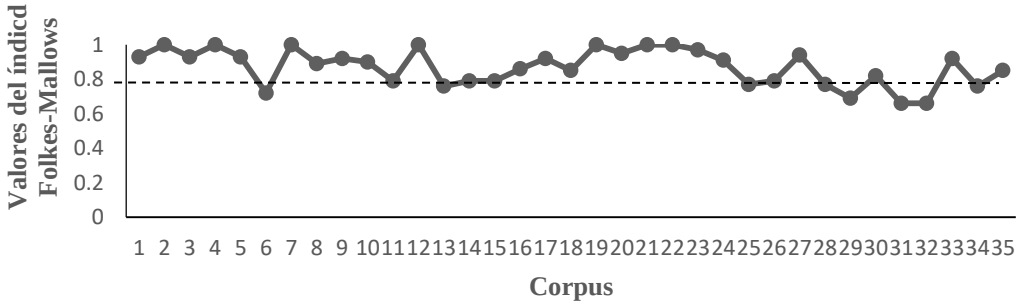


Figura 3-14 Valores para el índice Folkes-Mallows obtenidos al aplicar el método de agrupamiento basado en frases relevantes.

Los valores obtenidos para el índice Folkes-Mallows se presentan en la Figura 3-14. Para el índice Folkes-Mallows se obtienen valores por encima de 0.8 para la mayoría de los corpus sometidos a evaluación, lo cual indica que la calidad del agrupamiento es buena. Al igual que para la entropía, el índice Folkes-Mallows se afectó en los corpus que presentan cuatro clases.

En la Figura 3-15 se muestra el comportamiento de la medida F para los 35 corpus sometidos a evaluación. Como se puede observar los valores se mantiene por encima de

0.8 en 33 de los 35 corpus. Incluso para los corpus que presentan cuatro clases estos valores se mantienen por encima de 0.8.

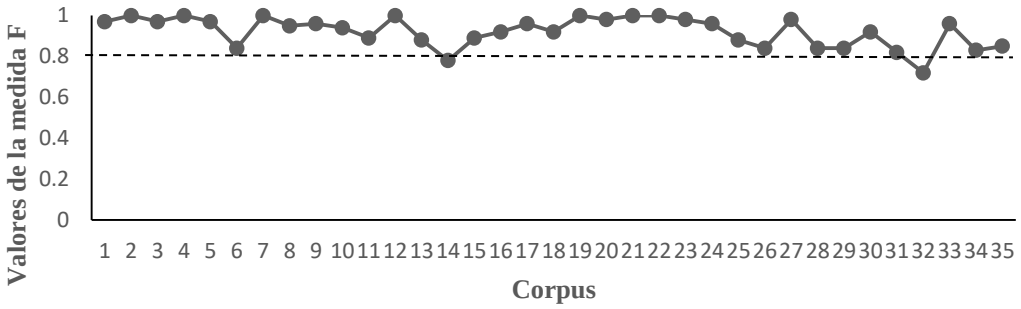


Figura 3-15 Valores obtenidos para la medida F al aplicar el método de agrupamiento basado en frases relevantes.

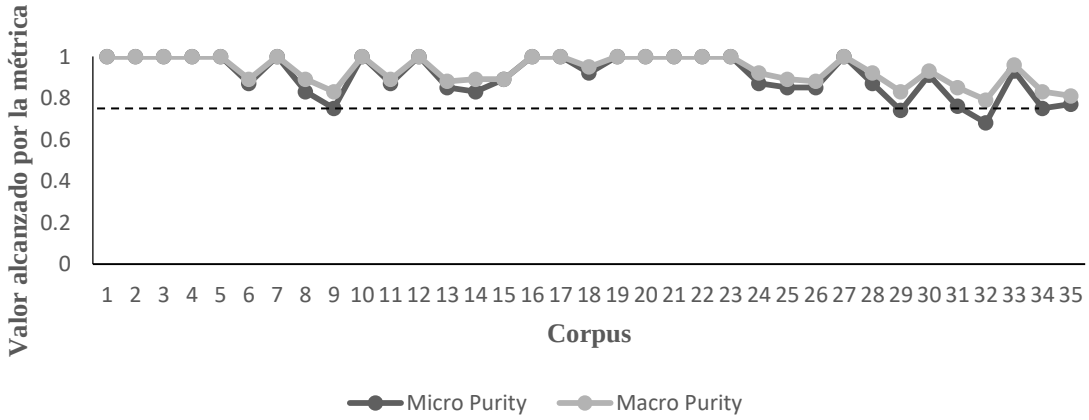


Figura 3-16 Valores obtenidos para las medidas Micro Purity y Macro Purity al aplicar el método de agrupamiento basado en frases relevantes.

En la Figura 3-16 se muestran los valores alcanzados para las métricas Micro Purity y Macro Purity. La mayoría de estos valores están por encima de 0.8 para ambas medidas, incluso para varios corpus se alcanza el valor máximo. Para los corpus con cuatro clases se nota una ligera disminución en los valores alcanzados para la medida Micro Purity.

En la Figura 3-17 se muestra los valores alcanzados para las medidas Precisión y Cubrimiento. Estos valores al igual que para las métricas anteriores se encuentran la mayoría por encima de 0.8, lo que refleja la alta semejanza de los grupos obtenidos con respecto a las clases de referencia.

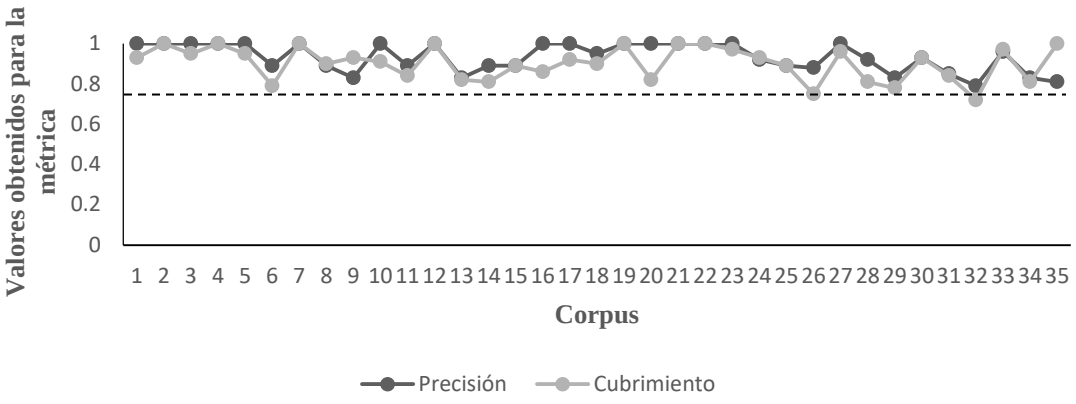


Figura 3-17 Valores obtenidos para las medidas Precisión y Cubrimiento al aplicar el método de agrupamiento basado en frases relevantes.

Como se puede observar en cada uno de los gráficos anteriores, en la mayoría de los corpus los valores que se obtuvieron para cada una de las métricas están por encima de 0.75. La disminución de la calidad del agrupamiento en los últimos corpus, manifestada en el descenso de los valores obtenidos para algunas de las medidas aplicadas, puede estar asociada al aumento del número de clases que conforman estos corpus. Aun así, estos valores se consideran aceptables. Además, es válido aclarar que no es común que las colecciones sometidas a procesos de agrupamiento contengan más de dos o tres clases. Lo común es que se tengan dos clases: (1) la clase a la que pertenecen los documentos que tratan el tema específico que es de interés para el usuario, (2) la clase de los documentos que a pesar de ser devueltos en el proceso de recuperación de información no están vinculados con el tema en cuestión que es de interés para el usuario. En última instancia esta segunda clase se puede ver con un conjunto de *k* subclases donde *k* es la cantidad de documentos que la integran. Sin embargo, considerarla de esta manera tampoco afecta el proceso de agrupamiento ya que se tiene un conjunto de documentos aislados, y un conjunto de documentos agrupados los cuales forman la clase que es en efecto de interés para el usuario.

3.7 Conclusiones parciales

- El estudio realizado en artículos científicos en idioma español demostró que el método propuesto es aplicable a artículos escritos en este idioma.

- La comparación del agrupamiento de artículos científicos, donde la selección de términos se realizó en cada unidad estructural de manera independiente, demostró que las referencias bibliográficas y el título aportan los términos más representativos en este tipo de documentos. Además, no existieron diferencias significativas entre tomar solo el título de cada referencia y tomar todo el contenido de las mismas, por lo cual la selección de términos en esta unidad estructural se puede reducir a que los términos se seleccionen únicamente en los títulos de las mismas.
- Se comparó el método de agrupamiento propuesto con la función OverallSimSux, para esta última se usaron los algoritmos k-Means y k-XStar. No se obtuvieron diferencias significativas entre el método propuesto y la función OverallSimSux con el algoritmo k-Means. Esto se considera un resultado satisfactorio dado que el método propuesto logra descubrir en el proceso iterativo la cantidad de grupos, mientras que el algoritmo k-Means requiere la cantidad de grupos como parámetro de entrada. En la comparación usando el algoritmo k-XStar sí se obtuvieron diferencias significativas para las métricas Cubrimiento, medida F e índice Folkes-Mallows. Estas diferencias fueron a favor del método propuesto en esta investigación.
- La aplicación del método de agrupamiento basado en frases relevantes, al conjunto de corpus tomado como caso de estudio demostró la factibilidad del mismo para el agrupamiento de artículos científicos.

CONCLUSIONES

- Del análisis de la literatura científica consultada se concluye que en la mayoría de los métodos de agrupamiento se hace uso de la co-citación de los artículos para obtener los grupos de documentos similares y los índices de co-citación no son libremente accesibles.
- La extracción de frases relevantes en las referencias bibliográficas y en los títulos de los artículos científicos, facilita la conformación de grupos similares sin tener que procesar todo el documento, por lo que se reduce la complejidad en el proceso de extracción de términos y la representación de los documentos.
- El método de agrupamiento propuesto, obtiene grupos similares a los que se alcanzan con la función SimRefBib y con la metodología OverallSimSux cuando a esta se le proporciona la cantidad de grupos esperados y supera los resultados alcanzados por la metodología cuando no se conoce la cantidad de grupos. Este método además reduce la complejidad computacional en la conformación de los grupos.
- La evaluación a través de los experimentos y los casos de estudios definidos, demostró que el método propuesto con frases relevantes es más eficiente que otros métodos de agrupamiento de artículos científicos reportados en la literatura. Al mismo tiempo la eficacia del mismo se comporta a la altura de estos métodos.

RECOMENDACIONES

- Aplicar el método de agrupamiento basado en la extracción de frases relevantes a artículos escritos en otros idiomas que no sea inglés, para analizar la factibilidad del uso del mismo en diversos idiomas.
- Adaptar el método propuesto para aplicarlo en el agrupamiento de Curriculum Vitaes

REFERENCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C. & ZHAI, C. 2012. A survey of text clustering algorithms. *In:* AGGARWAL, C. C. & ZHAI, C. (eds.) *Mining Text Data.* New york: Springer.

ALJABER, B., STOKES, N., BAILEY, J. & PEI, J. 2010. Document clustering of scientific texts using citation contexts. *Information Retrieval,* 13**,** 101-131.

AMADOR, L., GARCÍA, M. M., GÁLVEZ, D. & MAGDALENO, D. 2017. SemClustDML: algoritmo para agrupar artículos científicos basado en la información brindada por las referencias bibliográficas. *Revista Cubana de Ciencias Informáticas,* 11**,** 46-60.

AMADOR PENICHET, L., MAGDALENO GUEVARA, D. & GARCÍA LORENZO, M. M. 2018. New Similarity Function for Scientific Articles Clustering Based on the Bibliographic References. *Computación y Sistemas,* 22.

ARBELAITZ, O., GURRUTXAGA, I., MUGUERZA, J., PÉREZ, J. M. & PERONA, I. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition,* 46**,** 243-256.

ARCO GARCÍA, L. 2005. *Modelo para el agrupamiento de documentos afines y su ulterior resumen a través de la representación espacio vectorial de un corpus textual.* Universidad Central "Marta Abreu" de Las Villas.

ARCO, L., BELLO, R., MEDEROS, J. M. & PÉREZ, Y. 2006. Agrupamiento de documentos textuales mediante métodos concatenados. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial,* 10.

ARTHUR, D. & VASSILVITSKII, S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007. Society for Industrial and Applied Mathematics, 1027-1035.

BALA, C., BASU, T. & DASGUPTA, A. Automatic detection of k with suitable seed values for classic k-means algorithm using DE. Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on, 2015. IEEE, 759-765.

BANDYOPADHYAY, S. & SAHA, S. 2008. A point symmetry-based clustering technique for automatic evolution of clusters. *IEEE Transactions on Knowledge and Data Engineering,* 20**,** 1441-1457.

BANERJEE, A. & GHOSH, J. 2004. Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres. *IEEE Transactions on Neural Networks,* 15**,** 702-719.

BERNOTAS, M., KARKLIUS, K., LAURUTIS, R. & SLOTKIENĖ, A. 2015. The peculiarities of the text document representation, using ontology and tagging-based clustering technique. *Information Technology And Control,* 36.

BERRY, M. W. & CASTELLANOS, M. 2004. Survey of text mining. *Computing Reviews,* 45**,** 548.

BEZDEK, J. C., EHRLICH, R. & FULL, W. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences,* 10**,** 191-203.

BEZDEK, J. C., LI, W., ATTIKIOUZEL, Y. & WINDHAM, M. 1997. A geometric approach to cluster validity for normal mixtures. *Soft Computing,* 1**,** 166-179.

BEZDEK, J. C. & PAL, N. R. Cluster validation with generalized Dunn's indices. Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on, 1995. IEEE, 190-193.

BLOEHDORN, S., CIMIANO, P., HOTHO, A. & STAAB, S. An Ontology-based Framework for Text Mining. LDV Forum, 2005. 87-112.

BOYACK, K. W., SMALL, H. & KLAVANS, R. 2013. Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology,* 64**,** 1759-1767.

BRUN, M., SIMA, C., HUA, J., LOWEY, J., CARROLL, B., SUH, E. & DOUGHERTY, E. R. 2007. Model-based evaluation of clustering validation measures. *Pattern recognition,* 40**,** 807-824.

CALIŃSKI, T. & HARABASZ, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods,* 3**,** 1-27.

CASILLAS GARCÍA, J. E. 2015. Clustering semántico de textos.

CASTRO, A., SIFUENTES, E., GONZÁLEZ, S. & RASCÓN, L. H. 2014. Uso de Minería de Datos en el manejo de Información Geográfica. *Información tecnológica,* 25**,** 95-102.

CELEBI, M. E., KINGRAVI, H. A. & VELA, P. A. 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications,* 40**,** 200-210.

COBO, G., SEVILLANO, X., ALÍAS, F. & SOCORÓ, J. C. 2006. Técnicas de representación de textos para clasificación no supervisada de documentos. *Procesamiento del lenguaje natural,* 37.

COSTA, G. & ORTALE, R. 2018. Machine learning techniques for XML (co-) clustering by structure-constrained phrases. *Information Retrieval Journal,* 21**,** 24-55.

CHOU, C.-H., SU, M.-C. & LAI, E. 2004. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications,* 7**,** 205-220.

DESGRAUPES, B. 2013. Clustering indices. *University of Paris Ouest-Lab ModalX,* 1**,** 34.

DOMÍNGUEZ, Y. L., FUENTES, F. A., BRUZÓN, A. F. & BUENO, R. O. 2014. Optimizaciones al Algoritmo de Agrupamiento Compacto Jerárquico Dinámico. *Revista Cubana de Ciencias Informáticas,* 8**,** 59-65.

DUBES, R. C. 1987. How many clusters are best?-an experiment. *Pattern Recognition,* 20**,** 645-663.

FORSATI, R., MAHDAVI, M., SHAMSFARD, M. & MEYBODI, M. R. 2013. Efficient stochastic algorithms for document clustering. *Information Sciences,* 220**,** 269-291.

GARFIELD, E., MALIN, M. V. & SMALL, H. 2013. A system for automatic classification of scientific literature. *Journal of the Indian Institute of Science,* 57**,** 61.

GIL-GARCÍA, R. & PONS-PORRATA, A. 2010. Dynamic hierarchical algorithms for document clustering. *Pattern Recognition Letters,* 31**,** 469-477.

GONZÁLEZ-SOLER, L. J., PÉREZ-SUÁREZ, A. & CHANG-FERNÁNDEZ, L. 2015. Algoritmo incremental de agrupamiento con traslape para el procesamiento de grandes colecciones de datos (Overlapping clustering incremental algorithm for large data collections processing).

GUHA, S., RASTOGI, R. & SHIM, K. CURE: an efficient clustering algorithm for large databases. ACM Sigmod Record, 1998. ACM, 73-84.

HABIBI, M. & POPESCU-BELIS, A. 2015. Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* 23**,** 746-759.

HARTUV, E. & SHAMIR, R. 2000. A clustering algorithm based on graph connectivity. *Information processing letters,* 76**,** 175-181.

HERNÁNDEZ, J. K. F. 2016. *Descubrimiento de secuencias frecuentes y su aplicación a la clasificación de documentos,* Editorial Universitaria.

HOTHO, A., MAEDCHE, A., STAAB, S. & ZACHARIAS, V. On Knowledgeable Supervised Text Mining. Proceedings of Text Mining Workshop, Springer, 2002.

HOTHO, A., STAAB, S. & STUMME, G. 2003. Text clustering based on background knowledge. *Institute AIFB, Universität Karlsruhe*.

HU, C.-P., HU, J.-M., GAO, Y. & ZHANG, Y.-K. 2010. A journal co-citation analysis of library and information science in China. *Scientometrics,* 86**,** 657-670.

HUANG, A. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008. 49-56.

IEZZI, D. F. 2012. Centrality measures for text clustering. *Communications in Statistics-Theory and Methods,* 41**,** 3179-3197.

INGARAMO, D. A., ERRECALDE, M. L. & ROSSO, P. 2007. Medidas internas y externas en el agrupamiento de resúmenes científicos de dominios reducidos. *Procesamiento del lenguaje natural, nº 39 (sept. 2007); pp. 55-62*.

KARO, I. M. K., MAULANAADHINUGRAHA, K. & HUDA, A. F. A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function. Informatics and Computing (ICIC), 2017 Second International Conference on, 2017. IEEE, 1-6.

KARYPIS, G., HAN, E.-H. & KUMAR, V. 1999. Chameleon: Hierarchical clustering using dynamic modeling. *Computer,* 32**,** 68-75.

KIELA, D. & CLARK, S. A systematic study of semantic vector space model parameters. Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL, 2014. 21-30.

KIM, S. N., MEDELYAN, O., KAN, M.-Y. & BALDWIN, T. 2013. Automatic keyphrase extraction from scientific articles. *Language resources and evaluation,* 47**,** 723-742.

KUMAR, N. & SRINATHAN, K. Automatic keyphrase extraction from scientific documents using N-gram filtration technique. Proceedings of the eighth ACM symposium on Document engineering, 2008. ACM, 199-208.

KUMAR, S., GAO, X. & WELCH, I. Learning Under Data Shift for Domain Adaptation: A Model-Based Co-clustering Transfer Learning Solution. Pacific Rim Knowledge Acquisition Workshop, 2016. Springer, 43-54.

KUNA, H., REY, M., MARTINI, E., RAMBO, A. & PODKOWA, L. 2015. Avances en el Desarrollo de un Sistema de Recuperación de Información para Publicaciones Científicas del Área de Ciencias de la Computación. *Revista Latinoamericana de Ingenieria de Software,* 3**,** 47-55.

LANQUILLON, C. 2001. *Enhancing text classification to improve information filtering.* Otto-von-Guericke-Universität Magdeburg, Universitätsbibliothek.

LEBRET, R. & COLLOBERT, R. 2014. N-gram-based low-dimensional representation for document classification. *arXiv preprint arXiv:1412.6277*.

LIKAS, A., VLASSIS, N. & VERBEEK, J. J. 2003. The global k-means clustering algorithm. *Pattern recognition,* 36**,** 451-461.

LIN, Y.-S., JIANG, J.-Y. & LEE, S.-J. 2014. A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering,* 26**,** 1575-1590.

LÓPEZ, R. A. B., HERRERA, H. S. G. & GONZÁLEZ, J. D. S. 2016. Análisis de clústeres para la clasificación de datos económicos. *Revista Publicando,* 3**,** 267-275.

MAGDALENO, D. 2015. *Metodología para el agrupamiento de documentos semiestructurados.* Dr., Universidad Central "Marta Abreu" de Las Villas.

MAGDALENO GUEVARA, D., FUENTES, I. E., CABEZAS, M. & GARCÍA LORENZO, M. M. 2016. Recuperación de información para artículos científicos soportada en el agrupamiento de documentos XML. *Revista Cubana de Ciencias Informáticas,* 10**,** 57-72.

MAGDALENO GUEVARA, D., MIRANDA, Y., FUENTES, I. E. & GARCÍA, M. M. 2015. Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial,* 18**,** 69-80.

MALBERNAT, L. R., CLEMENS, M. P., VARELA, A. E. & URRIZAGA, M. Aplicación de técnicas de data mining en gestión de docentes de educación superior. XVII Workshop de Investigadores en Ciencias de la Computación (Salta, 2015), 2015.

MANNING, C. D. & SCHÜTZE, H. 1999. *Foundations of statistical natural language processing*, MIT press.

MEDEROS, A. A. L. & RUIZ, J. A. S. 2012. *Texminer: un modelo para el resumen automático y la desambiguación de textos científicos en el dominio de ingeniería de puertos y costas*, D-Universidad de La Habana.

MIESCH, A. T. 1976. Q-mode factor analysis of geochemical and petrologic data matrices with constant row-sums. US Govt. Print. Off.

MILLIGAN, G. W. & COOPER, M. C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50**,** 159-179.

MONROY MEDINA, O. A. Algoritmo de clustering basado en el concepto de densidad atómica. XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). 2016.

NALAWADE, R., SAMAL, A. & AVHAD, K. 2016. Improved Similarity Measure For Text Classification And Clustering. *International Research Journal of Engineering and Technology (IRJET),* 3**,** 214-219.

NASIR, J. A., VARLAMIS, I., KARIM, A. & TSATSARONIS, G. 2013. Semantic smoothing for text clustering. *Knowledge-Based Systems,* 54**,** 216-229.

NETO, J. L., SANTOS, A. D., KAESTNER, C. A., FREITAS, A. A. & NIEVOLA, J. C. A trainable algorithm for summarizing news stories. Proc. PKDD'2000 Workshop on Machine Learning and Textual Information Access, 2000.

PAL, N. R. & BISWAS, J. 1997. Cluster validation using graph theoretic concepts. *Pattern Recognition,* 30**,** 847-857.

PEI, J., HAN, J., MORTAZAVI-ASL, B., PINTO, H., CHEN, Q., DAYAL, U. & HSU, M.-C. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. icccn, 2001. IEEE, 0215.

PENNEY, G. P., WEESE, J., LITTLE, J. A., DESMEDT, P. & HILL, D. L. 1998. A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE transactions on medical imaging,* 17**,** 586-595.

PÉREZ-SUÁREZ, A., MARTÍNEZ-TRINIDAD, J. F., CARRASCO-OCHOA, J. A. & MEDINA-PAGOLA, J. E. 2013. An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition,* 46**,** 3040-3055.

PINTO, D., TOVAR, M., VILARIÑO, D., BELTRÁN, B., JIMÉNEZ-SALAZAR, H. & CAMPOS, B. 2010. BUAP: Performance of K-Star at the INEX'09 Clustering Task. *In:* GEVA, S., KAMPS, J. & TROTMAN, A. (eds.) *Focused Retrieval and Evaluation: 8th International Workshop of the Initiative for the Evaluation of*

XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, 2009, Revised and Selected Papers. Berlin, Heidelberg: Springer Berlin Heidelberg.

QIAN, Y. & ZHANG, K. A customizable hybrid approach to data clustering. Proceedings of the 2003 ACM symposium on Applied computing, 2003 New York. ACM, 485-489.

RAJESHWARI, P., SHANTHINI, B. & PRINCE, M. 2015. Hierarchical energy efficient clustering algorithm for WSN. *Middle East Journal of Scientific Research,* 23**,** 108-117.

ROMARY, P. L. L. Automatic Key Term Extraction from Scientific Articles in GROBID. SemEval, 2010. 4.

ROUSSEEUW, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* 20**,** 53-65.

SAITTA, S., RAPHAEL, B. & SMITH, I. F. A bounded index for cluster validity. International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2007. Springer, 174-187.

SEBASTIANI, F. A tutorial on automated text categorisation. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, 1999. Buenos Aires, AR, 7-35.

SERT, S. A., BAGCI, H. & YAZICI, A. 2015. MOFCA: Multi-objective fuzzy clustering algorithm for wireless sensor networks. *Applied Soft Computing,* 30**,** 151-165.

SHANNON, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review,* 5**,** 3-55.

SMALL, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology,* 24**,** 265-269.

SMALL, H. 1993. Macro-level changes in the structure of co-citation clusters: 1983–1989. *Scientometrics,* 26**,** 5-20.

SMALL, H. & SWEENEY, E. 1985. Clustering the science citation index® using co-citations: I. A comparison of methods. *Scientometrics,* 7**,** 391-409.

SRIKANT, R. & AGRAWAL, R. Mining sequential patterns: Generalizations and performance improvements. International Conference on Extending Database Technology, 1996. Springer, 1-17.

STEIN, B. & NIGGEMANN, O. On the nature of structure and its identification. International Workshop on Graph-Theoretic Concepts in Computer Science, 1999. Springer, 122-134.

SUÁREZ, A. P. & PAGOLA, J. E. M. A clustering algorithm based on generalized stars. International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2007. Springer, 248-262.

TINEO, R. J. M., SANDOVAL, E. A. P., BECERRA, C. I. V., VARGAS, E. P., APAZA, G. M. & SALINAS, E. A. 2016. MODELO DE CLUSTERING BASADO EN REDES NEURONALES PARA IDENTIFICAR EL PERFIL DE LOS ALUMNOS POR SEGMENTO ENFOCADO A LOS SERVICIOS DE TECNOLOGÍAS DE INFORMACIÓN DE LA UNIVERSIDAD PERUANA UNIÓN. *Revista de Investigación Business Intelligence,* 1.

TZORTZIS, G. & LIKAS, A. 2014. The MinMax k-Means clustering algorithm. *Pattern Recognition,* 47**,** 2505-2516.

VARGAS FLORES, S. I., LEDENEVA, Y. N., GARCÍA HERNÁNDEZ, R. A. & SIDOROV, G. 2016. Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos.

WANG, C.-D., LAI, J.-H. & ZHU, J.-Y. A conscience on-line learning approach for kernel-based clustering. Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010. IEEE, 531-540.

WANG, C. & BLEI, D. M. Collaborative topic modeling for recommending scientific articles. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011. ACM, 448-456.

WANG, X., ZHAO, Y., LIU, R. & ZHANG, J. 2013. Knowledge-transfer analysis based on co-citation clustering. *Scientometrics,* 97**,** 859-869.

WEST, J. D., WESLEY-SMITH, I. & BERGSTROM, C. T. 2016. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data,* 2**,** 113-123.

XIAO, J., LU, J. & LI, X. 2017. Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis,* 21**,** 1327-1338.

YAN, Y., CHEN, L. & TJHI, W.-C. 2013. Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Sets and Systems,* 215**,** 74-89.

YANG, Z., WANG, Y. & KITSUREGAWA, M. LAPIN: effective sequential pattern mining algorithms by last position induction for dense databases. International Conference on Database systems for advanced applications, 2007. Springer, 1020-1023.

ZAKI, M. J. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine learning,* 42**,** 31-60.

ANEXOS

Anexo 1. Descripción de los documentos usados como casos de estudio.

Clase de Referencia	Etiqueta	Área del conocimiento	Cantidad documentos	Idioma	Procedencia
Fuzzy Logic	FL	Ciencias técnicas	13	Inglés	ICT
SVM	SVM	Ciencias técnicas	19	Inglés	ICT
Association Rule	AR	Ciencias técnicas	14	Inglés	ICT
RST	RST	Ciencias técnicas	11	Inglés	ICT
Copula	Copula	Ciencias técnicas	10	Inglés	ICT
Belief Propagation	BP	Ciencias técnicas	9	Inglés	ICT
Modeling	Modeling	Ciencias técnicas	9	Inglés	Springer
Wireless sensor network	WSN	Ciencias técnicas	15	Inglés	Springer
Photosynthesis	PHO	Ciencias biológicas	18	Inglés	Springer
Oxidative Stress	OS	Ciencias biológicas	14	Inglés	Springer
Gene Expression	GE	Ciencias biológicas	16	Inglés	Springer
Micropropagation	MICRO	Ciencias biológicas	12	Inglés	Springer
Life Satisfation	LS	Ciencias sociales	13	Inglés	Springer
Quality of life	QL	Ciencias sociales	14	Inglés	Springer
Family firm	FF	Ciencias sociales	9	Inglés	Springer
Fuzzy	F(E)	Ciencias técnicas	12	Español	*
Clustering	C(E)	Ciencias técnicas	8	Español	*
**		Ciencias biológicas	27	Español	Biotecnología Vegetal

* Documentos recuperados con Google Scholar procedentes de diferentes revistas y clasificados siguiendo criterio de expertos.

** Documentos pertenecientes al Volumen 16, número 4 de la revista Biotecnología Vegetal.

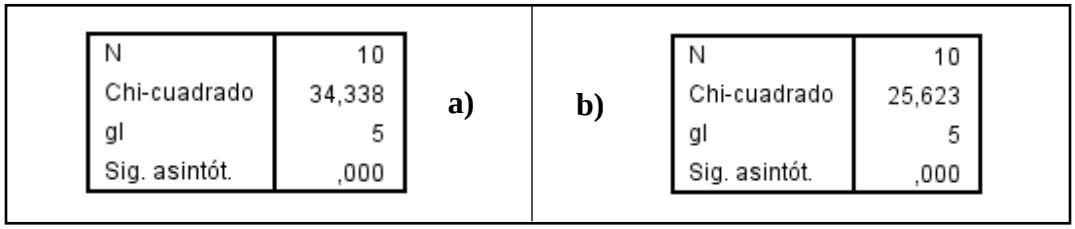
Anexo 2. Descripción de los casos de estudio utilizados.

No. Corpus	Cantidad de documentos	Cantidad de clases	Temas que trata
1	32	2	FL, SVM
2	25	2	RST, AR
3	32	2	RST, SVM
4	28	2	AR, FL
5	32	2	AR, SVM
6	32	2	PHO, OS
7	19	2	Copula, BP
8	29	2	Copula, SVM
9	27	2	LS, OS
10	30	2	FL, GE
11	30	2	OS, GE
12	24	2	Modeling, WSN
13	34	2	PHO, GE
14	20	2	F(E), C(E)
15	27	2	LS, QL
16	28	2	MICRO, GE
17	26	2	MICRO, OS
18	30	2	MICRO, PHO
19	28	2	FL, WSN
20	22	2	FF, LS
21	23	2	FF, QL
22	23	2	FL, Copula
23	31	3	RS, Copula, BP
24	38	3	Copula, SVM, BP
25	42	3	MICRO, OS, GE
26	46	3	FL, SVM, AR
27	44	3	AR, RST, SVM
28	45	3	RST, SVM, F
29	48	3	OS, GE, PHO
30	36	3	LS, QL, FF
31	60	4	OS, GE, PHO, MICRO
32	52	4	AR, SVM, C (E), F (E)
33	51	4	RST, SVM, Copula, BP
34	61	4	AR, RST, SVM, FL
35	44	4	Modeling, WSN, F(E), C(E)

Anexo 3 Resultados del test no paramétrico de Friedman al comparar el agrupamiento por unidades estructurales aplicando el algoritmo k-Means.

Tabla A1 Valores obtenidos para la medida Entropía al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,11	0,31	0,65	0,31	0,29	0,29
2	0,14	0,22	0,66	0,62	0,36	0,44
3	0	0,11	0,65	0,64	0,12	0,23
4	0,33	0,41	0,66	0,56	0,59	0
5	0,29	0,38	0,65	0,39	0,24	0,29
7	0	0,43	0,36	0,57	0	0
8	0	0,25	0,3	0,34	0	0,58
22	0,15	0,24	0,66	0,55	0	0,24
23	0	0,33	1	0,6	0,22	0
24	0	0,57	0,72	0,64	0,29	0,23



N	10
Chi-cuadrado	34,338
gl	5
Sig. asintót.	,000

a)

b)

N	10
Chi-cuadrado	25,623
gl	5
Sig. asintót.	,000

Figura A-1 Resultado de la prueba no paramétrica de Friedman para los valores de a) Entropía, b) Folkes-Mallows, obtenidos al aplicar el algoritmo k-Means a cada una de las unidades estructurales.

Tabla A2 Valores obtenidos para el índice Folkes-Mallows al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,94	0,83	0,7	0,83	0,77	0,77
2	0,92	0,84	0,66	0,66	0,77	0,63
3	1	0,94	0,7	0,65	0,94	0,88
4	0,74	0,74	0,67	0,61	0,58	1
5	0,77	0,78	0,7	0,72	0,82	0,77
7	1	0,7	0,71	0,68	1	1
8	1	0,88	0,84	0,7	1	0,54
22	0,91	0,83	0,66	0,58	1	0,83
23	1	0,78	0,48	0,61	0,82	1
24	1	0,65	0,63	0,59	0,7	0,84

Tabla A3 Valores obtenidos para la medida Micro Purity al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,95	0,91	0,74	0,91	0,84	0,84
2	0,94	0,9	0,69	0,74	0,88	0,76
3	1	0,95	0,74	0,7	0,97	0,94
4	0,84	0,86	0,7	0,71	0,74	1
5	0,84	0,87	0,74	0,85	0,88	0,84
7	1	0,84	0,83	0,72	1	1
8	1	0,92	0,91	0,93	1	0,72
22	0,94	0,89	0,7	0,76	1	0,89
23	1	0,86	0,57	0,72	0,88	1
24	1	0,75	0,77	0,72	0,78	0,89

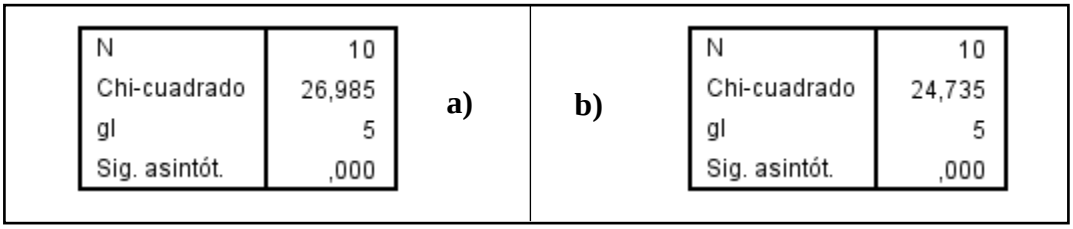


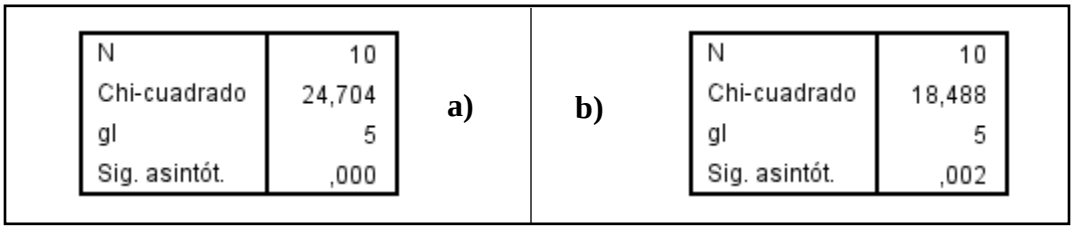
Figura A-2 Resultado de la prueba no paramétrica de Friedman para los valores de a) Micro Purity, b) Macro Purity, obtenidos al aplicar el algoritmo k-Means a cada una de las unidades estructurales.

Tabla A4 Valores obtenidos para la medida Macro Purity al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,71	0,82	0,92	0,82	0,86	0,89
2	0,96	0,9	1	0,89	0,89	1
3	1	0,94	0,9	0,88	0,76	0,79
4	0,82	0,89	0,94	0,91	1	1
5	0,86	0,86	0,93	0,86	0,79	1
7	0,95	0,96	1	0,83	1	1
8	1	0,95	1	0,88	1	1
22	1	1	0,98	0,8	1	1
23	0,95	0,88	0,97	0,85	0,78	1
24	0,9	0,87	0,88	0,92	1	0,87

Tabla A5 Valores obtenidos para la medida Precisión al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,96	0,91	0,81	0,91	0,88	0,88
2	0,96	0,92	0,77	0,8	0,88	0,82
3	1	0,96	0,81	0,7	0,98	0,94
4	0,88	0,86	0,78	0,78	0,74	1
5	0,88	0,87	0,8	0,84	0,91	0,88
7	1	0,84	0,88	0,77	1	1
8	1	0,92	0,93	0,9	1	0,72
22	0,95	0,92	0,77	0,86	1	0,92
23	1	0,89	0,68	0,88	0,92	1
24	1	0,77	0,87	0,85	0,85	0,91



N	10
Chi-cuadrado	24,704
gl	5
Sig. asintót.	,000

a)

b)

N	10
Chi-cuadrado	18,488
gl	5
Sig. asintót.	,002

Figura A-3 Resultado de la prueba no paramétrica de Friedman para los valores de a) Precisión, b) Cubrimiento, obtenidos al aplicar el algoritmo k-Means a cada una de las unidades estructurales.

Tabla A6 Valores obtenidos para la medida Cubrimiento al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,97	0,9	0,96	0,9	0,89	0,89
2	0,96	0,93	0,96	0,91	0,88	0,79
3	1	0,97	0,96	0,86	0,96	0,94
4	0,87	0,86	0,96	0,83	0,73	1
5	0,89	0,87	0,96	0,86	0,92	0,89
7	1	0,84	0,85	0,8	1	1
8	1	0,92	0,85	0,91	1	0,74
22	0,96	0,92	0,96	0,92	1	0,92
23	1	0,86	0,88	0,84	0,92	1
24	1	0,79	0,8	0,75	0,88	0,95

Tabla A7 Valores obtenidos para la métrica F-measure al aplicar el algoritmo de agrupamiento k-Means a cada una de las unidades estructurales.

Corpus	OverallSimSux	título	autores	resumen	cuerpo	referencias
1	0,97	0,91	0,67	0,91	0,88	0,88
2	0,96	0,92	0,66	0,66	0,88	0,75
3	1	0,97	0,67	0,65	0,97	0,94
4	0,86	0,86	0,76	0,62	0,66	1
5	0,88	0,88	0,67	0,85	0,91	0,88
7	1	0,84	0,84	0,75	1	1
8	1	0,93	0,89	0,7	1	0,58
22	0,96	0,91	0,66	0,68	1	0,91
23	1	0,87	0,48	0,61	0,9	1
24	1	0,8	0,64	0,6	0,82	0,92

N	10
Chi-cuadrado	30,541
gl	5
Sig. asintót.	,000

Figura A-4 Resultado de la prueba no paramétrica de Friedman para los valores de la medida F obtenidos al aplicar el algoritmo k-Means a cada una de las unidades estructurales.

Anexo 4. Resultados del test no paramétrico de Wilcoxon en la aplicación del algoritmo k-Means.

	referencias - abs	referencias - cuerpo	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-2,397[b]	-,420[c]	-2,608[b]	-1,186[b]	-1,355[c]	-2,807[c]	-2,599[c]	-2,666[c]	-1,632[b]
Sig. asintót. (bilateral)	,017	,674	,009	,236	,176	,005	,009	,008	,103

Figura A-5 Resultados del test no paramétrico de Wilcoxon para la medida Entropía, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-Means.

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,210[b]	-1,990[c]	-1,887[c]	-,296[c]	-1,521[b]	-2,549[b]	-2,701[b]	-2,670[b]	-,771[c]
Sig. asintót. (bilateral)	,833	,047	,059	,767	,128	,011	,007	,008	,441

Figura A-6 Resultados del test no paramétrico de Wilcoxon para el índice Folkes-Mallows, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-Means.

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,281[b]	-2,244[c]	-1,887[c]	,000[d]	-1,521[b]	-2,521[b]	-2,666[b]	-2,670[b]	-,654[c]
Sig. asintót. (bilateral)	,779	,025	,059	1,000	,128	,012	,008	,008	,513

Figura A-7 Resultados del test no paramétrico de Wilcoxon para la medida F, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-Means.

a)

	referencias - OverallSimSux	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-1,521[b]	-,350[b]	-1,305[c]	-,102[b]	-,535[c]	-2,805[b]	-1,788[c]	-2,530[b]	-1,226[c]
Sig. asintót. (bilateral)	,128	,726	,192	,919	,592	,005	,074	,011	,220

b)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,281[b]	-1,887[c]	-1,988[c]	-,534[c]	-1,521[b]	-2,805[b]	-1,943[b]	-1,897[b]	-1,177[c]
Sig. asintót. (bilateral)	,779	,059	,047	,594	,128	,005	,052	,058	,239

c)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,281[b]	-1,785[c]	-2,143[c]	-,239[c]	-1,524[b]	-2,501[b]	-2,501[b]	-2,547[b]	-,971[c]
Sig. asintót. (bilateral)	,779	,074	,032	,811	,128	,012	,012	,011	,332

d)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSux	titulo - OverallSimSux	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,281[b]	-1,841[c]	-1,989[c]	-,534[c]	-1,521[b]	-2,805[b]	-1,943[b]	-2,261[b]	-1,177[c]
Sig. asintót. (bilateral)	,779	,066	,047	,594	,128	,005	,052	,024	,239

Figura A-8 Resultados del test no paramétrico de Wilcoxon para las medidas a) Cubrimiento, b) Precisión, c) Micro Purity y d) Macro Purity, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-Means.

Anexo 5. Resultados del test no paramétrico de Friedman en la comparación del agrupamiento por unidades estructurales aplicando el algoritmo k-XStar.

a)

N	10
Chi-cuadrado	24,719
gl	5
Sig. asintót.	,000

b)

N	10
Chi-cuadrado	37,983
gl	5
Sig. asintót.	,000

c)

N	10
Chi-cuadrado	38,833
gl	5
Sig. asintót.	,000

Figura A-9 Resultado de la prueba no paramétrica de Friedman para los valores de a) Entropía b) Folkes-Mallows y c) F-measure obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

a)

N	10
Chi-cuadrado	37,040
gl	5
Sig. asintót.	,000

b)

N	10
Chi-cuadrado	11,593
gl	5
Sig. asintót.	,041

Figura A-10 Resultado de la prueba no paramétrica de Friedman para los valores de a) Cubrimiento b) Precisión, obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

a)

N	10
Chi-cuadrado	13,962
gl	5
Sig. asintót.	,016

b)

N	10
Chi-cuadrado	11,384
gl	5
Sig. asintót.	,044

Figura A-11 Resultado de la prueba no paramétrica de Friedman para los valores de a) Micro Purity b) Macro Purity, obtenidos al aplicar el algoritmo k-XStar a cada una de las unidades estructurales.

Anexo 6. Resultados del test no paramétrico de Wilcoxon en la aplicación del algoritmo k-XStar.

	referencias - OverallSimSu x	referencias - titulo	referencias - autores	referencias - abs	referencias - cuerpo	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-1,338[b]	-1,606[b]	,000[c]	-2,018[b]	-1,219[b]	-,281[d]	-2,255[b]	-,791[d]	-,536[b]
Sig. asintót. (bilateral)	,181	,108	1,000	,044	,223	,779	,024	,429	,592

Figura A-12 Resultados del test no paramétrico de Wilcoxon para la medida Precisión, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-XStar.

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSu x	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-1,367[b]	-2,346[b]	-,423[b]	-1,400[b]	-1,263[b]	-,351[c]	-2,247[b]	-1,680[c]	-,178[b]
Sig. asintót. (bilateral)	,172	,019	,672	,161	,206	,726	,025	,093	,859

Figura A-13 Resultados del test no paramétrico de Wilcoxon para la medida Micro Purity, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-XStar.

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSu x	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-1,367[b]	-2,298[b]	,000[c]	-1,606[b]	-1,338[b]	-,281[d]	-2,255[b]	-1,367[d]	-,297[b]
Sig. asintót. (bilateral)	,172	,022	1,000	,108	,181	,779	,024	,172	,766

Figura A-14 Resultados del test no paramétrico de Wilcoxon para la medida Macro Purity, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-XStar.

a)

	referencias - OverallSimSu x	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-2,194[b]	-1,688[c]	-2,293[c]	-2,807[c]	-1,123[c]	-2,807[b]	-2,803[b]	-2,312[b]	-1,225[b]
Sig. asintót. (bilateral)	,028	,091	,022	,005	,262	,005	,005	,021	,221

b)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSu x	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-,105[b]	-2,601[b]	-,677[b]	-2,073[b]	-,700[b]	-2,383[c]	-2,497[b]	-2,075[c]	-1,955[b]
Sig. asintót. (bilateral)	,917	,009	,498	,038	,484	,017	,013	,038	,051

c)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSu x	cuerpo - titulo	abs - titulo	autores - titulo	titulo - OverallSimSu x
Z	-1,377[b]	-2,142[b]	-2,805[b]	-1,362[b]	-1,580[c]	-,408[b]	-2,073[c]	-2,805[c]	-2,807[c]
Sig. asintót. (bilateral)	,169	,032	,005	,173	,114	,683	,038	,005	,005

d)

	referencias - cuerpo	referencias - abs	referencias - autores	referencias - titulo	referencias - OverallSimSu x	titulo - OverallSimSu x	autores - titulo	abs - titulo	cuerpo - titulo
Z	-1,530[b]	-2,040[b]	-2,803[b]	-,764[b]	-1,988[c]	-2,807[c]	-2,807[c]	-2,312[c]	-,533[c]
Sig. asintót. (bilateral)	,126	,041	,005	,445	,047	,005	,005	,021	,594

Figura A-15 Resultados del test no paramétrico de Wilcoxon para las medidas a) Cubrimiento, b) Entropía, c) Folkes-Mallows y d) F-measure, en la comparación del agrupamiento de las unidades Referencias y Título con el resto de las unidades estructurales aplicando el algoritmo k-XStar.

Anexo 7. Valores obtenidos para cada métrica al aplicar los algoritmos k-Means y k-XStar teniendo en cuenta solo los títulos de las referencias.

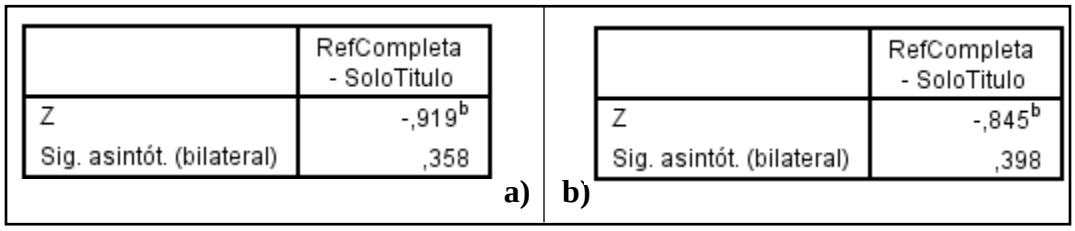
Tabla A8 Valores obtenidos al aplicar el algoritmo k-Means.

Corpus	Entropía	Folkes-Mallows	Micro Purity	Macro Purity	Medida F	Precisión	Cubrimiento
1	0,37	0,69	0,79	0,84	0,81	0,84	0,84
2	0	1	1	1	1	0,89	0,89
3	0,18	0,88	0,91	0,93	0,94	0,93	0,95
4	0,28	0,8	0,88	0,91	0,89	0,91	0,9
5	0,33	0,73	0,81	0,86	0,84	0,86	0,87
7	0	1	1	1	1	1	0,95
8	0	1	1	1	1	1	0,97
22	0,11	0,91	0,93	0,95	0,95	0,95	0,92
23	0,46	0,61	0,79	0,79	0,77	0,79	0,81
24	0,31	0,76	0,85	0,88	0,87	0,88	0,88

Tabla A9 Valores obtenidos al aplicar el algoritmo k-XStar.

Corpus	Entropía	Folkes-Mallows	Micro Purity	Macro Purity	Medida F	Precisión	Cubrimiento
1	0,14	0,6	0,87	0,91	0,74	0,91	0,66
2	0	0,79	1	1	0,85	1	0,79
3	0	0,67	1	1	0,76	1	0,71
4	0	0,8	1	1	0,87	1	0,8
5	0	0,7	1	1	0,79	1	0,74
7	0	0,85	1	1	0,91	1	0,84
8	0	0,94	1	1	0,97	1	0,93
22	0	0,55	1	1	0,68	1	0,63
23	0,29	0,53	0,77	0,85	0,68	0,79	0,81
24	0	0,78	1	1	0,87	1	0,81

Anexo 8. Resultados del test no paramétrico de Wilcoxon teniendo en cuenta solamente las referencias bibliográficas.



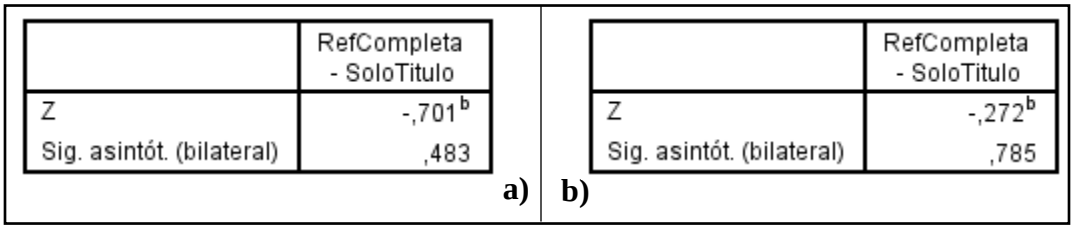
	RefCompleta - SoloTitulo
Z	-,919[b]
Sig. asintót. (bilateral)	,358

a)

	RefCompleta - SoloTitulo
Z	-,845[b]
Sig. asintót. (bilateral)	,398

b)

Figura A-16 Resultado de la prueba no paramétrica de Wilcoxon para los valores de Cubrimiento obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.



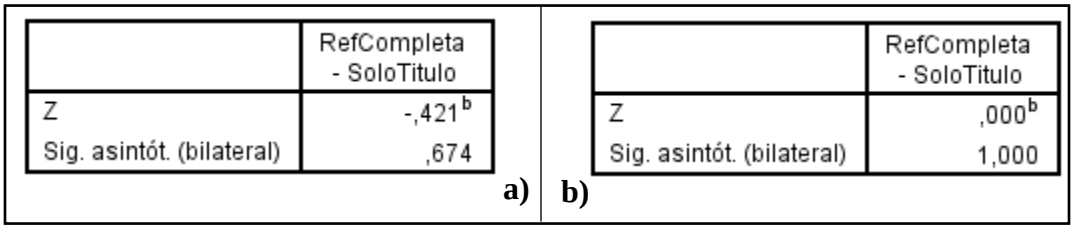
a)

	RefCompleta - SoloTitulo
Z	-,701[b]
Sig. asintót. (bilateral)	,483

b)

	RefCompleta - SoloTitulo
Z	-,272[b]
Sig. asintót. (bilateral)	,785

Figura A-17 Resultado de la prueba no paramétrica de Wilcoxon para los valores de Precisión obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.



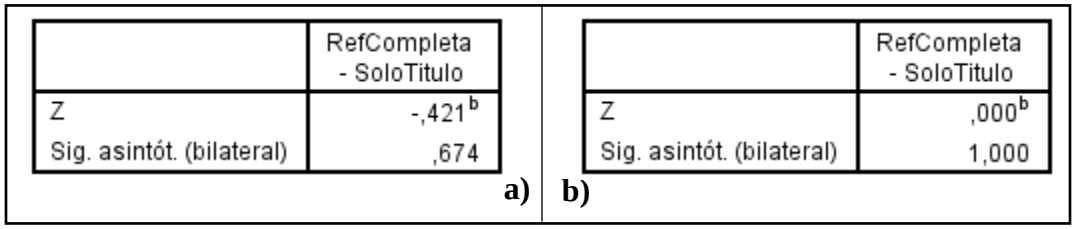
a)

	RefCompleta - SoloTitulo
Z	-,421[b]
Sig. asintót. (bilateral)	,674

b)

	RefCompleta - SoloTitulo
Z	,000[b]
Sig. asintót. (bilateral)	1,000

Figura A-18 Resultado de la prueba no paramétrica de Wilcoxon para los valores de Micro Purity obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.



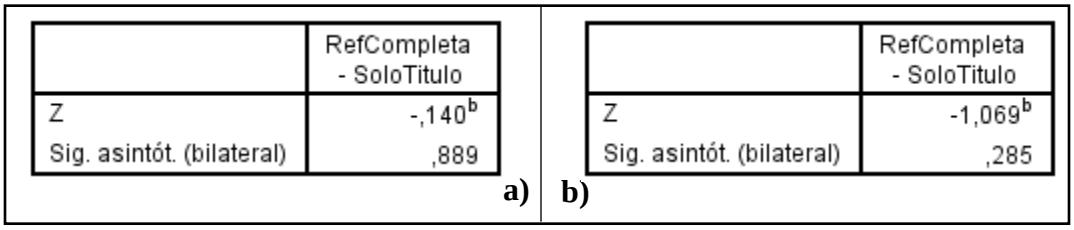
a)

	RefCompleta - SoloTitulo
Z	-,421[b]
Sig. asintót. (bilateral)	,674

b)

	RefCompleta - SoloTitulo
Z	,000[b]
Sig. asintót. (bilateral)	1,000

Figura A-19 Resultado de la prueba no paramétrica de Wilcoxon para los valores de Macro Purity obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.



a)

	RefCompleta - SoloTitulo
Z	-,140[b]
Sig. asintót. (bilateral)	,889

b)

	RefCompleta - SoloTitulo
Z	-1,069[b]
Sig. asintót. (bilateral)	,285

Figura A-20 Resultado de la prueba no paramétrica de Wilcoxon para los valores de Entropía obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.

a)

	RefCompleta - SoloTitulo
Z	-,169[b]
Sig. asintót. (bilateral)	,866

b)

	RefCompleta - SoloTitulo
Z	-1,266[b]
Sig. asintót. (bilateral)	,205

Figura A-21 Resultado de la prueba no paramétrica de Wilcoxon para los valores del índice Folkes-Mallows obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.

a)

	RefCompleta - SoloTitulo
Z	-,169[b]
Sig. asintót. (bilateral)	,866

b)

	RefCompleta - SoloTitulo
Z	-1,051[b]
Sig. asintót. (bilateral)	,293

Figura A-22 Resultado de la prueba no paramétrica de Wilcoxon para los valores de la medida F obtenidos al aplicar el algoritmo a) k-Means y b) k-XStar a las referencias teniendo en cuenta solo el título y teniendo en cuenta todo el texto de las referencias.

Anexo 9. Resultados del test no paramétrico de Friedman en la comparación del método propuesto con la función OverallSimSux y la función SimRefBib.

a)

N	10
Chi-cuadrado	5,522
gl	3
Sig. asintót.	,137

b)

N	10
Chi-cuadrado	7,087
gl	3
Sig. asintót.	,069

Figura A-23 Resultado de la prueba no paramétrica de Friedman para los valores de la medida a) Micro Purity y b) Macro Purity obtenidos al comparar el método propuesto con la función OverallSimSux usando los algoritmos k-Means y k-XStar y con la función SimRefBib.

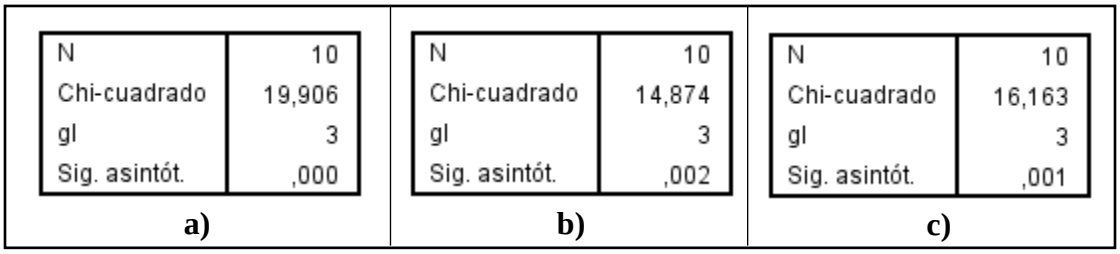
a)

N	10
Chi-cuadrado	4,662
gl	3
Sig. asintót.	,198

b)

N	10
Chi-cuadrado	4,563
gl	3
Sig. asintót.	,207

Figura A-24 Resultado de la prueba no paramétrica de Friedman para los valores de la medida a) Entropía y b) Precisión obtenidos al comparar el método propuesto con la función OverallSimSux usando los algoritmos k-Means y k-XStar y con la función SimRefBib.



N	10
Chi-cuadrado	19,906
gl	3
Sig. asintót.	,000

a)

N	10
Chi-cuadrado	14,874
gl	3
Sig. asintót.	,002

b)

N	10
Chi-cuadrado	16,163
gl	3
Sig. asintót.	,001

c)

Figura A-25 Resultado de la prueba no paramétrica de Friedman para los valores de la medida a) Cubrimiento, b) Folkes-Mallows y c) F-measure al comparar el método propuesto con la función OverallSimSux usando los algoritmos k-Means y k-XStar y con la función SimRefBib.