Procedimiento para identificar similitud de perfiles de usuarios en revistas mediante minería de textos

[Procedure for identification similarity of user profiles in journals through text mining]

[Um método para identificar semelhança dos perfis de usuário em revistas através da mineração de texto]

[1] [2] [3]
, ,

Artículo de investigación científica y tecnológica[4]

Resumen

El aumento de las publicaciones científicas convierte en un desafío la labor de identificar árbitros para evaluar los artículos. Lograr una buena calidad del proceso editorial requiere que la identificación de evaluadores sea un proceso rápido y apropiado. Los editores en ocasiones utilizan la intuición para determinar, sin métodos cuantitativos, que tan apropiado podría ser un evaluador. Todo esto provoca que los evaluadores seleccionados no sean siempre los idóneos. En esta investigación se presenta un procedimiento para facilitar la identificación de evaluadores en las revistas gestionadas con el Open Journal System (OJS), constituyendo un paso de avance en la eliminación de los métodos cualitativos en la identificación de evaluadores y sus similitudes. Se utilizaron métodos teóricos y empíricos, incluyendo técnicas y herramientas de minería de texto. La aplicación del procedimiento en la revista Minería y Geología, tomada como caso de estudio, evidenció su efectividad.

Palabras clave: *agrupamiento jerárquico; minería de textos; perfiles de usuarios; similitud.*

Abstract

Increasing scientific publications becomes challenging work to identify referees to evaluate the articles. Achieve good quality of the editorial process requires the identification of evaluators is a quick and appropriate process. Publishers sometimes use intuition to determine, without quantitative methods, as appropriate could be an evaluator. All this causes the selected evaluators are not always suitable. This research presents a procedure to facilitate the identification of evaluators in journals managed with the Open Journal System (OJS), constituting a step forward in the elimination of qualitative methods in identifying evaluators and their similarities. theoretical and empirical methods were used, including tools and techniques of text mining. How case study took the journal Mining and Geology, for the application of the process, demonstrating its effectiveness.

Keywords: *hierarchical clustering; text mining; user profiles; similarity.*

Resumo

Aumentando publicações científicas torna-se um desafio de identificar os árbitros trabalho para avaliar os artigos. Alcançar uma boa qualidade do processo editorial requer a identificação de avaliadores é um processo rápido e adequado. Publishers, por vezes, usar a intuição para determinar, sem métodos quantitativos, conforme o caso poderia ser um avaliador. Tudo isto faz com que os avaliadores selecionados nem sempre são adequados. Esta pesquisa apresenta um procedimento para facilitar a identificação dos avaliadores em revistas gerenciados com o Jornal Open System (OJS), constituindo um passo em frente na eliminação de métodos qualitativos na identificação de avaliadores e suas

[1]
[2]
[3]
[4] .

semelhanças. Foram utilizados métodos teóricos e empíricos, incluindo as ferramentas e técnicas de mineração de texto. A aplicação do procedimento no Jornal Minas e Geologia, tomado como um estudo de caso, mostrou a sua eficácia.
Palavras-chave*: agrupamento hierárquico; mineração de texto; perfis de usuário; similitude*.

0. Introducción

«*Las revistas científicas, desde su establecimiento han sabido ostentar el título de difusoras por excelencia del conocimiento científico, encontrando en las Tecnologías de la Información y las Comunicaciones (TIC) una vía para llegar a un mayor número de lectores en todo el orbe*» (Marbot & Rojas, 2015, 49). En la actualidad la mayoría de las revistas son gestionadas por sistemas informáticos que permiten el envío en línea de los artículos, la selección de los evaluadores[5] y el chequeo de las diferentes etapas por las que transitan las contribuciones.

«*Entre las aplicaciones informáticas más utilizadas para la gestión de revistas científicas se encuentran el Open Journal Systems (OJS), el Sistema Electrónico de Gestión Editorial (SEGE) y el Quark Publishing System 7 (QPS 7)*» (Rodríguez & Leiva, 2009, 61). Mediante el *OJS* se colecta un conjunto importante de datos, principalmente de carácter textual, que debidamente procesados por herramientas informáticas pueden ser de utilidad a los consejos editoriales. La minería de textos es especialmente apropiada para este propósito. «*La minería de textos permite identificar relaciones y modelos en la información no estructurada, así como proveer de una visión selectiva y perfeccionada de la información contenida en documentos de textos y sacar consecuencias para la acción, detectar patrones interesantes y no triviales, e incluso, información sobre el conocimiento almacenado en las mismas*» (Tan, 1999, 1) y (Hotho, Nürnberger & Paaß, 2005, 4).

«*La minería de textos necesita para lograr sus propósitos, combinar varias técnicas, de ahí que sea un campo multidisciplinario que incluye la recuperación de información, el análisis de textos, la extracción de información, el agrupamiento, la construcción de resúmenes, la categorización, la clasificación, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos*» (Tan, 1999, 1). Uno de los principales problemas que deben resolver los procesos de gestión de las revistas científicas es la identificación de posibles evaluadores. El *OJS* tiene implementado algunos módulos que contribuyen a su identificación, pero poseen las siguientes limitaciones:

– Las búsquedas está concebida principalmente para la obtención de documentos por una necesidad de información dada y no para la identificación de posibles revisores o expertos en una temática determinada.

– La no existencia de un mecanismo u opción que permita realizar un agrupamiento jerárquico de los primeros autores de artículos respecto al resumen y las palabras claves de un artículo determinado.

[5] Experto, *referee* o persona con influencia en una o ciertas materias porque es considerada una autoridad en ellas.

Estas limitaciones dificultan la identificación de autores de artículos de la propia plataforma que podrían servir como posibles evaluadores, propiciando un lento proceso de selección, que se desconozcan algunos de los posibles candidatos y de conocerse, no se sepa con claridad la estructura jerárquica de estos respecto al tema del artículo científico a evaluar.

1. Fundamento teórico

El aumento del número de artículos científicos, cada vez más colaborativos e interdisciplinarios, identificar revisores adecuados se ha convertido en un gran desafío, además de ser una tarea muy demorada. Los evaluadores de los trabajos científicos influyen en el progreso personal del investigador y determinan nuevas oportunidades de investigación. Ejercen cierto control sobre la calidad del trabajo y promueven una investigación innovadora. Animan a la diseminación de dicha innovación y sirven para jerarquizar bajo el principio de la originalidad, a investigadores, publicaciones e instituciones.

Se puede decir que el modo en sentido general en el que se realiza la identificación de posibles evaluadores en revistas gestionadas por el *OJS* sigue siendo similar al procedimiento tradicional de las editoriales de revistas científicas, donde la elección de referees (o evaluadores) es una de las atribuciones tradicionales de los editores, al suponerse que un buen editor debe estar al corriente del desarrollo en su área de conocimiento y por tanto, sabe qué expertos están cualificados para evaluar un trabajo determinado.

Relacionado con los tópicos conformación e identificación de la similitud de perfiles de usuarios, estudios de Escobar (2007), Bedoya (2013) y Rodríguez et al. (2016), sirvieron de base en el desarrollo de la presente investigación. Sin embargo, no existe un procedimiento específico para identificar la similitud de perfiles de usuarios en el *OJS*. Por ello, establecer un procedimiento para el análisis de datos textuales de perfiles de usuarios que permita identificar posibles evaluadores en revistas científicas gestionadas por el *OJS*, utilizando gran parte de la base teórica del modelo de espacio vectorial y el método de agrupamiento jerárquico es el objetivo de la investigación.

El procedimiento propuesto está conformado por varias fases o etapas generales que permiten lograr la identificación de los posibles evaluadores por medio de valores de similitud y agrupamiento jerárquico. Para su validación se aplicó en la revista Minería y Geología.

2. Metodología

Se realizó una revisión bibliográfica para el estudio de los diferentes enfoques y tendencias, en materia de técnicas de minería de textos, conformación y similitud

de perfiles de usuarios. Para ello se emplearon métodos de investigación teóricos y empíricos, a saber:

– Métodos teóricos: - Análisis y síntesis, para procesar la información en la elaboración de los fundamentos teóricos; - Histórico lógico, para el estudio de la evolución del problema y conocer los resultados alcanzados tras la aplicación de otros posibles acercamientos en cuanto a identificar la similitud de perfiles de usuarios del *OJS*, e - Hipotético deductivo, para la elaboración de la hipótesis y la deducción de los resultados de la investigación.

– Métodos empíricos: -_Entrevista, como punto de partida, para estudiar el estado de opiniones de especialistas sobre la identificación de posibles evaluadores de artículos científicos, así como las posibilidades de uso del procedimiento propuesto, y - Observación científica, en el diagnóstico e implantación de los resultados y la aseveración de su evolución.

3. Resultados y discusión

El procedimiento propuesto para la identificación de posibles evaluadores en revistas gestionadas por el *OJS* parte una fase inicial donde se conforman los perfiles de usuarios; tres fases intermedias donde se realiza una representación espacio vectorial, selección de rasgos, similitud y agrupamiento, hasta una fase final donde se evalúan los resultados obtenidos, como se muestra en la figura 1.

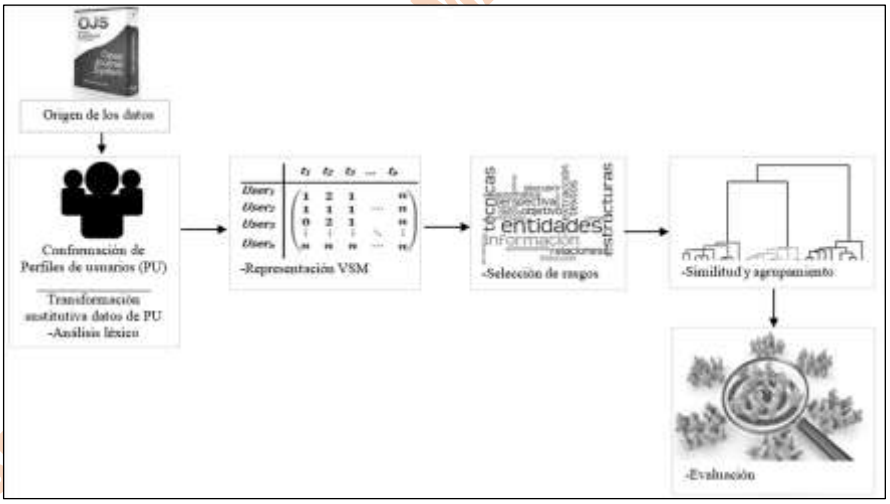


Figura 1. Esquema general del procedimiento propuesto

3.1 Operaciones de conformación y transformación

El conjunto de datos, de preferencia textual, pertenecientes a perfiles de usuarios de investigadores es tratado como un corpus. Arco et al. (2007, 21-22) exponen que existen dos grandes tipos de operaciones con los corpus textuales: operaciones de conformación y operaciones de transformación. El primer tipo tiene el objetivo de conformar el propio corpus mediante la adición de textos, el

ordenamiento de estos, su delimitación y segmentación, en tanto, el segundo tipo, se ejecutan sobre un corpus ya conformado y se dividen en dos clases: transformaciones genéricas y transformaciones específicas que incluyen dos subclases: transformaciones descriptivas y transformaciones sustitutivas.

En el caso de las operaciones de conformación, la creación del corpus se propone mediante el ordenamiento y delimitación de textos que incluye:

– Desambiguación de nombres de investigadores en revistas científicas. La ambigüedad en el nombre de los autores es un problema que afecta la efectividad del resultado final del procedimiento. «*Este problema se refiere a la posibilidad de representar el nombre de los autores de diferentes formas en los metadatos bibliográficos acopiados en los repositorios digitales. Puede manifestarse de dos formas: (1) nombres iguales que no se refieren al mismo autor y (2) nombres diferentes, que se refieren al mismo autor*» (Alonso, Hidalgo & Leiva, 2014, 133). Para crear por cada autor un identificador único (*id*) y lograr la generación de perfiles de usuarios reales, de forma tal que no exista ambigüedad en los nombres de los autores de artículos, hay que homogenizar los datos de la tabla *authors* de la base de datos del sistema *OJS* eliminando incongruencias en el nombre de los autores.

- Elección de los atributos del perfil de usuario. En la elección de los atributos que conformarán el perfil de usuario, el sistema propuesto, debe tener en cuenta las peculiaridades fundamentales por las que será posible la identificación de posibles expertos. Para obtener los atributos se partió de Rodríguez (2013, 178-181) ajustada a las exigencias que impone el *OJS*: nombre y apellidos, grado científico o académico, resumen y palabras claves de artículos publicados. En la figura 2 se muestra el proceso para la conformación de los perfiles de usuarios. Desde una fase inicial en el que los usuarios interactúan con el sistema *OJS*, hasta una fase final donde se le realizan a los datos almacenados en la base de datos una serie de operaciones que hacen posible queden conformados los perfiles de usuarios.

- Creación del perfil de usuario. Se eligió la formación del perfil de usuario de posibles evaluadores mediante la combinación de los métodos explícito e implícito propuestos por Samper (2005, 56).

En el caso de las operaciones de transformación mediante las transformaciones sustitutivas se propone la realización de un análisis léxico que incluye:

- Remover etiquetas html, normalizar el espaciado y codificar el texto a utf-8.
- Eliminación de *stopwords* (nombre que reciben las palabras sin significado como: artículos, pronombres, preposiciones, u otros). Estas palabras son utilizadas para eliminar términos comunes que no aportan ninguna información sobre el contenido o la materia propiamente del perfil de usuario.
- Lematización que es el proceso mediante el cual se asigna a cada palabra en un texto el lema correspondiente.

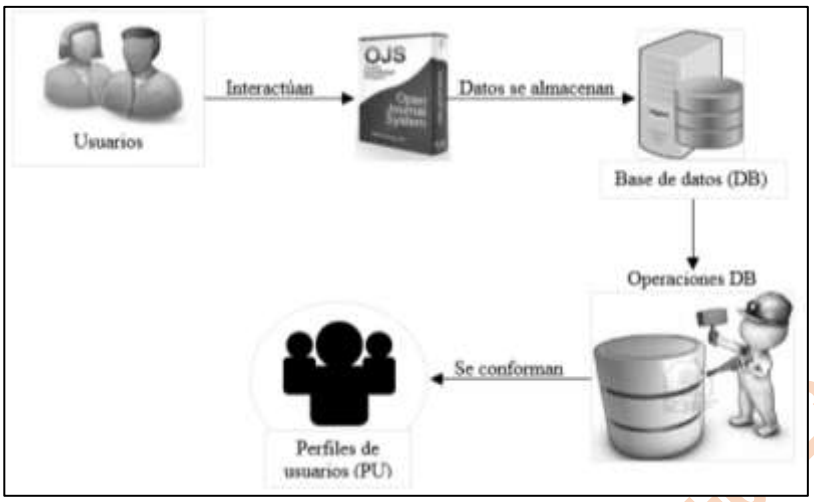


Figura 2. Proceso para la conformación de los perfiles de usuario

3.2 Representación espacio-vectorial (VSM)

Los perfiles de usuario se generan a partir de un grupo de operaciones mediante un lenguaje de manipulación de datos, conformando una nueva tabla en la base de datos del propio sistema *OJS*, que contiene un conjunto de perfiles de usuarios *(U)* y términos *(T)*, en la que cada usuario U_i contiene un número de términos. De esta forma, es posible representar a cada usuario como un vector perteneciente a un espacio n-dimensional, siendo n el número de términos del conjunto T que conforman el perfil: $U_i = (t_{i1}, t_{i2}, t_{i3}; \dots\dots\dots; t_{in});\ t_{ij} \in [0,1], j = 1, \dots, n$ (1)
Donde cada uno de los elementos t_{ij} puede representar la ausencia o relevancia del término t_j en el perfil del usuario u_i.

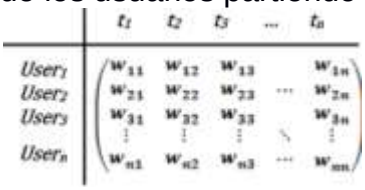
El proceso de construcción de los vectores-usuarios genera la representación de los usuarios extrayendo la información de los perfiles. Con ello se determinarán los pesos de cada término extraído de su perfil en el vector usuario u_i. Su función sería: $F: U \times T \rightarrow [0,1]$ (2)

La representación de cada vector-usuario tiene n componentes, de los cuales los a que estén referenciados en el perfil corresponde un valor diferente de 0, mientras que los que no estén referenciados adquieren el valor 0. Llegado a este punto es necesario determinar la importancia o peso de cada término en el vector-usuario. El cálculo de la importancia o peso de cada término se conoce como ponderación y se calcula frecuentemente según la siguiente función: $W_{i,j} = tf_{i,j} \times idf_j$ (3)
Donde: $tf_{i,j}$:es la frecuencia de aparición del término t_j en el perfil de usuario u_i
n_i:indica el número de perfiles en los que aparece el término t_j
idf:es la función inversa de n_i
Así, $idf_j = log\left(\frac{U}{n_i}\right)$, siendo U el número total de perfiles de usuarios.

Calculando los pesos de los términos en los perfiles aplicando la siguiente función:

$$W_{i,j} = tf_{i,j} \times log\left(\frac{U}{n_i}\right) \qquad (4)$$

Se obtiene una matriz de peso con los términos en cada uno de los perfiles de usuarios, quedando establecida en una tabla la matriz de los términos correspondiente a cada uno de los usuarios partiendo de su perfil.



3.3 Selección de rasgos

«*La selección de rasgos usada para representar un dominio tiene un efecto profundo en la calidad del modelo producido. Los rasgos bien seleccionados pueden mejorar la exactitud de las técnicas de minería de textos sustancialmente y reducir la cantidad de datos necesarios para obtener el nivel de funcionamiento deseado*» (Forman, 2003, 1290-1291). A partir de lo planteado, se proponen los siguientes criterios de selección de rasgos en dominios textuales para favorecer la rapidez y exactitud del procedimiento:

– Eliminar todos los términos cuyas frecuencias superan los umbrales superior e inferior especificados, debido a que el poder de resolución es máximo en rango medio de frecuencias de aparición de las palabras, tal y como puede observarse en la figura 3. «*El poder de resolución será la habilidad de los términos de indexación para convertirse en ítems relevantes*» Vegas (1999, citado por Samper, 2005, 14). Con ello se logra una reducción considerable de los datos y por ende un procesamiento más rápido y efectivo en la conformación de los grupos.

– Eliminar todos los términos cuya frecuencia de documentos es menor que un umbral predeterminado, pues los términos que aparecen solamente en muy pocos perfiles improbablemente llevan o contienen poca información general de la clase específica y algunas veces tienden a ser ruidosos, además, porque usar términos de aparición infrecuentes no es estadísticamente confiable.

– Implementar medidas que cuantifiquen la calidad de los términos, considerando aquellos que sobrepasen un umbral determinado.

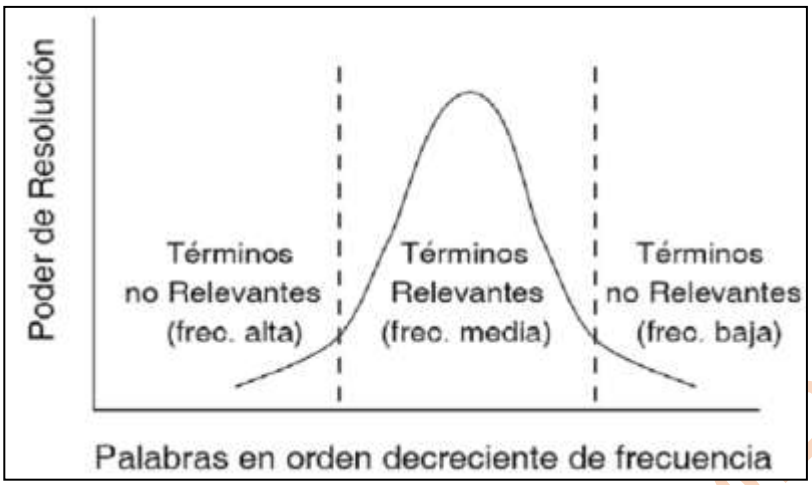


Figura 3. Poder de resolución de los términos de un documento (Vegas, citado por Samper, 2005, 14)

3.4 Similitud y agrupamiento de perfiles de usuario

Existen muchas medidas de similitud que pueden ser utilizadas para el agrupamiento. «*Las que han reportado los mejores resultados en dominios textuales son: similitud de Dice, Jaccard y Coseno*» (Frakes & Baeza, 1992). «*Entre ellas, la similitud Coseno ha sido la más utilizada para comparar vectores de frecuencias de documentos en un vocabulario de* n *términos*» (Korfhage, 1977). Por lo que se propone su uso.

$$Fcos(A,B) = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A^2{}_j \cdot \sum_{j=1}^{n} B^2{}_j}} \quad (5)$$

La relación coseno medirá el coseno del ángulo entre perfiles de usuarios y categorías, ya que éstos se representarán como vectores en un espacio multidimensional de dimensión t. Así, se puede expresar la medida de similitud entre un perfil de usuario p_i y una categoría c_k, siendo n el número de términos, como:

$$sim(p_i, c_k) = \frac{\vec{p_i} \cdot \vec{c_k}}{|\vec{p_i}| \cdot |\vec{c_k}|} = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A^2{}_j \cdot \sum_{j=1}^{n} B^2{}_j}} \quad (6)$$

De esta manera puede agruparse perfiles de usuarios jerárquicamente por una categoría determinada.

A partir de la de similitud del Coseno expuesta en la expresión 5 se puede obtener una matriz de similitud de usuarios.

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

Donde cada elemento δ_{ij} de M representa la similitud entre el estímulo i y el estímulo j. Quedando determinada la matriz de similitud de los usuarios, de forma tal que pueden ser identificados los niveles de compatibilidad y establecer conglomerados entre ellos.

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

«*Las estrategias jerárquicas (aglomerativas o divisivas) construyen una jerarquía de agrupamientos, representada tradicionalmente por un árbol llamado dendrograma*» (Pascual, 2010, 40). Se propone el uso de las estrategias aglomerativas, ya que son computacionalmente más rápidas que las divisivas. «*Para estimar la distorsión con respecto a la matriz de similitud o distancia original, el dendrograma resultante se propone se evalúe mediante el coeficiente de correlación cofenético (CPCC)*» (Estrada et al., 2010, 405).

3.5 Validación de la efectividad del procedimiento

El agrupamiento es un proceso subjetivo; el mismo conjunto de datos comúnmente necesita ser agrupado de formas diferentes dependiendo de su aplicación. Esta subjetividad hace el agrupamiento difícil y más aún, su validación. Para un mismo conjunto de objetos, si se aplican diferentes algoritmos de agrupamiento se pueden obtener resultados muy diferentes, por ello surge la necesidad de evaluar las estructuraciones. «*Estas medidas de evaluación (índices) se espera que sean objetivas y no tengan ninguna preferencia sobre algún algoritmo en particular. Existen tres categorías de índices de validación: índices externos, índices relativos e índices internos*» (Brun et al., 2007, 808).

En la presente investigación se propone el uso de los índices externos, ya que estos usan como patrón para compararse una estructuración específica, la cual es obtenida a partir de una información previa acerca de los datos, donde este patrón es visto como la estructuración real o verdadera. Un agrupamiento obtenido por un clasificador es mejor en la medida que éste se parece más a dicho patrón. «*Uno de los índices externos más usados es la medida F (F-measure)*» (Rijsbergen, 1979, 113). Sin embargo, un índice de validación externo puede ser cualquier medida de similitud entre estructuraciones, siempre calculando la similitud existente entre la estructuración obtenida por cierto clasificador, respecto a una estructuración conocida, la cual es asumida como correcta o natural para el conjunto de objetos analizados.

3.6 Caso de estudio

Se consideró cómo caso de estudio una colección formada por cincuenta perfiles de usuario previamente etiquetados, pertenecientes a la revista Minería y Geología, para ilustrar el funcionamiento y la efectividad del procedimiento. Luego de realizar un grupo de transformaciones sustitutivas se procede a efectuar una Representación Espacio Vectorial (*VSM*) y selección de rasgos (Tabla 1).

Tabla 1. Representación VSM a partir del análisis léxico realizado, selección de rasgos y pesado de los términos por medio de la medida frecuencia de término – frecuencia inversa de documento (tf-idf)

Términos	Usuarios														
	u1	u2	u3	u4	u5	u6	u7	u8	u9	u10	u11	u12	u13	u14	u15
vertical	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,007	0,005	0,000
vias	0,000	0,000	0,000	0,000	0,023	0,000	0,000	0,000	0,000	0,000	0,000	0,088	0,000	0,000	0,000
vincul	0,013	0,000	0,020	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
visc	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,069	0,000	0,000	0,000	0,000	0,005	0,000
visibl	0,000	0,000	0,000	0,004	0,000	0,000	0,000	0,000	0,000	0,005	0,000	0,000	0,000	0,000	0,000
vist	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,007	0,005	0,000
volcan	0,000	0,000	0,020	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
volum	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,016	0,012	0,000	0,000	0,000	0,000

Obtenida una representación *VSM* se realiza el agrupamiento de los usuarios, por medio de la aplicación de la medida de similitud del coseno disponible en la expresión (5). En la tabla 2 se muestra cómo quedarían los usuarios y su relevancia en relación con las siguientes categorías: geoestadística, yacimientos lateríticos, tectónica y minería responsable.

Tabla 2. Usuarios y su relevancia por medio de categorías preestablecidas

Usuarios	Términos			
	Geoestadística	Yacimientos lateríticos	Tectónica	Minería responsable
Arioza-Iznaga		0,1880		
Ávila-Torres				
Barea-Pérez		0,0685	0,2019	0,0253
Cobiella-Reguera		0,0149	0,0732	
Coello-Velazquez		0,045		
Cuador-Gil	0,3277	0,0926		
García-Pujadas				
Góngora-Leyva		0,0118		
Laurencio-Alfonso				
Legrá-Lobaina	0,0223	0,0529		0,0224
Martínez-Vargas	0,0744	0,0555		
Montero-Peña				0,1275
Rodríguez-Infante		0,0308	0,1818	0,0218
Torres-Tamayo		0,0445		
Zulueta-Torres				0,0319

La matriz de similitud entre los usuarios del sistema es obtenida luego de aplicar una de las medidas de similitud existentes entre vectores a los pesos de los términos de cada perfil de usuario respecto a otro. Como resultado de la aplicación de la función coseno disponible en la expresión (5) a partir de los valores del peso de cada término, se obtiene una matriz simétrica de similitudes entre usuarios como se ilustra en la Tabla 3, donde la intersección de los usuarios un mayor valor, significa que son más similares entre sí.

Tabla 3. Valores de matriz de similaridad entre usuarios usando la función del coseno

	Ariosa	Ávila	Barea	Cobiella	Coello	Cuador	García	Góngora	Laurencio	Legrá	Martínez	Montero	Rodríguez	Torres	Zuleta
Ariosa	NA	0,0558	0,1420	0,1966	0,2127	0,2178	0,1600	0,1662	0,0875	0,2445	0,2276	0,1301	0,2147	0,1624	0,0912
Ávila	0,0558	NA	0,1219	0,0556	0,1316	0,0692	0,0539	0,0965	0,0556	0,1126	0,1210	0,0000	0,0637	0,1386	0,0193
Barea	0,1420	0,1219	NA	0,1901	0,1815	0,1498	0,1486	0,1491	0,0443	0,2201	0,1993	0,0702	0,2259	0,1661	0,1107
Cobiella	0,1966	0,0556	0,1901	NA	0,1825	0,1401	0,1114	0,1170	0,0636	0,2363	0,2138	0,0990	0,2703	0,1704	0,0883
Coello	0,2127	0,1316	0,1815	0,1825	NA	0,2212	0,1665	0,1563	0,1147	0,2611	0,2274	0,0994	0,1922	0,2899	0,0697
Cuador	0,2178	0,0692	0,1498	0,1401	0,2212	NA	0,1051	0,1076	0,0905	0,3192	0,2640	0,0717	0,1954	0,2016	0,1006
García	0,1600	0,0539	0,1486	0,1114	0,1665	0,1051	NA	0,1163	0,0469	0,1968	0,1502	0,1163	0,1612	0,1445	0,0489
Góngora	0,1662	0,0965	0,1491	0,1170	0,1563	0,1076	0,1163	NA	0,0811	0,2136	0,1706	0,0893	0,1768	0,3328	0,1377
Laurencio	0,0875	0,0556	0,0443	0,0636	0,1147	0,0905	0,0469	0,0811	NA	0,0894	0,1117	0,0000	0,0535	0,1226	0,0253
Legrá	0,2445	0,1126	0,2201	0,2363	0,2611	0,3192	0,1968	0,2136	0,0894	NA	0,3847	0,1250	0,2476	0,2681	0,1534
Martínez	0,2276	0,1210	0,1993	0,2138	0,2274	0,2640	0,1502	0,1706	0,1117	0,3847	NA	0,0738	0,2252	0,2400	0,1163
Montero	0,1301	0,0000	0,0702	0,0990	0,0994	0,0717	0,1163	0,0893	0,0000	0,1250	0,0738	NA	0,1297	0,0740	0,1502
Rodríguez	0,2147	0,0637	0,2259	0,2703	0,1922	0,1954	0,1612	0,1768	0,0535	0,2476	0,2252	0,1297	NA	0,1988	0,1322
Torres	0,1624	0,1386	0,1661	0,1704	0,2899	0,2016	0,1445	0,3328	0,1226	0,2681	0,2400	0,0740	0,1988	NA	0,1259
Zuleta	0,0912	0,0193	0,1107	0,0883	0,0697	0,1006	0,0489	0,1377	0,0253	0,1534	0,1163	0,1502	0,1322	0,1259	NA

Para estimar la distorsión con respecto a la matriz de similitud o distancia original, se evaluó el *CPCC* para doce combinaciones, obteniéndose el mejor valor para la métrica euclidiana y el método de unión promedio. A partir de ello se procede a realizar un agrupamiento jerárquico, quedando finalmente como se ilustra en la figura 3.

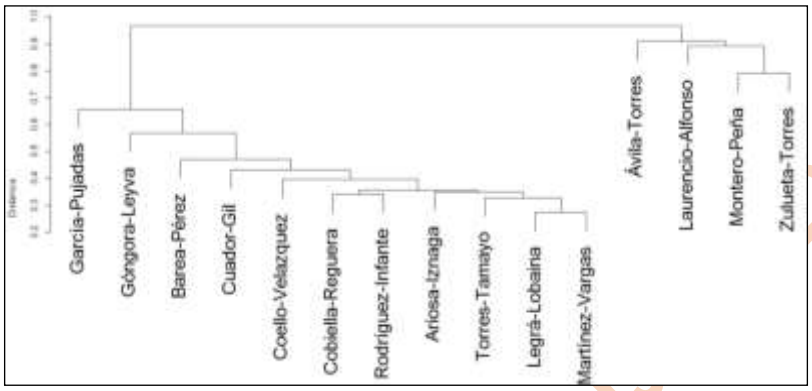


Figura 3. Representación gráfica del agrupamiento utilizando la métrica euclidiana y como método de unión el promedio

Para evaluar la calidad del agrupamiento se utilizaron cuatro clases: geoestadística, tectónica, yacimientos lateríticos y minería responsable. Se obtiene el valor de 0,73 cómo valor promedio de *f-measure*, evidenciándose un desempeño admisible en cuanto a la calidad del agrupamiento.

Con la aplicación del procedimiento propuesto al caso de estudio se evidenció cómo es posible obtener por niveles de relevancia categorías de perfiles de usuarios y por medio del agrupamiento jerárquico conocer la similitud existente entre perfiles generando grupos similares, facilitando todo esto la identificación de revisores en revistas científicas gestionadas con el *OJS*.

El procedimiento creado combina los pasos propuestos por Samper (2005, 56-57) y Tan (1999, 2-3), con un impacto social significativo ya que no se encontraron antecedentes de un procedimiento con las peculiaridades del propuesto en la presente investigación.

4. Conclusiones

El procedimiento presentado para identificar similitud de perfiles de usuarios en el *OJS*, es un paso de avance para facilitar la elección de evaluadores en revistas científicas y la eliminación de los métodos cualitativos que mediante información de tipo subjetivo han surtido de expertos a numerosas actividades de investigación y desarrollo. Al aplicarse a un caso de estudio se pudo constatar su aplicabilidad y efectividad, por lo que puede ser utilizado como herramienta o referencia a otros investigadores. Es un procedimiento que combina varias técnicas y que muestra una secuencia progresiva, que puede ser mejorado por medio de un estudio más

profundo en el proceso de selección de rasgos, buscando obtener mejores resultados respecto a su efectividad. Se pretende a partir de este, implementar una aplicación web que permita la integración con el *OJS* y obtener, por diversos criterios, grupos de usuarios específicos.

Referencias Bibliográficas

ALONSO SIERRA, Luis Enrique; HIDALGO DELGADO, Yusniel & LEIVA MEDEROS, Amed Abel (2014). Desambiguación del nombre de los autores en revistas científicas [en línea]. En: Revista Cubana de Ciencias Informáticas, Vol. 8, No. 3. La Habana (Cuba): Universidad de las Ciencias Informáticas. p. 131-150. e-ISSN: 2227-1899 <http://rcci.uci.cu/index.php?journal=rcci&page=article&op=viewFile&path[]=607&path[]=285> [consulta: 12/10/2015].

ARCO GARCÍA, Leticia; BELLO PÉREZ, Rafael; LLANES ABEIJÓN, Manuel; VALDÉS VERA, Libernys; MEDEROS MARTÍNEZ, Juan Manuel & PÉREZ OLMOS, Yoisy (2007). CorpusMiner 1.0: Herramienta para el agrupamiento de documentos [en línea]. En: Revista Cubana de Ciencias Informáticas, Vol. 1, No. 2 (abr). La Habana (Cuba): Universidad de las Ciencias Informáticas. p. 18-31. e-ISSN: 2227-1899 <http://rcci.uci.cu/index.php?journal=rcci&page=article&op=viewFile&path[]=12&path[]=11> [consulta: 20/12/2015].

BEDOYA LEIVA, Óscar Fernando (2013). Clasificación difusa para descubrir perfiles de usuarios en la web [en línea]. En: Revista Educación en Ingeniería, Vol. 8, No. 16. Bogotá (Colombia): Asociación Colombiana de Facultades de Ingeniería. p. 94-104. e-ISSN: 1900-8260. <http://www.educacioneningenieria.org/index.php/edi/article/download/272/175> [consulta: 14/05/2016].

BRUN, Marcel; SIMA, Chao; HUA, Jianping; LOWEY, James; CARROLL, Brent; SUH, Edward & DOUGHERTY, Edward R. (2007). Model-based evaluation of clustering validation measures [on line]. In: Pattern Recognition, Vol. 40, No. 3. p. 807-824. e-ISSN: 0031-3203. <http://www.sciencedirect.com/science/article/pii/S0031320306003104> [consult: 20/12/2016].

ESCOBAR JERIA, Víctor Heughes (2007). Minería Web de uso y perfiles de usuario: aplicaciones con lógica difusa [en línea]. Tesis doctoral (Doctor en Informática). Universidad de Granada (España), Departamento de Ciencias de la Computación e Inteligencia Artificial. <http://decsai.ugr.es/Documentos/tesis_dpto/100.pdf> [consulta: 23/01/2015].

ESTRADA CASTILLÓN, Eduardo; SCOTT MORALES, Laura; VILLARREAL QUINTANILLA, José A.; JURADO YBARRA, Enrique; COTERA CORREA, Mauricio; CANTÚ AYALA, César & GARCÍA PÉREZ, Jaime (2010). Clasificación de los pastizales halófilos del noreste de México asociados con perrito de las praderas (Cynomys mexicanus): diversidad y endemismo de especies [en línea]. En: Revista Mexicana de Biodiversidad, Vol. 81, No. 2. Distrito Federal (México): Universidad Nacional Autónoma de México. p. 401-416. e-ISSN: 2007-8706. <http://www.scielo.org.mx/pdf/rmbiodiv/v81n2/v81n2a14.pdf> [consulta: 21/05/2016].

FORMAN, George (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification [on line]. In: Journal of Machine Learning Research, No. 3. p. 1289-1305. e-ISSN: 1533-7928. <http://www.jmlr.org/papers/volume3/forman03a/forman03a_full.pdf> [consult: 12/11/2015].

FRAKES, Williams B. & BAEZA YATES, Ricardo (eds.) (1992). Information Retrieval: Data Structures and Algorithms. New York (USA): Financial Times / Prentice Hall. 464 p. ISBN: 978-0134638379

HOTHO, Andreas; NÜRNBERGER, Andreas & PAAß, Gerhard (2005). Brief survey of text mining

[online]. In: Journal for Language Technology and Computational Linguistics, LCCL, Vol. 20, No. 1 (may). Mannheim (Germany): The German Society for Computational Linguistics and Language Technology, GSCL. p. 19-62. ISSN: 2190-6858 <http://www.jlcl.org/2005_Heft1/19-62_HothoNuernbergerPaass.pdf> [consult: 23/05/2016]

KORFHAGE, Robert R. (1977). Information storage and retrieval [online]. New York (USA): Wiley. 368 p. ISBN: 978-0-471-14338-3. <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471143383.html> [consult: 23/11/2015].

MARBOT DÍAZ, Evelyn & ROJAS BENÍTEZ, José Luis (2015). Herramienta para la evaluación de una publicación científica digital [en línea]. En: Ciencias de la Información, Vol. 46, No. 2 (may-ago). La Habana (Cuba): Instituto de Información Científica y Tecnológica. p. 49-55. e-ISSN: 0864-4659 <http://www.redalyc.org/articulo.oa?id=181441052008> [consulta: 23/05/2016].

PASCUAL GONZÁLEZ, Damaris (2010). Algoritmos de Agrupamiento basados en densidad y validación de clusters [en línea]. Tesis Doctoral. Castellón (España): Universitat Jaume I, Departamento de Lenguajes y Sistemas Informáticos. 183 p. <https://dialnet.unirioja.es/servlet/tesis?codigo=21464>, <http://www.cerpamid.co.cu/sitio/files/DamarisTesis.pdf> [consulta: 15/12/2015]

RODRÍGUEZ BÁRCENAS, Gustavo (2013). Red de Inteligencia Compartida Organizacional como soporte a la toma de decisiones. Tesis Doctoral (Doctor en Ciencias de la Información). Granada (España): Universidad de Granada, Departamento de Información y Comunicación. 352 p.

RODRÍGUEZ BÁRCENAS, Gustavo; CEVALLOS, Alex; RUBIO PEÑA, Jorge & TORRES TAMAYO, Enrique (2016). Levels of similarity in user profiles based cluster techniques and multidimensional scaling [online]. In: International Journal of Systems Applications, Engineering & Development, Vol. 10. New York (USA): North Atlantic University Union. p. 56-64. e-ISSN: 2074-1308. <http://www.naun.org/main/UPress/saed/2016/a202014-058.pdf> [consult: 24/05/2016].

RODRÍGUEZ ROCHE, S. y LEIVA RAMOS, A. (2009). Las tecnologías de información en la actividad editorial: tendencias, contextos y perspectivas [en línea]. En: Acimed. vol. 20, no. 5. La Habana (Cuba): Editorial Ciencias Médicas. p. 56-65. ISSN: 1024-9435 <http://scielo.sld.cu/pdf/aci/v20n5/aci051109.pdf> [consulta: 25/10/2015].

SAMPER, Juan José (2005). Estudio y evaluación de un sistema inteligente para la recuperación y el filtrado de información de internet [en línea]. Tesis Doctoral (Doctor en Informática). Granada (España): Universidad de Granada. 142 p. <http://hera.ugr.es/tesisugr/15764552.pdf> [consulta: 23/03/2016].

TAN, Ah-Hwee (1999). Text Mining: The state of the art and the challenges, [on line]. In: PAKDD Workshop on Knowledge discovery from Advanced Databases, KDAD'99 (26/04/1999) Beijing (China): Kent Ridge Digital Labs. Proceedings, p. 1-6. <http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf> [consult: 14/11/2015].

VAN RIJSBERGEN, Cornelius Joost (1979). Information Retrieval [online]. 2 ed. Oxford (UK): Butterworth-Heinemann. 224 p. ISBN: 978-0408709293 <http://openlib.org/home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf> [consult: 16/04/2016].