



Revista e-Ciencias de la Información  
ISSN: 1659-4142  
[revista.ebci@ucr.ac.cr](mailto:revista.ebci@ucr.ac.cr)  
Universidad de Costa Rica  
Costa Rica

Rodríguez-Bárcenas, Gustavo

Método de algoritmo de clúster para el análisis del perfil de investigadores científicos  
Revista e-Ciencias de la Información, vol. 12, núm. 2, 2022, Julio-Diciembre, pp. 160-184  
Universidad de Costa Rica  
San José, Costa Rica

DOI: <https://doi.org/10.15517/eci.v12i2.50456>

Disponible en: <https://www.redalyc.org/articulo.oa?id=476871069010>

- ▶ Cómo citar el artículo
- ▶ Número completo
- ▶ Más información del artículo
- ▶ Página de la revista en redalyc.org

redalyc.org

Sistema de Información Científica Redalyc  
Red de Revistas Científicas de América Latina y el Caribe, España y Portugal  
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto



UNIVERSIDAD DE  
COSTA RICA

EBCI

Escuela de  
Bibliotecología y Ciencias  
de la Información



# e-Ciencias de la Información

Método de algoritmo de clúster  
para el análisis del perfil de  
investigadores científicos

*Gustavo Rodríguez-Bárcenas*

Recibido: 17/03/2022 | Corregido: 28/06/2022 | Aceptado: 29/06/2022

e-Ciencias de la Información, volumen 12, número 2, Jul-Dic 2020

DOI: <http://dx.doi.org/10.15517/eci.v12i2.50456>

ISSN: 1649-4142



¿Cómo citar este artículo?

Rodríguez-Bárcenas, G. (2022). Método de algoritmo de clúster para el análisis del perfil de investigadores científicos. *e-Ciencias de la Información*, 12(2). doi: [10.15517/eci.v12i2.50456](http://dx.doi.org/10.15517/eci.v12i2.50456)

# Método de algoritmo de clúster para el análisis del perfil de investigadores científicos

Cluster algorithm method for profile analysis of scientific researchers

Gustavo Rodríguez-Bárcenas<sup>1</sup> 

1

## RESUMEN

El aumento de la producción científica convierte en un desafío la tarea de identificar patrones y rasgos particulares que caractericen a los investigadores. Lograr establecer niveles de compatibilidad y similaridad entre actores en un contexto de investigación científica a partir de sus perfiles requiere de un proceso rápido y apropiado. El objetivo de este artículo es evaluar los niveles de similaridad, distancia euclíadiana y compatibilidad entre vectores de investigadores, a partir de algoritmos de agrupamiento, escalamiento multidimensional, principios del modelo espacio-vectorial y atributos de sus perfiles científicos, considerando las terminologías que se abordan en su producción científica. Se utilizaron métodos teóricos y empíricos, incluyendo técnicas y herramientas de minería de texto. La aplicación del procedimiento en el Centro de Estudios de la Energía y Tecnología Avanzada de Cuba (CEETAM) y la Universidad Técnica de Cotopaxi (UTC) en Ecuador, evidenció su efectividad. Como resultado se pudo identificar los profesionales con mayores niveles de coincidencia en áreas a fines y líneas de investigación, lo que propicia el establecimiento de Comunidades Colectivas de Conocimientos; se pudo demostrar que los métodos empleados pueden ser integrados a las TIC, resultando en la obtención de relaciones perceptuales entre los investigadores y expresando los grupos que se forman a partir de conglomerados de observaciones en cada subcategoría y dominios de conocimientos de los dos casos de estudio analizados.

**Palabras Clave:** análisis de conglomerados, perfiles de usuario, modelo de espacio vectorial.

Método de algoritmo de clúster para el análisis del perfil de investigadores científicos  
Gustavo Rodríguez-Bárcenas

1 Universidad Técnica de Cotopaxi, Latacunga, Cotopaxi, ECUADOR, [gustavo.rodriguez@utc.edu.ec](mailto:gustavo.rodriguez@utc.edu.ec)

## ABSTRACT

The increase in scientific production makes it a challenge to identify particular patterns and traits that characterize researchers. Establishing levels of compatibility and similarity between actors in a scientific research context from their profiles requires a rapid and appropriate process. The objective of this article is to evaluate the levels of similarity, Euclidean distance and compatibility between vectors of researchers, based on clustering algorithms, multidimensional scaling, principles of the vector-space model and attributes of their scientific profiles, considering the terminologies addressed in their scientific production. Theoretical and empirical methods were used, including text mining techniques and tools. The application of the procedure in the Advanced Energy and Technology Study Center from Cuba and the Cotopaxi Technical University from Ecuador, evidenced its effectiveness. As a result, it was possible to identify professionals with higher levels of coincidence in areas and lines of research, which favors the establishment of Collective Communities of Knowledge; it was possible to demonstrate that the methods used can be integrated to ICT, resulting in obtaining perceptual relationships between researchers and expressing the groups formed from clusters of observations in each subcategory and knowledge domains of the two case studies analyzed.

2

**Keywords:** *cluster analysis, user profiles, vector space model.*

## 1. INTRODUCCIÓN

El elemento clave de cualquier sistema de información y la razón de ser de cualquier entidad que se dedique a la prestación de servicios de información es el usuario que satisface estas necesidades, intereses y demandas de información. Para toda oferta de información se convierte en un conocimiento crítico del usuario, que es considerado el alfa y el omega de dichas ofertas. El usuario es el protagonista de la pantalla del ordenador, es el principio y el fin del ciclo de transferencia de información: pide, analiza, evalúa y recrea la información (Bárcenas et al., 2016; Day, 2011).

En los estudios de Day (2011) y Sun (2012) se resumen al usuario como:

- Persona relacionada, real o potencialmente, con el uso de sistemas de información.
- Actores sociales interactuantes y en comunicación, en una sociedad en constante cambio y conflicto.
- Seres humanos relacionados socialmente, que pertenecen a diferentes clases sociales y poseen capitales culturales, hábitos y visiones diferentes del mundo.

Hoy en día la web es un medio utilizado por los cibernautas para intercambiar ideas. Así, la inteligencia colectiva se ha definido como la suma de estas pero a niveles personales para crear un sistema colaborativo inclusivo, que suma el conocimiento de varios individuos con el fin de generar uno colectivo que se entrega de forma sencilla en una democracia (Martelete y Braz, 2004); este concepto también puede relacionarse con su integración a escenarios procedimentales de lo que se conoce como de enjambre, ya que se ocupa del diseño y desarrollo de sistemas interactivos multiagente que cooperan para cumplir un objetivo específico que no puede ser alcanzado por un solo agente (Almufti, 2019).

Todo sistema informático de alguna manera se desarrolla para satisfacer las necesidades de formación o información de los usuarios que hoy en día trabajan para implementar nuevos métodos que permitan una mejor identificación y representación de la información y el conocimiento y por otro lado, las personas que poseen dicho conocimiento ya sea de forma implícita o explícita, con el fin de que los usuarios puedan inferir posicionamientos y establecer relaciones en base al análisis resultante de las representaciones obtenidas en los mismos, siguiendo las premisas que identifican el contexto de las TIC en la actualidad y la referencia a un nuevo paradigma de representación y visualización de información.

Se evidencia como una cuestión real el inmenso desarrollo científico y tecnológico, esto constituye uno de los aspectos que caracterizan el fin del segundo milenio y que también marcará el siglo XXI. Los niveles de avance alcanzado por la investigación científica y las ulteriores aplicaciones tecnológicas en los últimos tiempos se consideran elevado, un crecimiento tan grande ha tenido muchos y variados efectos no sólo en campos particulares de la investigación y la industria, sino en una gama enorme de aspectos de la vida cotidiana, llegando a transformar los hábitos y costumbres de sociedades enteras.

En este sentido, el aumento de la producción científica convierte en un desafío la tarea de identificar patrones y rasgos particulares que caractericen a las personas investigadores. Lograr establecer niveles de compatibilidad y similaridad entre actores en un contexto de investigación científica a partir de sus perfiles requiere de un proceso eficientes y eficaces. Una problemática acentuada en muchas instituciones de educación superior donde tienen algún tipo de modelo de adopción de tecnologías, es identificar patrones en su comunidad que conlleven a tomar decisiones respecto a su capital intelectual y usarlo en aras de su propio desarrollo, ello será posible si se analizan adecuadamente los perfiles de sus profesionales investigadores a partir de su producción científica. Los perfiles presentan rasgos particulares, cada uno tiene sus propias características orientado hacia sus intereses y necesidades, de acuerdo con su desarrollo cognoscitivo, de su experiencia, lo cual los hacen únicos, de éstos pueden derivarse innumerables estudios adaptando principios del modelo espacio-vectorial para minería de texto, análisis de clústeres y escalamiento multidimensional, así como métodos para recopilar información adecuada y tomando la base léxico semántica que encierran estos, precisamente esto constituye el principal propósito de la investigación.

El objetivo del presente trabajo es evaluar los niveles de similaridad, distancia euclíadiana y compatibilidad entre vectores de investigadores, a partir de algoritmos de agrupamiento, escalamiento multidimensional, principios del modelo espacio-vectorial y atributos de perfiles de investigadores considerando su producción científica, que permita la identificación de patrones y rasgos particulares de acuerdo a las terminologías que los caracterizan.

## 2. REFERENTE TEÓRICO

### 2.1 Definición de los perfiles de usuario en los sistemas informáticos

Para Samper (2005) citado por Bárcenas et al. (2016) perfil es una palabra que viene del latín *profilare*, que significa diseñar contornos. Un perfil es un modelo de un objeto, una representación compacta que describe sus características más importantes, que puede crearse en la memoria de un ordenador y puede utilizarse para representar el objeto en las tareas de cálculo.

También se reconoce el perfil de origen, derivado de la psicología, entendido como un conjunto de medidas diferentes de una persona o grupo, cada una de las cuales se expresa en la misma unidad de medida. Es decir, ciertas características de un individuo se miden mediante pruebas que dan diferentes puntuaciones, estas puntuaciones son su perfil, que se utiliza con fines de diagnóstico según lo analizado de los trabajos de Rodríguez-García et al. (2014) además las investigaciones de Velásquez et al. (2014). Teniendo en cuenta el enfoque anterior se puede entender el perfil del usuario como un conjunto de rasgos distintivos que lo caracterizan.

Se han realizado estudios en esta área, destacando el caso de Ratheeshkumar et al. (2018) donde se utilizaron los perfiles de los usuarios como parte de un evento de datos sociales en la web para descifrar las implicaciones semánticas de las consultas y captar las necesidades de datos de los clientes. También acorde a la función de investigador científico pueden existir perfiles que sirven para identificar en cuáles líneas ha incursionado y con cuáles áreas de conocimientos se ha relacionado.

Los algoritmos diseñados para resolver este problema de información basan sus cálculos en perfiles de usuario en los que se guardan representaciones de los intereses de los usuarios como el caso de la investigación de Degemmis et al. (2006), este autor realiza una investigación aplicando el algoritmo clásico de Rocchio para la categorización de textos, capaz de descubrir las preferencias de los usuarios a partir del análisis de la descripción de textos de artículos en el catálogo online de sitios web de comercio electrónico. Los experimentos se llevaron a cabo en varios conjuntos de datos, y los resultados se compararon con los obtenidos a partir de una programación lógica inductiva (ILP) y una probabilística.

En el caso de un perfil de usuario de un sistema de software, puede comprender tanto los datos personales como las características del sistema informático, así como: los patrones de comportamiento, los intereses personales y las preferencias. Este modelo de usuario se representa mediante una estructura de datos adecuada para su análisis, recuperación y utilización. En términos informáticos; un perfil de usuario es la representación de un conjunto de características que describen a una persona, en su papel de usuario de un sistema. Se almacena en la mayoría de los casos en forma de pares atributo-valor. El sistema almacena, analiza y pone a disposición esta información.

Cada persona investigadora tiene sus propios intereses y necesidades, en función de su desarrollo cognitivo, del entorno en el que se desenvuelve y de su experiencia vital, que la hacen única, de los perfiles de investigadores se pueden determinar el nivel de interacción entre estos, en función de los campos de experiencia recogidos en su perfil, el nivel de compatibilidad o la distancia entre ellos, los clústeres de investigadores que responden a los parámetros definidos en su perfil. El perfil de un investigador contiene un conjunto de datos, en su mayoría de naturaleza textual, aunque los avances tecnológicos han llevado a incorporar imágenes de texto, gráficos, etc., el abanico de información que se recoge es cada vez mayor, investigando para ello la naturaleza textual o las terminologías que los compongan.

El compendio de terminologías, palabras, intereses que caracterizan a un investigador pueden obtenerse por varios métodos aplicando las tecnologías, estas características pueden ser almacenadas en un sistema de base de datos, considerando esto una base de datos está compuesta por tablas, campos, claves y relaciones entre ellas, una tabla puede considerarse como una matriz en la que cada fila puede representar un investigador y cada columna indica la presencia o no de un término, palabra o interés obtenido en su perfil correspondiente, en lo adelante término.

Para Bárcenas (2016), podemos considerar una matriz de datos de perfiles de investigadores ( $U$ ), compuesta por los investigadores  $u_i$ , donde se ha obtenido un conjunto de términos ( $T$ ), formado por  $n$  términos  $t_j$ , en la que cada investigador  $u_i$  contiene un número de términos, como resultado de los campos suscritos en su perfil. Así, es posible representar a cada investigador como un vector perteneciente a un espacio  $n$ -dimensional, siendo  $n$  el número de términos ingresados en el perfil que forman el conjunto  $T$ :

$$u_i = (t_{i1}; t_{i2}; t_{i3}; \dots; \dots; t_{in})$$

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

5

Método de algoritmo de clúster para el análisis del perfil de investigadores científicos  
Gustavo Rodríguez-Bárcenas

### 3. METODOLOGÍA

#### 3.1 Enfoque

El enfoque de la investigación es mixto debido a que se obtiene y analiza los datos (cuantitativo) aplicando las técnicas K-Media y MDS y el Clustering Aglomerativo Jerárquico constituyendo herramientas que permiten la visualización de la información considerando la similitud y las distancias; con esos resultados se aplica la inducción (cualitativo) para derivar las posibles explicaciones basadas en los fenómenos observados. Se efectúa la creación del perfil de investigador en dos instituciones de educación superior como son el Centro de Estudio de la Energía y Tecnologías de Avanzada de la Universidad de Moa (CEETAM), en Cuba y la Universidad Técnica de Cotopaxi (UTC) en Ecuador a través de un sistema informático desarrollado para ambos casos y el análisis de la frecuencia de aparición de las terminologías que representan la información suministrada por estos. Finalmente se determina la similitud o proximidad con entre investigadores

### 3.2 Otros criterios metodológicos considerados

Para la presente investigación, se consideraron los siguientes criterios:

- a. Como unidad de análisis se toman las producciones científicas de las personas profesionales de la Universidad Técnica de Cotopaxi del Ecuador y El Centro de Estudio de la Energía y Tecnología de Avanzada de la Universidad de Moa, en Cuba, siendo ambas instituciones de educación superior públicas, autónomas, laicas y gratuitas.
- b. Se realizó una investigación descriptiva pues se analiza y caracteriza el fenómeno relacionados con los contextos de los perfiles de investigadores de dos instituciones de educación superior, se evalúan los niveles de compatibilidad entre los actores involucrados, alineándose a las nuevas tendencias mundiales de los preceptos de minería de texto, para la obtención de patrones que permiten relacionar a los autores intelectuales de producción científica en varios dominios de conocimiento.

6

### 3.3 Consideraciones para la creación del perfil de usuario

Para crear el perfil del investigador se toman las premisas descritas por Samper (2005) se toma como norma seguir el método explícito ya que se requiere que el perfil se construya a partir del propio análisis y valoración que hace el propio usuario de él mismo, de acuerdo con sus intereses y motivaciones, para ello también se consideran los siguientes criterios:

- a. Adquisición de datos: la adquisición de los datos se toma como información explícita del método de referencia.
- b. Representación del perfil: se utiliza el método de razonamiento inductivo ya que se avanza de lo particular a lo general, por lo que se monitorea la interacción del usuario con el sistema, esto permitirá reutilizar la información en su perfil.
- c. *Feedback* del usuario: se considera el método de *feedback* explícito, ya que se obtiene según Samper (2005) preguntando directamente al investigador mediante formularios desarrollados en el sistema informático. Se le puede pedir que rellene un cuestionario o que emita un juicio de valor sobre algo, o simplemente que edite su perfil añadiendo nuevos parámetros relacionados con su producción científica.

### 3.4 Campos del perfil de los investigadores

Se tomó como referencia un usuario tipo, o sea uno que se desenvuelve en un contexto investigativo, para ello se muestran los datos que definirán el perfil de usuario del sistema. Junto a ellos es referido al conjunto de experiencias, laborales y educacionales, de una persona o usuario. Para la confección de la matriz de términos serán usados los campos que describen el perfil del investigador donde mayor relevancia exista, como son en la identificación de sus conocimientos, especialidades, etc., estos se relacionan a continuación:

- Nombre de la formación.
- Nombre de la formación adicional.
- Especialidades.
- Temas de interés y descriptores temáticos.
- Resúmenes y palabras claves de los libros o capítulos de libros publicados.
- Resúmenes y palabras claves de los artículos publicados.
- Palabras claves de los trabajos presentados en eventos, seminarios y conferencias.

### 3.5 Proceso de extracción de términos del perfil basados en estadística

#### a) Peso relacionado con los términos en perfiles de usuarios

Según la expresión (1) el proceso de construcción de los vectores – investigadores en las matrices derivado de la base de datos de perfiles comprenderá la extracción de los términos con los cuales se realizará la representación de los investigadores, extrayendo los contenidos de información para conformar el corpus de análisis. La tarea fundamental de este método estará dada por la asociación automática de la representación de cada usuario en función de los contenidos de información de este, o sea, determinar los pesos de cada término extraído de su perfil en el vector investigador  $u_i$ . Su función será:

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

La representación de cada componente del vector-investigador, de los que están referenciados en el perfil tendrá un valor diferente de 0, mientras que los que no están referenciados tendrán un valor nulo o 0. La frecuencia de aparición de un término en un perfil de un investigador de cierta forma determina su importancia en él, sugiriendo que dichas frecuencias pueden ser utilizadas para resumir el área de conocimiento en que se mueve el usuario o los principales intereses del mismo.

Siguiendo lo que describe el espacio vectorial para el Modelo de Sistemas de Recuperación de Información acorde a los trabajos de Al-Anzi y AbuZeina (2018), así como de Kastrati y Imran (2019), y dando continuidad a los métodos usados para almacenar los términos recogidos en el perfil de cada usuario, se continúa con el proceso de selección, a ello le sigue determinar la importancia o peso de cada término en el vector-investigador. El cálculo de la importancia o peso de cada término se conoce como ponderación del término.

Gerald Salton (1989), utiliza este concepto de peso en su modelo de recuperación basado en el espacio vectorial. En dicho modelo, se forma una matriz término / documento que representa la base de datos. Cada vector de la matriz representa un documento; cada elemento del vector tendrá valor 0 (cero) si dicho documento no contiene el término; o el valor del peso del término si lo contiene.

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

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

Donde  $N$  es el número de documentos de la colección, y  $n_i$  el número de documentos donde se menciona al término  $i$ -ésimo, si asociamos al caso de la presente investigación a  $N$  con  $U$  como el número de investigadores de la base de datos de perfiles, y  $n_i$  como el número de investigadores que contienen en su perfil el término  $i$ , entonces es posible determinar la importancia o peso de cada término en su perfil, quedando para ello una matriz de pesos como muestra la siguiente matriz.

	$t_1$	$t_2$	$t_3$	...	$t_n$
$User_1$	$w_{11}$	$w_{12}$	$w_{13}$		$w_{1n}$
$User_2$	$w_{21}$	$w_{22}$	$w_{23}$	...	$w_{2n}$
$User_3$	$w_{31}$	$w_{32}$	$w_{33}$		$w_{3n}$
	:	:	:	..	:
$User_n$	$w_{n1}$	$w_{n2}$	$w_{n3}$	...	$w_{nn}$

### b) Similitud entre los investigadores

Se tiene en consideración el cálculo de similitud entre los vectores que componen la matriz de peso, que en esencia son los vectores-investigadores, para obtener el grado de relevancia de un investigador  $u_i$  según su perfil con respecto a los demás que componen la matriz, es posible establecer la similaridad entre los vectores de esta matriz, o sea cada vector lo constituirá un investigador y será posible determinar la similitud de cada investigador con respecto a los demás. El sistema toma un valor real que será tanto mayor cuanto más similares sean los investigadores que se analizan. Existen diferentes funciones para medir la similitud entre vectores, todas ellas están basadas en considerar a ambos como puntos en un espacio  $n$ -dimensional, la función del coseno es una de ellas:

Función del coseno:

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

Donde  $A_i$  y  $B_j$  son, respectivamente, los pesos asociados al término  $t_j$  en la representación de los investigadores  $A$  y  $B$ . Las funciones típicas de similitud generan valores entre 0 para elementos sin similitud y 1 para elementos completamente iguales. Una matriz de similitud puede quedar representada simétricamente, donde cada elemento  $\delta_{ij}$  de  $M$  representa la similaridad entre el estímulo  $i$  y el estímulo  $j$  como se muestra en  $M$ :

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

9

### 3.6 Escalamiento multidimensional para representar perceptualmente a los investigadores

El escalamiento multidimensional (MDS) es una técnica de representación espacial que trata de visualizar sobre un mapa un conjunto de estímulos cuya posición relativa se desea analizar (Machado y Lopes, 2020). El método estará centrado en obtener una representación espacial que visualice la relación perceptual entre los distintos investigadores, de manera que se podrá observar qué investigadores se encuentran cercanos o lejanos entre ellos. Esto es posible debido a la transformación de la similitud entre ellos en distancias susceptibles de ser representadas en un espacio multidimensional.

El procedimiento, en términos muy generales, sigue algunas ideas básicas en la mayoría de las técnicas. El punto de partida es una matriz de similaridad entre  $n$  objetos, con el elemento  $\delta_{ij}$  en la fila  $i$  y en la columna  $j$ , que representa la similaridad del objeto  $i$  al objeto  $j$ . También se fija el número de dimensiones,  $p$ , para hacer el gráfico de los objetos en una solución particular. Generalmente el camino que se sigue según Ashby (2014); Cambria et al. (2014); Dunn-Rankin et al. (2014); Žilinskas (2007) es:

1. Arreglar los  $n$  objetos en una configuración inicial en  $p$  dimensiones, esto es, suponer para cada objeto las coordenadas  $(x_1, x_2, \dots, x_p)$  en el espacio de  $p$  dimensiones.
2. Calcular las distancias euclidianas entre los objetos de esa configuración, esto es, calcular las  $d_{ij}$  que son las distancias entre el objeto  $i$  y el objeto  $j$ .

$$d(O_i, O_j) = \sqrt{\sum_{k=1}^n (x_k(O_i) - x_k(O_j))^2}$$

Donde  $O_i$  y  $O_j$  son los objetos para los cuales se desea calcular la distancia,  $n$  es el número de características de los objetos del espacio y  $x_k(O_i)$ ,  $x_k(O_j)$  es el valor del atributo  $k$ -ésimo en los objetos  $O_i$  y  $O_j$ , respectivamente.

De tal manera también debe verificarse los tres axiomas siguientes:

- $d(x, y) \geq 0 \quad \forall x, y \in X, y d(x, y) = 0$  (*si y solo si*  $x = y$ )
- $d(x, y) = d(y, x) \quad \forall x, y \in X$  (*simetría*)
- $d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in X$  (*desigualdad triangular*)

3. Hacer una regresión de  $d_{ij}$  sobre  $\delta_{ij}$ . Esta regresión puede ser lineal, polinomial o monótona. Utilizando el método de los mínimos cuadrados se obtienen estimaciones de los coeficientes  $a$  y  $b$ , y de ahí puede obtenerse lo que genéricamente se conoce como una "disparidad".

$$\hat{d}_{ij} = \hat{a} + \hat{b}\delta_{ij}$$

Si se supone una regresión monótona, no se ajusta una relación exacta entre  $d_{ij}$  y  $\delta_{ij}$ , sino se supone simplemente que si  $\delta_{ij}$  crece, entonces  $d_{ij}$  crece o se mantiene constante.

4. A través de algún estadístico conveniente, se mide la bondad de ajuste entre las distancias de la configuración y las disparidades. Existen diferentes definiciones de este estadístico, pero la mayoría surge de la definición del llamado índice de stress.

Uno de los criterios más utilizados es el siguiente:

$$STRESS1 = \sqrt{\frac{\sum \sum (d_{ij} - \hat{d}_{ij})^2}{\sum \sum d_{ij}^2}}$$

Todas las sumatorias sobre  $i$  y  $j$  van de 1 a  $p$  y las disparidades dependen del tipo de regresión utilizado en el tercer paso del procedimiento.

5. Las coordenadas  $(x_1, x_2, \dots, x_p)$  de cada objeto se cambian ligeramente de tal manera que la medida de ajuste se reduzca. Quedan representadas la matriz de distancia ( $D$ ) y la matriz de coordenadas ( $X$ ) de los estímulos en un espacio de  $n$  dimensiones (para el caso de la investigación solo 2 dimensiones).

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2n} \\ d_{31} & d_{32} & d_{33} & \dots & d_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \dots & d_{nn} \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

### 3.7 Análisis de clúster para identificar conglomerados de investigadores

El clúster puede establecer la jerarquía en términos de grupos de investigadores, basándose en las matrices de similitud y distancias obtenidas. Así, cada investigador puede identificarse con el grupo al que pertenece en función de su distancia o similaridad.

Los algoritmos de partición tratan de descubrir clúster reubicando iterativamente puntos entre subconjuntos. El algoritmo *k-means* según investigaciones realizadas por Sun et al. (2019); Du (2019) y Joshi et al. (2020) es uno de los más simples y conocidos algoritmos de agrupamiento. Está basado en la optimización del error cuadrático, que sigue una forma fácil para dividir una base de datos dada en *k* grupos fijados a priori. La idea principal es definir *k centroides* (uno para cada grupo) y, luego, ubicar los restantes puntos en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada clúster y reubicar nuevamente los puntos en cada grupo. El proceso se repite hasta que no haya cambios en la distribución de los puntos de una iteración a la siguiente.

11

Método de algoritmo de clúster para el análisis del perfil de investigadores científicos  
Gustavo Rodríguez-Bárcenas

## 4. RESULTADOS

Los resultados obtenidos responden como se mencionó a un caso de estudio específico relacionado con el CEETAM que se encarga de todo lo relacionado con la Eficiencia Energética y Uso Racional de la Energía (EEURE), los perfiles han sido creados como referencia a algunos investigadores de este centro de estudios. El conocimiento más importante de la investigación en esta área se refleja en el perfil construido por él mismo; otro contexto de investigación es el caso de los investigadores de la UTC, específicamente las carreras universitarias de Medio Ambiente y Turismo, de esta manera se asocia un corpus terminológico a cada perfil de investigación de estas Instituciones de Educación Superior, así el primer caso de Cuba y el segundo caso de Ecuador.

Para una mejor comprensión se han considerado solo algunos investigadores del CEETAM y de las carreras de Medio Ambiente y Turismo de la UTC, de manera que puedan visualizar los resultados para ser presentados de una mejor forma.

### 4.1 Caso: Centro de Estudio de la Energía y Tecnologías de Avanzada de la Universidad de Moa, Cuba

En el listado que se muestra a continuación contiene información de algunas personas investigadores seleccionadas al azar, constituyen sólo una muestra intencional con el objetivo de revelar la funcionalidad del caso, por lo que el número de términos y otros elementos constituyen procedimientos de cálculo de la similitud y la distancia descritos en los métodos.

1. *egongora*: especialista en termodinámica, refrigeración y climatización.
2. *rmontero*: especialista gestión eficiente de la energía total.
3. *iromero*: especialista en máquinas eléctricas.
4. *alegra*: especialista en modelación matemática, simulación y metodología de la investigación.
5. *lrpuron*: especialista en inteligencia artificial aplicada a los procesos industriales.
6. *yretirado*: especialista en el secado de mineral con el uso de la energía solar térmica.
7. *grbarcenas*: especialista en tecnologías de la información y la comunicación en los procesos.
8. *yaguilera*: especialista en redes y comunicaciones informáticas.
9. *dgonzalezr*: especialista en informática 1.
10. *eromero*: especialista en informática 2.

A partir de la selección de campos tomados en cuenta en los 10 usuarios previamente seleccionados, se obtienen un total de 470 elementos léxico-semánticos realizados entre términos y frases que identifican, dominios de conocimiento de especialidad, palabras clave, entre otros. Contando las ocurrencias de cada término en los perfiles de los usuarios seleccionados, se obtiene la frecuencia de los términos en estos perfiles, su magnitud está representada por la Tabla 1.

TABLA 1. MATRIZ DE FRECUENCIA DE LOS TÉRMINOS (TFI) EN LOS PERFILES DE LOS USUARIOS<sup>2</sup>

NO.	ID TÉRMINOS	39	40	41	42	43	44	47	49	50	51
		FRECUENCIAS OBTENIDAS									
1	Acceso Remoto	0	0	0	0	0	0	2	2	1	1
2	accionamiento	0	0	2	0	0	0	0	0	0	0
3	adherencia en menas lateríticas	1	0	0	0	0	0	0	0	0	0
4	agrupamiento	0	0	0	0	0	0	1	0	0	0
5	agua	1	0	0	0	0	0	0	0	0	0
6	Agua Caliente	1	0	0	0	0	0	0	0	0	0
7	Agua caliente sanitaria	0	1	0	0	0	0	0	0	0	0
8	Ajuste de Curvas	0	0	0	1	0	0	0	0	0	0
9	Algoritmo Iterativo	0	0	0	1	0	0	0	0	0	0
10	Algoritmos	0	0	1	1	0	0	2	2	0	0
...	...	...	...	...	...	...	...	...	...	...	...
448	utilidad del error de estimación	0	0	0	1	0	0	0	0	0	0
449	valores de una variable	0	0	0	1	0	0	0	0	0	0
450	Vapor	1	0	0	0	0	1	0	0	0	0
451	variabilidad	0	0	0	1	0	0	0	0	0	0
452	Variogramas adaptativos	0	0	0	1	0	0	0	0	0	0
453	variogramas dinámicos	0	0	0	1	0	0	0	0	0	0
454	velocidad del viento	0	0	0	2	0	0	1	0	0	0
455	Ventilación	1	0	0	0	0	1	0	0	0	0
456	Video conferencia	0	0	0	0	0	0	1	1	1	0
457	Virtualización	0	0	0	0	0	0	1	1	1	1
458	Volúmenes de Sólidos Minerales Irregulares	0	0	0	1	0	0	0	0	0	0
459	volúmenes geólogo - mineros	0	0	0	1	0	0	0	0	0	0
460	Voz sobre IP	0	0	0	0	0	0	1	1	1	0
461	Web	0	0	0	0	0	0	1	1	3	2
462	Web 2.0	0	0	0	0	0	0	1	1	1	1
463	wikis	0	0	0	0	0	0	1	1	1	1
464	Yacimiento	0	0	0	3	0	0	0	0	0	0
465	yacimiento Merceditas	0	0	0	1	0	0	0	0	0	0
466	Yacimiento Punta Gorda	0	0	0	3	0	0	0	0	0	0
467	Yacimientos Lateríticos	0	0	0	2	0	0	0	0	0	0
468	yacimientos lateríticos cubanos	0	0	0	1	0	0	0	0	0	0
469	yacimientos lateríticos de Ni	0	0	0	1	0	0	0	0	0	0
470	Zimbra	0	0	0	0	0	0	1	1	1	0

2 Se han colocado algunas muestras debido a la extensión de la tabla

La expresión (2) suponiendo que el número de usuarios seleccionados ( $N$ ) es igual a 10, la matriz de pesos ( $W$ ) de los elementos léxicos contenidos en los perfiles de los usuarios del sistema. A partir de la aplicación de la función coseno en la ecuación (4) y de los valores de peso obtenidos a través, se obtiene como resultado una matriz simétrica de similitudes entre investigadores, como se observa en la Tabla 2.

**TABLA 2. MATRIZ SIMILITUD USANDO LA FUNCIÓN DEL COSENO.**

ID	39	40	41	42	43	44	47	49	50	51
39	1	0.081	0.085	0.062	0.023	0.158	0.010	0.002	0	0.001
40	0.081	1	0.142	0.016	0.432	0.078	0.011	0.019	0	0
41	0.085	0.142	1	0.011	0.158	0.018	0.006	0.014	0.001	0.001
42	0.062	0.016	0.011	1	0.012	0.003	0.016	0.013	0.012	0.008
43	0.023	0.432	0.158	0.012	1	0.063	0.038	0.023	0	0
44	0.158	0.078	0.018	0.003	0.063	1	0.037	0.005	0	0
47	0.010	0.011	0.006	0.016	0.038	0.037	1	0.259	0.219	0.175
49	0.002	0.019	0.014	0.013	0.023	0.005	0.259	1	0.647	0.391
50	0	0	0.001	0.012	0	0	0.219	0.647	1	0.690
51	0.001	0	0.001	0.008	0	0	0.175	0.391	0.690	1

Dadas las similitudes obtenidas y las valoraciones empíricas realizadas por el autor, totalmente intencionadas, se plantea un nivel de compatibilidad considerando las variables lingüísticas siguientes:

#### Valor de intervalo: (S)

- $S = 0$  – Incompatibles (I).
- $0 < S < 0.1$  – Compatibilidad Extremadamente Baja (CEB).
- $\leq S < 0.25$  – Compatibilidad Muy Baja (CMuyB).
- $0.25 \leq S < 0.5$  – Compatibilidad Moderadamente Baja (CmoderadamtB).
- $S = 0.5$  - Medianamente Compatibles (MC).
- $0.5 < S < 0.75$  – Compatibilidad Moderadamente Alta (CmoderadamtA).
- $0.75 \leq S \leq 0.99$  – Compatibilidad Muy Alta (CmuyA).
- $S = 1$  – Totalmente Compatibles (C).

Se obtuvo como resultado el nivel de compatibilidad entre los investigadores seleccionados:

- **Las compatibilidades son muy bajas (CMuyB):** entre los usuarios *egongora* (ID:39) (especialista en termodinámica, refrigeración y climatización) y *yretirado* (ID:44) (especialista en secado de mineral con el uso de energía solar térmica); entre *rmontero* (ID:40) (especialista en gestión total eficiente de la energía) e *iromero* (ID:41) (especialista en máquinas eléctricas); entre *iromero* (ID:41) y *Irpuron* (ID:43) (especialista en inteligencia artificial aplicado a los procesos industriales); entre *dgonzalezr* (ID:50) (especialista en informática 1) y *grbarcenas* (ID:47) (especialista en TIC y gestión del conocimiento) y entre *iromero* (ID:41) (especialista en informática 2) y *grbarcenas* (ID:47).

- **Compatibilidad extremadamente baja (CEB):** entre el usuario *egongora* (ID:39) y los demás usuarios exceptuando a *yretirado* (ID:44); entre *rmontero* (ID:40) y *alegra* (ID:42) (especialista en modelación matemática, simulación y metodología de la investigación), *yretirado* (ID:44), *grbarcenas* (ID:47) y *yaguilera* (ID:49) (especialista en redes de computadoras); entre *iromero* (ID:41) y los demás usuarios con excepción de *Irpuron* (ID:43); entre *alegra* (ID:42) y los demás usuarios; entre *Irpuron* (ID:43) y *yretirado* (ID:44), *grbarcenas* (ID:47) y *yaguilera* (ID:49) y entre *yretirado* (ID:44) y *yaguilera* (ID:49).
- **Incompatibilidad (I):** entre los usuarios *rmontero* (ID:40), *Irpuron* (ID:43) y *yretirado* (ID:44) con *dgonzalezs* (ID:50) y *eromero* (ID:51).
- **Compatibilidades moderadamente bajas (CmoderadamtB):** entre los usuarios *rmontero* (ID:40) y *Irpuron* (ID:43); entre *grbarcenas* (ID:47) y *yaguilera* (ID:49); entre *yaguilera* (ID:49) y *eromero* (ID:51).
- **Compatibilidad moderadamente alta (CmoderadamtA):** entre los usuarios *yaguilera* (ID:49) y *dgonzalezs* (ID:50) y entre los usuarios *dgonzalezs* (ID:50) y *eromero* (ID:51).

15

Otro resultado es la existencia de una serie de actores seleccionados que son titulados de las mismas especialidades, pero que representan dominios de conocimiento algo distantes, ejemplo de esto son los usuarios *yaguilera* y *rmontero* ya que ambos son titulados de Ingeniería Eléctrica respectivamente, sin embargo *rmontero* representa el dominio de EERUE y *yaguilera* el dominio de sistemas telemáticos, solo une su formación y por tanto la compatibilidad es extremadamente baja con una similitud de 0.019, otros casos son el del investigador *grbarcenas* respecto a *egongora* y *yretirado*, los tres son graduados de Ingeniería Mecánica con una similitud de 0,010 y 0,037 respectivamente, *grbarcenas* respecto a ellos representa un dominio de conocimiento diferente, sin embargo, entre *yretirado* y *egongora* hay una similitud de 0,158 representando ambos dominios de conocimiento similares.

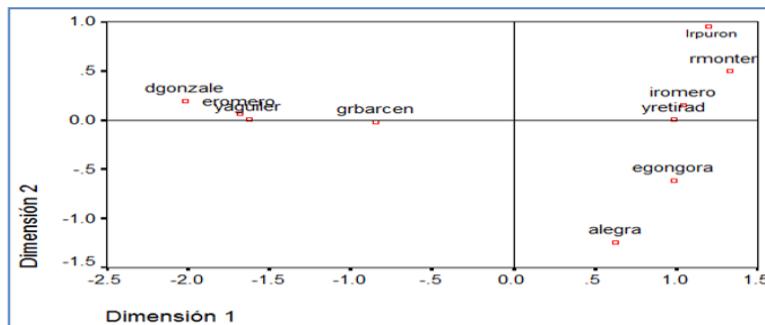
A partir de los procedimientos metodológicos para la visualización del MDS se obtuvo dos grupos, uno (grupo de la izquierda) conformado por *dgonzalezs*, *eromero*, *yaguilera* y *grbarcenas*, que representan una comunidad colectiva de conocimiento relacionados con las TIC y su aplicación; el otro grupo conformado por el resto (grupo de la derecha) representa una comunidad colectiva de conocimiento vinculada a la Eficiencia Energética y Uso Racional de la Energía (EERUE), en ambos grupos se muestran dos investigadores que de alguna manera constituyen fronteras, estos son *grbarcenas* y *alegra*, esto es resultado de la diversidad de áreas de conocimientos en que ambos incursionan.

Los resultados de la prueba del sistema fueron comparados con el software profesional SPSS, la matriz de distancia entre los actores seleccionados (Tabla 3), a partir de esto y de los procedimientos metodológicos se obtienen coordenadas en dos dimensiones, dando como resultado la representación de la Figura 1.

TABLA 3. MATRIZ DE DISTANCIA EUCLIDIANA ENTRE LOS INVESTIGADORES.

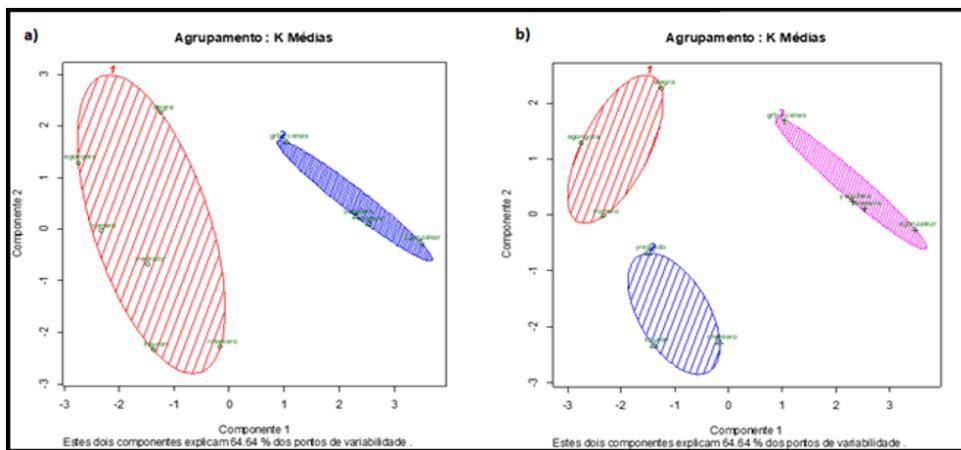
	EGONGORA	RMONTERO	IROMERO	ALEGRA	LRPURON	YRETIRADO	GRBARENAS	YAGUILERA	DGONZALEZR	EROMERO
egongora	0	1.368	1.31	1.34	1.432	1.195	1.461	1.631	1.725	1.641
rmontero	1.368	0	1.245	1.462	0.806	1.363	1.508	1.659	1.77	1.691
iromero	1.31	1.245	0	1.413	1.228	1.396	1.467	1.618	1.724	1.642
alegra	1.34	1.462	1.413	0	1.467	1.416	1.44	1.601	1.693	1.613
lrpuron	1.432	0.806	1.228	1.467	0	1.386	1.479	1.651	1.765	1.687
yretirado	1.195	1.363	1.396	1.416	1.386	0	1.423	1.621	1.719	1.637
grbarcenas	1.461	1.508	1.467	1.44	1.479	1.423	0	1.153	1.28	1.266
yaguilera	1.631	1.659	1.618	1.601	1.651	1.621	1.153	0	0.585	0.867
dgonzalezr	1.725	1.77	1.724	1.693	1.765	1.719	1.28	0.585	0	0.51
eromero	1.641	1.691	1.642	1.613	1.687	1.637	1.266	0.867	0.51	0

FIGURA 1: CONFIGURACIÓN DE ESTÍMULOS DERIVADA EN DOS DIMENSIONES.



Los resultados obtenidos en el caso de la utilización del algoritmo de agrupamiento de K-Medias, guarda relación con el métodos aplicados de MDS, en la figura 2 se establecieron dos particiones, quedando el clúster a) y b) con la aglomeración de los investigadores respectivos a cada área de conocimiento e intereses; para el caso de la b) se establecieron tres particiones con el objetivo de realizar un análisis más detallado y poder identificar grupos con intereses comunes, los mejores niveles de similitudes y distancia euclidiana están entre los investigadores *dgonzalezr* y *yaguilera* con una distancia de 0.585, también entre *dgonzalezr* y *eromero* con una distancia entre ellos de 0.510; por otro lado los investigadores *lrpuron* y *rmontero* con 0.806, con respecto al resto de los usuarios y su relación, estos se encuentran en una posición favorable.

**FIGURA 2. GRÁFICOS CORRESPONDIENTES AL USO DEL ALGORITMO K-MEDIAS ESPECIFICANDO A) 2-PARTICIONES; B) 3-PARTICIONES.**



17

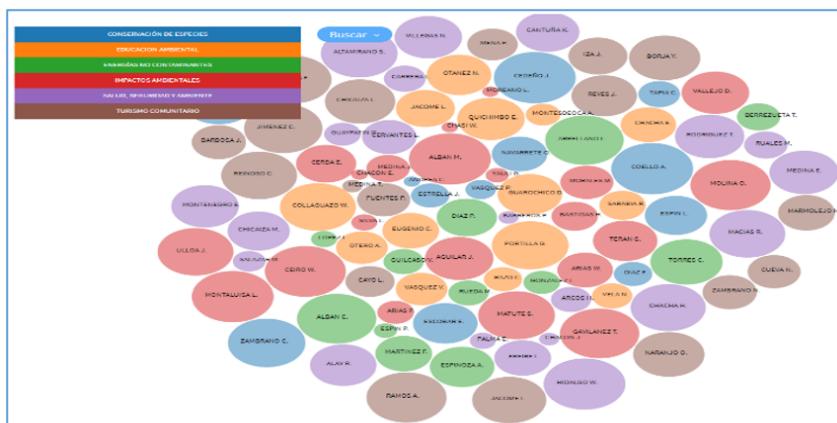
Método de algoritmo de clúster para el análisis del perfil de investigadores científicos  
Gustavo Rodríguez-Bárcenas

Comparando los resultados de las figuras 1 y 2, se aprecia que el escalamiento multidimensional y el algoritmo k-means son efectivos para establecer grupos con una estrecha relación de similitud y distancia, el caso de estudio analizado contempla a pocos usuarios como un demo. Estos algoritmos han sido implementados en la UTC, estableciendo una herramienta efectiva para visualizar la producción científica de los investigadores de la Universidad Técnica de Cotopaxi, mostrando excelentes resultados en la identificación de comunidades de expertos en áreas específicas, así como para conocer la relación entre los investigadores. Los investigadores son agrupados de acuerdo a su campo de producción científica por líneas de investigación, áreas de conocimiento, entre otros indicadores relevantes para el dominio científico.

## 4.2 Caso: Universidad Técnica de Cotopaxi, Ecuador

La Figura 3 muestra la producción científica en términos de artículos científicos, libros y conferencias, los círculos de colores representan las sublíneas de investigación de las carreras de Medio Ambiente y Turismo, los radios de los círculos representan la cantidad de producción científica de acuerdo a los indicadores mencionados. Investigadores como Coello, Arrellano, Montaluisa, Cedeño y otros son los líderes en estas especialidades en cambio Moreano, Chasi, Yauli y otros tienen radios más bajos por lo que su producción científica ha sido baja. En la Figura 4 hay muchos resultados deducibles, como cuál es la sublínea de investigación según su color con mayor productividad o qué nivel de proximidad tiene una sublínea de investigación respecto a las demás.

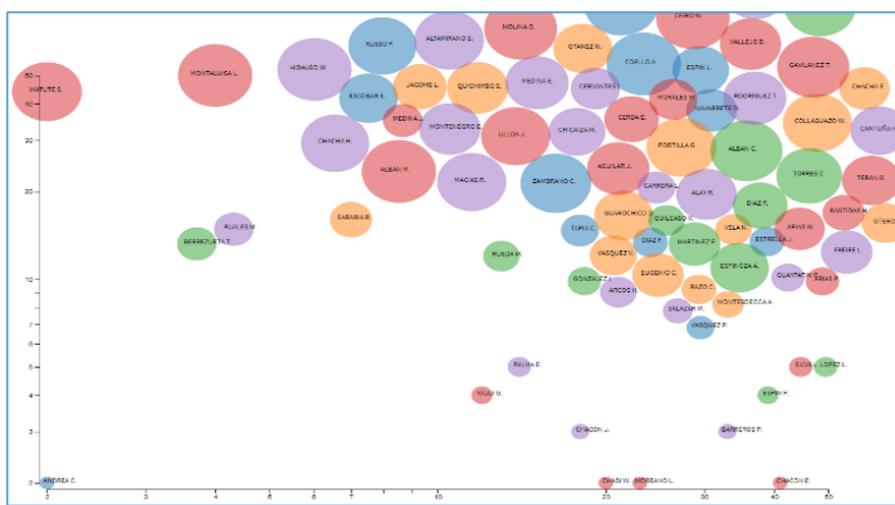
**FIGURA 3. PRODUCCIÓN CIENTÍFICA DE LAS PERSONAS INVESTIGADORAS DE ACUERDO CON LA LÍNEA DE INVESTIGACIÓN DE LA UNIVERSIDAD: CULTURA, PATRIMONIO Y SABERES ANCESTRALES.**



18

En el mismo caso de la sub-línea de investigación universitaria “Cultura, patrimonio y conocimientos ancestrales” pero en esta oportunidad representando similitud. En la Figura 4 como resultado de la aplicación del algoritmo k-means podemos observar que la producción científica de los investigadores Chacón E., Espín, Silva y López han sido baja sin embargo han trabajado en investigaciones relacionadas, por lo que conforman un pequeño clúster entre ellos; asimismo, podemos observar que los investigadores Moreano, Chasi, Chacón A., Medina, Yauli y Palma conforman otro clúster con distancia separada y baja producción científica, existiendo cierta similitud en sus áreas de investigación; sin embargo, la mayor relación se encuentra en la nube superior derecha, con una zona de mayor concentración de investigadores con características similares.

**FIGURA 4. RELACIÓN DE LOS INVESTIGADORES SEGÚN LA LÍNEA DE INVESTIGACIÓN DE LA UNIVERSIDAD: CULTURA, PATRIMONIO Y CONOCIMIENTOS ANCESTRALES.**



### 4.3 Resumen de ambos casos de estudio

El proceso de conciliación implica la extracción de perfiles terminológicos para la relación entre los actores del CEETAM. Por lo tanto, automatizar completamente este proceso era una tarea compleja debido al elevado número de interacciones necesarias. Sin embargo, es de destacar que estas acciones hacen uso de las TIC, principalmente describiendo la *World Wide Web*, para ver desde las distancias y los niveles de similitud entre los actores la compatibilidad, respondiendo a las nuevas tendencias de los entornos virtuales en este ámbito, saber que facilita las redes informales de los actores de la organización, como se refiere Marteleto y Braz (2004) en sus trabajos relacionados con las redes sociales, la inteligencia colectiva y el capital social.

Es importante indicar que el método de aglomeración debe ser utilizado con base en el tipo de datos y el objetivo de caracterización, previo a la visualización de la información es necesario establecer un conjunto de elementos terminológicos y conceptuales de los investigadores, lo cual depende de la oportunidad de proporcionar información al momento de crear su perfil empatando explícitamente con Samper (2005). La herramienta desarrollada para ambas instituciones de investigación ayuda al investigador a conocer los grupos a los que pertenece según la experiencia en estudios similares.

El estudio realizado por Avinash et al. (2020) relacionado con el tema *"Mapping Scientific Collaboration: A Bibliometric Study of Rice Crop Research in India"*, uno de sus objetivos era construir el perfil de la colaboración nacional e internacional de los científicos indios especializados en arroz, el autor del estudio aplicó tanto el análisis de coautores como el de redes sociales para captar a grandes rasgos los entresijos de la colaboración que opera en el campo de la investigación sobre el cultivo del arroz en la India. Este estudio utiliza técnicas bibliométricas para medir las colaboraciones científicas. En comparación con la investigación llevada a cabo en el CEETAM y la UTC, se aplican conceptos más relacionados con dos algoritmos de k-media y MDS, así como otras técnicas matemáticas implementadas a través de un sistema que permite la identificación de grupos de investigadores y el nivel de compatibilidad según su producción científica.

La investigación ha conseguido demostrar que métodos como los algoritmos de clúster, técnicas como el MDS y los espacios vectoriales pueden ser aplicados en contextos científicamente productivos y representados por investigadores de instituciones universitarias. En el caso de la presente investigación, el CEETAM de Cuba y la UTC de Ecuador han establecido las similitudes y distancias entre los investigadores, y el dendrograma ubica a los investigadores de la muestra según estos criterios en un escenario identificable; por otro lado, los investigadores de la UTC muestran un patrón de desempeño productivo marcado por sus publicaciones; no se han tomado como fuentes bases de datos relevantes como Scopus, Wos y otras, sino que estas han sido suministradas y validadas por los propios investigadores. Autores como Haris et al. (2019) han realizado estudios donde evalúan el desempeño investigativo de los profesionales de Bosnia y Herzegovina, considerando las citas y el índice h y el análisis de correlación estadística, dando resultados interesantes, pero comparados con el uso de algoritmos estos pueden ser más eficientes.

K-means es un algoritmo de clasificación no supervisado que agrupa los objetos en k grupos en función de sus características. La agrupación se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster, vemos en ambos casos de estudios (CEETAM y UTC) como se conforman grupos de investigadores; la visualización de información permite mostrar gráficamente las relaciones, esto puede distinguir la correcta toma de decisiones en cuanto a la construcción de redes de investigadores de acuerdo a temáticas, campos, líneas de investigación y con ello poder enfrentar nuevas investigaciones y comunidades académicas, donde podrán ser considerados de acuerdo a lo publicado y lo generado en sus perfiles.

## 5. CONCLUSIÓN

20

El análisis de datos aplicando las técnicas K-Media y MDS y el Clustering Aglomerativo Jerárquico constituyen herramientas notables que permiten la visualización de la información considerando la similitud y las distancias. Esto permite establecer niveles de compatibilidad entre actores en un contexto de investigación, alineándose a las nuevas tendencias mundiales de preceptos de machine learning para conocer redes informales de investigadores.

Parte de los resultados en el CEETAM es un dendrograma contrastado con un MDS que permite obtener un clúster más representado por los investigadores *dgonzalezr, eromero* y *yaguilera* y estos vinculados a *grbarcenas*; de igual manera se identifica el vínculo de alegría con el clúster formados por *egongora, yretirado, rmontero, lrpuron e iromero*, se observa jerárquicamente el vínculo entre todos estos investigadores con diferentes niveles de compatibilidad.

El MDS obtuvo dos grupos, uno formado por *dgonzalezr, eromero, yaguilera* y *grbarcenas*, que representan una red de investigadores vinculados a las áreas de las TIC y su aplicación; el resto representan una red de investigadores vinculados a la EEURE.

La implementación de estos algoritmos k-media en una plataforma Web en la UTC podría representar la producción científica de artículos científicos, libros y ponencias en congresos en las sublíneas de investigación en las especialidades de Medio Ambiente y Turismo, donde destacan investigadores como *Coello, Arrellano, Montaluisa* y *Cedeño* que son líderes en estas especialidades.

En la línea de investigación, de la Universidad "Cultura, patrimonio y saberes ancestrales" el algoritmo k-media, mostró que los investigadores *Chacón, Espín, Silva y López* estos han tratado investigaciones que están relacionadas entre sí, y forman un pequeño clúster entre ellos; también se obtuvieron los investigadores *Moreano, Chasi, Chacón, Medina, Yauli y Palma*. Aunque están algo distantes, existe una similitud visual y cualitativa en las áreas investigadas; sin embargo, la mayor relación se encuentra en la nube superior derecha, con una zona de mayor concentración de investigadores con características similares.

## 6. REFERENCIAS

Ahn, J. (2011). The effect of social network sites on adolescents' social and academic development: Current theories and controversies. *Journal of the American Society for Information Science and Technology*, 62(8), 1435-1445. <https://doi.org/10.1002/asi.21540>

Al-Anzi, F. S. y AbuZeina, D. (2018). Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach. *Information Processing & Management*, 54(1), 105-115. <https://doi.org/10.1016/j.ipm.2017.10.003>

Almufti, S., Marqas, R. y Ashqi, V. (2019). Taxonomy of bio-inspired optimization algorithms. *Journal Of Advanced Computer Science & Technology*, 8(2), 23. doi: <http://dx.doi.org/10.14419/jacst.v8i2.29402>

Ashby, F. G. (2014). *Multidimensional models of perception and cognition*. Psychology Press.

Avinash, K., Sambit, M. y Pradip, S. (2020). Mapping Scientific Collaboration: A Bibliometric Study of Rice Crop Research in India. *Journal of Scientometric Research*, 9(1). <http://dx.doi.org/10.5530/jscires.9.1.4>

Bárcenas, G. R., Culqui, A. C., Peñaherrera, J. R., Beltrán, S. C. y Tamayo, E. T. (2016). Levels of Similarity in User Profiles Based Cluster Techniques and Multidimensional Scaling. *International Journal of Systems Applications, Engineering & Development*, 10(2016), 56-64.

Cambria, E., Song, Y., Wang, H. y Howard, N. (2014). Semantic multidimensional scaling for open-domain sentiment analysis. *Intelligent Systems (IEEE)*, 29(2), 44-51.

Day, R. E. (2011). Death of the user: Reconceptualizing subjects, objects, and their relations. *Journal of the American Society for Information Science and Technology*, 62(1), 78-88. <https://doi.org/10.1002/asi.21422>

Degennmis, M., Lops, P., Ferilli, S., Di Mauro, N., Basile, T. M. A. y Semeraro, G. (2006). Text learning for user profiling in e-commerce. *International Journal of Systems Science*, 37(13), 905-918. <https://doi.org/10.1080/00207720600891794>

Dunn-Rankin, P., Knezek, G. A., Wallace, S. R. y Zhang, S. (2014). *Scaling methods*. Psychology Press.

Haris, M., Arnela, P., Edin, M. y Mahira, M. (2019). In Search of a Silver Bullet: Evaluating Researchers' Performance in Bosnia and Herzegovina. *Journal of Scientometric Research*, 8(3). <http://dx.doi.org/10.5530/jscires.8.3.27>

Joshi, R., Prasad, R., Mewada, P. y Saurabh, P. (2020). Modified LDA Approach For Cluster Based Gene Classification Using K-Mean Method. *Procedia Computer Science*, 171, 2493-2500. <https://doi.org/10.1016/j.procs.2020.04.270>

Kastrati, Z. y Imran, A. S. (2019). Performance analysis of machine learning classifiers on improved concept vector space models. *Future Generation Computer Systems*, 96, 552-562. <https://doi.org/10.1016/j.future.2019.02.006>

López-Herrera, A. G. (2006). *Modelos de sistemas de recuperación de información documental basados en información lingüística difusa* [Tesis doctoral, Universidad de Granada].

Machado, J. T. y Lopes, A. M. (2020). Multidimensional scaling and visualization of patterns in prime numbers. *Communications in Nonlinear Science and Numerical Simulation*, 83, 105-128. <https://doi.org/10.1016/j.cnsns.2019.105128>

Marteleto, R. M. y de Oliveira e Silva, A. B. (2005). Redes e Capital Social: o enfoque da informação para o desenvolvimento local. *Ciência Da Informação*, 33(3). <https://doi.org/10.18225/ci.inf.v33i3.1032>

Ratheeshkumar, A., Rajkumar, M., Balakrishnan, S. y Kalaiarasan, R. (2018). An Effective Method for Mapping Web User Profile based on Domain Ontology. *International Journal of Engineering & Technology*, 7(4.19), 1-4.

Rodríguez-García, M., Valencia-García, R.I., Alcaraz-Mármol, G. y Carralero, C. (2014). Open Idea: An intelligent platform for managing innovative ideas. *Procesamiento de Lenguaje Natural*, 53, 147-150.

Salton, G. (1989). *Automatic Text Processing*. Addison Wesley.

Samper. (2005). *Estudio y evaluación de un sistema inteligente para recuperación y el filtrado de información de Internet*. [Tesis de Doctorado, Universidad de Granada].

Sun, J. (2012). Why different people prefer different systems for different tasks: An activity perspective on technology adoption in a dynamic user environment. *Journal of the American Society for Information Science and Technology*, 63(1), 48-63. <https://doi.org/10.1002/asi.21670>

Sun, S., Song, H., He, D. y Long, Y. (2019). An adaptive segmentation method combining MSRCR and mean shift algorithm with K-means correction of green apples in natural environment. *Information Processing in Agriculture*, 6(2), 200-215. <https://doi.org/10.1016/j.inpa.2018.08.011>

Velásquez, E., Cardona, A. y Peña, A. (2014). Vector Model for Cognitive State Inference of Patients in Coma-Derived States. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 13, 65-81. <https://doi.org/10.4304/risti.13.65-81>

Žilinskas, J. (2007). Reducing of search space of multidimensional scaling problems with data exposing symmetries. *Information Technology And Control*, 36(4), 377-382

# e-Ciencias de la Información



¿Dónde se encuentra indexada e-Ciencias de la Información?



DOAJ

Para más información ingrese a nuestra [lista completa de indexadores](#)

¿Desea publicar su trabajo?  
Ingrésese [aquí](#)

O escríbanos a la siguiente dirección  
[revista.ebc@ucr.ac.cr](mailto:revista.ebc@ucr.ac.cr)