UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA DE COMPUTACIÓN

Desarrollo de un Sistema Web Basado en Matrices de Distancias para Seleccionar Artículos Académicos Similares en la Revisión de la Literatura Científica

Trabajo de titulación previo a la obtención del
Título de Ingenieros en Ciencias de la Computación

AUTORES: KEVIN PATRICIO CADENA PEÑA
ALEXIS NICOLAS VILLAVICENCIO GARCIA
TUTOR: DIEGO FERNANDO VALLEJO HUANGA

Quito - Ecuador
2024

CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Kevin Patricio Cadena Peña con documento de identificación N.º 1727322552 y Alexis Nicolas Villavicencio Garcia con documento de identificación N.º 1723438543; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 29 de febrero del 2024

Atentamente,

Kevin Patricio Cadena Peña
1727322552

Alexis Nicolas Villavicencio Garcia
1723438543

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Nosotros, Kevin Patricio Cadena Peña con documento de identificación No. 1727322552 y Alexis Nicolas Villavicencio Garcia con documento de identificación No. 1723438543, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: "Desarrollo de un Sistema Web basado en Matrices de Distancias para Seleccionar Artículos Académicos Similares en la Revisión de la Literatura Científica.", el cual ha sido desarrollado para optar por el título de: Ingenieros en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 29 de febrero del 2024

Atentamente,

Kevin Patricio Cadena Peña
1727322552

Alexis Nicolas Villavicencio Garcia
1723438543

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Diego Fernando Vallejo Huanga con documento de identificación N° 1720162708, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: DESARROLLO DE UN SISTEMA WEB BASADO EN MATRICES DE DISTANCIAS PARA SELECCIONAR ARTÍCULOS ACADÉMICOS SIMILARES EN LA REVISIÓN DE LA LITERATURA CIENTÍFICA, realizado por Kevin Patricio Cadena Peña, con documento de identificación N.º 1727322552 y por Alexis Nicolas Villavicencio Garcia con documento de identificación N.º 1723438543, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 29 de febrero del 2024

Atentamente,

Ing. Diego Fernando Vallejo Huanga, MSc
1720162708

Desarrollo de un Sistema Web Basado en Matrices de Distancias para Seleccionar Artículos Académicos Similares en la Revisión de la Literatura Científica

1st Kevin Patricio Cadena Peña
kcadenap1@est.ups.edu.ec

2nd Alexis Nicolas Villavicencio Garcia
avillavicenciog1@est.ups.edu.ec

3rd Diego Vallejo-Huanga
dvallejoh@ups.edu.ec

***Resumen*—Debido al aumento constante de las publicaciones de artículos científicos en diversas bibliotecas virtuales a nivel mundial cada año, la búsqueda de documentos académicos relevantes para investigaciones se ha convertido en un proceso complejo que en ocasiones suele requerir una gran cantidad de tiempo y esfuerzo. Este artículo de investigación aborda esta problemática mediante el desarrollo de un sistema web que facilita a los usuarios la identificación de los artículos científicos más relevantes para sus investigaciones. El sistema permite los usuarios cargar un conjunto de datos que incluye los atributos títulos, palabras clave y resúmenes de cada documento extraído. Cada artículo científico fue sometido al proceso NLP con el fin de depurar y homogeneizar la información. Para medir la divergencia entre los documentos científicos, se calculan matrices de similitud mediante el uso de dos métricas diferentes que ensamblan los tres atributos en una sola matriz ponderada. Los resultados obtenidos se presentan a través de una interfaz web, que incluye una tabla de resumen, un mapa de calor, un diagrama de dispersión bidimensional y un grafo. La experimentación se llevó a cabo utilizando un conjunto de datos compuesto por 192 artículos científicos recopilados de Springer, IEEE y Scopus, abarcando las áreas de Ciencias Exactas, Ciencias de la Computación, Medicina y Ciencias Sociales. Los resultados revelaron una mayor similitud entre documentos que comparten la misma temática de estudio.**

***Palabras Clave*—Coseno Vectorial, Artículos Científicos, Matriz de Similitudes, Jaccard, Inteligencia Artificial, Sistema Web, Revisión de la Literatura Científica.**

***Abstract*—Due to the constant increase in publications of scientific articles in various virtual libraries worldwide every year, searching for relevant academic documents for research has become a complex process that sometimes requires significant time and effort. This research article addresses this problem by developing a web system that makes it easier for users to identify the most relevant scientific articles for their research. The system allows users to upload a set of data that includes the attributes, titles, keywords, and abstracts of each extracted document. Each scientific article was subjected to the NLP process to purify and homogenize the information. To measure the divergence between scientific documents, similarity matrices are calculated using two different metrics that assemble the three attributes into a single weighted matrix. The results are presented through a web interface, including a summary table, a heat map, a two-dimensional scatter diagram, and a graph. The experimentation was carried out using a data set composed of 192 scientific articles from Springer, IEEE, and Scopus, covering the areas of Exact Sciences, Computer Sciences, Medicine, and Social Sciences. The results revealed greater similarity between documents that share the same study topic.**

***Keywords*—Vector Cosine, Scientific Papers, Similarity Matrix, Jaccard, Artificial Intelligence, Web System, Review of Scientific Literature.**

I. Introducción

Una de las formas más importantes y ancestrales de transmisión del conocimiento, es la forma verbal [1]. Sin embargo, en la actualidad, la escritura ha tomado un valor relevante en los procesos de aprendizaje [2] y es una de las características más necesitadas en el contexto científico. Entonces, la comunicación escrita se ha convertido en una habilidad crucial en el ámbito académico, que permite la organización de ideas, transmisión de información de manera coherente y documentación de investigaciones, con el objetivo de que sean comprensibles para la ciudadanía [3].

Los artículos científicos, según Martinson y Anders, tienen como objetivo comunicar de manera efectiva los descubrimientos de una investigación fomentando debates en los resultados obtenidos [4]. La investigación científica es un proceso que debe seguir pautas de claridad, concisión y veracidad, asegurando la comprensión de los lectores, permitiendo promover el avance continuo de la ciencia [5]. En la última década se ha evidenciado un aumento en la producción de artículos académicos, alcanzando aproximadamente 5.14 millones de nuevos documentos al año, con un incremento del 28.7% en revistas académicas [6]. Durante los años 2015 y 2019 la colaboración científica internacional ha hecho que aumente del 22% al 24% la contribución de publicaciones en revistas académicas [7]. Esta tendencia ha permitido involucrar a cerca de 13 millones de personas anualmente, entre académicos, profesores y estudiantes de educación superior, en procesos de publicación de diversos campos de investigación [8].

En el año 2020, según estudios realizados en América Latina, se demostró que Brasil ha generado un total de 387464 publicaciones académicas en *Web of Science*, 191285 en *Current Contests Connect*, 161923 en *MEDLINE*, 90537 en *SciELO* y 664 en *Korean Jounal Database*, posicionando así a Brasil como el país latinoamericano con mayor contribución

científica [9]. Por otro lado, Ecuador se ubica en la séptima posición, a nivel latinoamericano, con 20816 publicaciones académicas en la plataforma *Web of Science* entre los años 2015 y 2020 [9], con un aumento de la producción científica del 0.49% al 2.27% [10].

El concepto Revisión de Literatura Sistemática (SLR) es una técnica utilizada para medir y categorizar la información en función de su relevancia, dentro las publicaciones científicas [11]. Este proceso es desafiante, laborioso y demanda grandes cantidades de recursos temporales, ergo, en la actualidad los investigadores hacen uso de diversas tecnologías como el aprendizaje de máquina para su automatización [12]. El proceso SLR, tuvo su origen en áreas de ciencias sociales, pero actualmente es utilizado en diferentes campos de investigación como la salud, informática, gestión e ingenierías [11]. Los investigadores recurren al proceso SLR debido a su metodología explícita, planificada, responsable y justificable que garantiza una revisión imparcial, precisa, auditable, reproducible y actualizable, que son elementos clave en la investigación académica rigurosa [13].

Entonces, se han generado varios desafíos para los investigadores que realizan el proceso de revisión de literatura científica, de forma manual, ya que este es un procedimiento altamente complejo y que deriva en un gran consumo de tiempo [14]. Diferentes aplicaciones de Inteligencia Artificial (IA), en conjunto con el Procesamiento Natural del Lenguaje (NLP, por sus siglas en inglés), han contribuido a la automatización de procesos ya que estos sistemas pueden aprender de la experiencia, reconocer patrones, tomar decisiones autónomas y adaptarse a nuevas situaciones [15]. Estos procesos de automatización, a través de IA y NLP, también se han utilizado para la revisión de literatura científica. Entonces, estas tecnologías permiten analizar grandes volúmenes de datos en lenguaje natural [16], empleando métodos de análisis de similitud del contenido de los documentos. No obstante, para lograr un alto nivel de precisión, es importante considerar diversas características de los documentos académicos, como el nombre del autor, el identificador de objeto digital (DOI), las palabras clave, el resumen y el contenido, tal como lo indican Akhil y George [17].

Este artículo científico propone la creación de un sistema web, que permita optimizar los procesos de búsqueda de documentos académicos, con el objetivo de mejorar la experiencia de la comunidad científica en el proceso de revisión de la literatura. El sistema utilizará técnicas de IA, NLP y permitirá a los usuarios cargar archivos de texto plano, desde los cuales se extraerán de forma automática títulos, palabras clave y resúmenes de los documentos académicos. Luego, la información se presentará de manera resumida en tablas y se mostrarán los los documentos con mayores similitudes utilizando algoritmos de aprendizaje no supervisado. Los resultados se visualizarán en gráficos para tener una representación intuitiva y facilitar la identificación de patrones y relaciones entre los documentos.

II. Trabajos Relacionados

Existen varios trabajos que abordan el proceso de clasificación y agrupamiento de documentos basados en métricas de similitud. El uso de técnicas de IA y NLP ha mejorado el rendimiento y precisión en los procesos de búsqueda de documentos científicos similares.

La selección de métricas de distancia para evaluar la similitud entre documentos constituye un desafío debido a la amplia variedad de técnicas y algoritmos disponibles. Según Cai et al. [18], la elección de una métrica adecuada al dominio específico del campo es determinante, ya que impacta de manera significativa en el rendimiento y la precisión del análisis. La evaluación de medidas de similitud se presenta como un desafío continuo, dado el constante desarrollo de diversas técnicas y medidas a lo largo del tiempo, todas orientadas a medir la similitud con un alto grado de exactitud [19]. En una exploración de la similitud entre artículos académicos de diversas disciplinas, Almas et al. [20] desarrollaron un sistema que utiliza NLP y etiquetado *Part-Of-Speech* (POS). Al utilizar únicamente títulos y resúmenes como datos de entrada, este sistema proporciona al usuario los cinco artículos más similares. Su método fundamentado en la similitud del coseno entre vectores, desde un sistema de pesado TF-IDF, alcanzó una precisión del 85.54%.

En el estudio de Moya et al. [21], se propuso la creación de un prototipo de sistema de recomendación orientado a grupos de investigación en instituciones de educación superior. Este enfoque consideró aspectos como perfiles de usuarios, áreas del conocimiento y medidas de similitud. La construcción del sistema se llevó a cabo utilizando el lenguaje de programación *Python* y el entrenamiento de una red neuronal con *Tensorflow*. El resultado final fue un sistema de recomendación capaz de considerar tanto la afinidad de grupos como las preferencias personales de los usuarios.

Dada la ingente cantidad de artículos científicos en la red, es menester identificar aquellos algoritmos y métricas de similitud que permiten tener una búsqueda y recomendación optimizada de documentos académicos. Así, Magara et al. [22] exploraron algoritmos y métricas de similitud, con técnicas de clasificación no lineal y algoritmos como *Particionamiento Recursivo* (rpart), *Random Forest*, y *Boosted Machine Learning*, para comparar su rendimiento. El *performance* de los algoritmos fue probado en un conjunto de datos proveniente de la Universidad de Ghent, utilizando validación cruzada. Los resultados mostraron una precisión promedio del 80.73%, con la aproximación de *rpart* y un tiempo de ejecución de 2.3546 segundos.

Otro enfoque para procesos de recomendación está basado en la autoría y co-autoría de los documentos científicos. Chen et al. [23] presentaron un algoritmo de enlace semántico que incorpora una red de información heterogénea ponderada. Este artículo usa la construcción de una Red de Información Heterogénea (HIN, por sus siglas en inglés) con distintos vértices como artículos y autores, y varios tipos de relaciones de enlace semántico como citación, escritura y coautoría. Entonces, la

recomendación de una artículo similar se realiza mediante el aprendizaje de representación de red y la combinación lineal de similitudes multimodales. La evaluación, basada en el conjunto de datos de *ACL Anthology Network*, muestra un aumento en el *recall* de 6.9% a 8.4% , en comparación con los algoritmos PV-DBOW, PW, PWFC, MMRQ y BM25.

Explorando las redes de multinivel de citas y relaciones de autores, Waleed et al. [24] desarrollaron una estrategia distintiva. Su metodología prioriza la red de citas para identificar artículos relevantes y autores clave, donde se incorpora la evaluación de la importancia de cada artículo mediante medidas de centralidad. La experimentación para validar su propuesta fue evaluada en un conjunto de datos de *AMiner* con 2092356 trabajos de investigación, 8024869 citas y 1712433 autores. Este nuevo enfoque fue comparado con otros sistemas de búsqueda como *Google Scholar* y *MSCN*, destacando su capacidad para recomendar artículos de alta calidad, sin depender del número de citas o la fecha de publicación del artículo de interés.

Nuestro artículo aborda el proceso de recomendación de documentos científicos similares mediante el desarrollo de un sistema web que utiliza técnicas de IA, NLP y matrices de similitudes. El sistema web tiene la capacidad de proporcionar al usuario información de manera visual, para un mejor análisis y comparación de datos. Por medio de diferentes tecnologías como el procesamiento de datos, representación gráfica y técnicas de NLP se optimiza el proceso de SLR para la selección de artículos académicos relevantes.

III. Materiales y Métodos

La metodología empleada en esta investigación es de carácter descriptiva y cuantitativa de tipo experimental. Inicia con la recopilación de datos textuales provenientes de casas editoriales como ScienceDirect, IEEE Xplore, MDPI, Springer, Taylor & Francis, Cengage Learning y Wiley. La recuperación de la información busca obtener artículos científicos que contengan tres metadatos para ser explotados, específicamente, títulos, palabras clave y resúmenes de los artículos académicos.

Como resultado de este proceso de recuperación de la información se obtuvo un fichero en formato CSV que posteriormente interactuará con el sistema web. Dentro del prototipo se calculan los niveles de similitud entre los *papers* y se generan visualizaciones que serán mostradas al usuario en la web. Este proceso de obtención de datos se limita a la búsqueda de artículos en cuatro áreas: Ciencias Sociales, Ciencias de la Computación, Medicina y Ciencias Exactas. El resumen del proceso metodológico se muestra en la Figura 1.

La arquitectura del sistema web tiene dos componentes; el *frontend* que se desarrolló con el *framework* Angular, incorporando bibliotecas de código abierto como PrimeNG y PrimeFlex para mejorar la adaptabilidad del sistema a distintos dispositivos. Además, se integró amCharts5, una herramienta que facilita la creación de gráficos interactivos. El segundo componente es el *backend* implementado en Python, con el uso de la biblioteca NLTK para llevar a cabo el proceso de

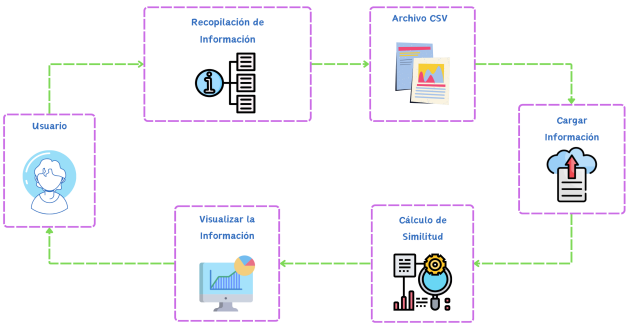

Fig. 1: Diagrama general del funcionamiento del sistema web para recuperar artículos científicos similares

NLP. La comunicación entre el *frontend* y el *backend* se logró mediante APIs, para consolidar la integración y funcionalidad del sistema web.

A. Recuperación de la Información y Generación del Dataset

El conjunto de datos recopilado consta de 48 artículos científicos por cada una de las cuatro áreas consideradas, dando un total de $i = \{1, ..., 192\}$ artículos científicos recopilados. Sobre este *dataset* se realizó los experimentos para medir la eficiencia de los algoritmos de NLP y el funcionamiento general del sistema.

Para que un artículo se indexe en nuestro conjunto de datos se utilizó un criterio de inclusión basado en un mínimo de 15 referencias, i.e, que el artículo científico debe tener esta cantidad de referencias en su bibliografía. Luego, con el objetivo de no procesar todo el corpus del artículo científico se extrajeron, únicamente, tres elementos de cada *paper*: el Título T_i, las Palabras Clave PC_i y el Resumen R_i. Estos elementos se organizaron en un archivo CSV, donde cada fila representa un artículo y las columnas son los atributos título, palabras clave y resumen, respectivamente. Cada elemento de la colección de documentos científicos, desde el campo del NLP, es considerado como una instancia documental Doc_i.

B. Metodología de Desarrollo del Software

El desarrollo del software se llevó a cabo en un equipo personal con las siguientes características: un procesador Intel Core i7 de quinta generación, 16 GB de memoria RAM, 1 TB de almacenamiento y 4 núcleos. Con el objetivo de gestionar eficientemente las tareas del proyecto, se optó por dividir el desarrollo en tres etapas. Esta decisión se basó en la optimización de la asignación de recursos en cada fase del proyecto, con la finalidad de mejorar la eficacia del proceso de desarrollo y ejecución.

En la primera etapa, la recopilación de información demandó aproximadamente 30 horas de búsqueda y selección de artículos científicos que cumpla con los requisitos para su indexación en el *dataset*. El archivo resultante se almacenó localmente en la computadora de experimentación y se respaldo en un repositorio en *Github*.

Para las siguientes dos etapas del proyecto, dado que se implementó una aplicación web, se utilizó un patrón de arquitectura cliente-servidor, compuesto por dos elementos: el *frontend* y el *backend* comunicados mediante APIs que se encargan de todas las interacciones y el procesamiento de la información, como se observa en la Figura 2.

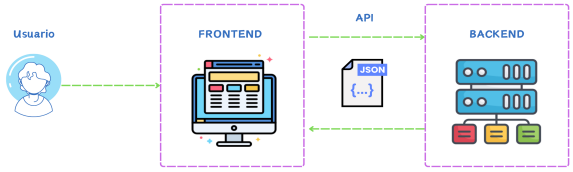

Fig. 2: Arquitectura de alto nivel del software desarrollado

El *backend*, que representa a la segunda etapa de desarrollo del software, se implementó mediante el lenguaje de programación Python versión 3.11.4 con el *framework* Flask para la creación de la API que sirve de comunicación. La API se encarga de recibir la información proveniente del *dataset*, en formato CSV. La información obtenida, posteriormente, se somete a un proceso de NLP utilizando la biblioteca NLTK de Python. La etapa de pre-procesamiento de NLP en el texto incluye tokenización, conversión a letras minúsculas, eliminación de caracteres especiales, supresión de *stopwords* y *stemming* para reducir las palabras a su base o raíz. El pre-procesamiento de los corpus de los Doc_i permite, a-posteriori, calcular bajo un mismo formato las similitudes entre cada instancia documental y el resto del *dataset* para generar una matriz de similitudes.

El cálculo de la matriz de similitudes entre documentos considera métricas de similitud, que proporcionan medidas cuantitativas para comparar la cercanía semántica entre documentos. Dos enfoques comúnmente utilizados son el índice de Jaccard y la similitud de coseno vectorial. Estas métricas se utilizarán en la experimentación de este proyecto de investigación para evaluar la divergencia entre los artículos académicos recopilados. Su aplicación permitirá la identificación de trabajos relacionados durante el proceso de SLR. El coeficiente de Jaccard evalúa la similitud entre dos conjuntos de datos, calculando la relación entre el tamaño de la intersección de los conjuntos y el tamaño de la unión de los dos conjuntos. Los elementos de los conjuntos, para este trabajo, son los *tokens* de cada Doc_i. El rango para este coeficiente es $J : X \times Y \in [0, 1]$, donde 1 representa que ambos conjuntos son iguales y 0 que los dos documentos son completamente diferentes.

Para evaluar la similitud entre los Títulos T_i y las Palabras Clave PC_i, se utilizó la medida de similitud de *Jaccard*. Esta elección se justifica debido a que tanto los títulos como las palabras clave suelen tener un número pequeño de términos o *tokens* sin repetir, siendo esta medida de similitud adecuada para reducir el coste computacional del sistema. La Ecuación 1 muestra la forma de calcular el coeficiente de Jaccard para T_i y PC_i.

$$J(T_i, T_j) = \frac{\text{card}(T_i \cap T_j)}{\text{card}(T_i \cup T_j)}$$
$$J(PC_i, PC_j) = \frac{\text{card}(PC_i \cap PC_j)}{\text{card}(PC_i \cup PC_j)} \quad (1)$$

La segunda métrica de similitud utilizada, coseno vectorial, evalúa el grado de semejanza entre dos vectores al examinar el ángulo formado entre ellos, en lugar de su magnitud. Su rango de definición es cos: $X \times Y \in [0, 1]$. Para aplicar esta métrica, es necesario vectorizar cada uno de los documentos mediante la creación de una bolsa de palabras y ponderar su importancia mediante la técnica *Term Frequency-Inverse Document Frequency* (TF-IDF). Esta técnica proporciona un puntaje de significancia a cada *token* para determinar su relevancia en Doc_i. Una vez finalizado este proceso se calcula la similitud entre los vectores que representan a los Doc_i.

El atributo Resumen R_i del *dataset* de documentos, comúnmente, contienen una gran cantidad de términos o *tokens*, donde la probabilidad de encontrar términos repetidos es más alta, ergo, para medir la similitud entre un par de R_i se empleó la medida de similitud del coseno Vectorial. Esta métrica resulta útil para comparar documentos extensos y capturar la similitud en función de la distribución de términos en un espacio vectorial y está definida por la Ecuación 2.

$$\cos(\vec{R_i}, \vec{R_j}) = \frac{\vec{R_i} \cdot \vec{R_j}}{\|\vec{R_i}\| \cdot \|\vec{R_j}\|} \quad (2)$$

Entre cada instancia documental se calcula las divergencias, bajo la metodología descrita, y se genera una matriz de similitudes para cada uno de los tres atributos MS_{T_i}, MS_{PC_i} y MS_{R_i}. Estas matrices de similitudes individuales se combinan en una única matriz consolidada de similitudes, denotada por MDT.

Algoritmo 1 Matriz de Similitud Total

Entrada: T_i, PC_i, R_i
Salida: MDT
Paso 1: NLP
for $i \leftarrow 1, n$ **do**
  $T_i \leftarrow NLP(T_i)$;
  $PC_i \leftarrow NLP(PC_i)$;
  $R_i \leftarrow NLP(R_i)$;
end for
Paso 2: Matriz de Similitud
$MS_{T_i} \leftarrow Jaccard(T_i)$;
$MS_{PC_i} \leftarrow Jaccard(PC_i)$;
$MS_{R_i} \leftarrow Coseno(R_i)$;
Paso 3: Ponderación
$MS_{T_i} \leftarrow (MS_{T_i} * 0.2)$;
$MS_{PC_i} \leftarrow (MS_{PC_i} * 0.3)$;
$MS_{R_i} \leftarrow (MS_{R_i} * 0.5)$;
Paso 4: Matriz de Similitud Total
$MDT \leftarrow (MS_{T_i} + MS_{PC_i} + MS_{R_i})$;

La unión de estas matrices se realiza mediante una combinación ponderada, asignando porcentajes de relevancia a cada una. Para esta investigación, considerando la importancia

de cada atributo descrito en la literatura científica, las ponderaciones fueron de 20% para T_i, 30% en las PC_i y 50% para R_i. El proceso de creación de la matriz *MDT* sigue el pseudocódigo representado en el Algoritmo 1. La Figura 3 resume el proceso metodológico descrito para la creación de las matrices de similitudes.

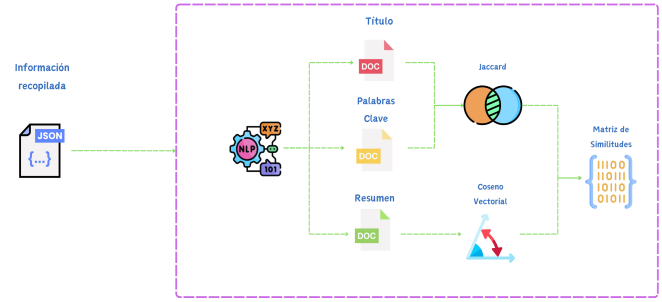

Fig. 3: Cálculo de la matriz de similitudes con procesos de NLP y métricas de divergencia

Este enfoque combinado de medidas permitirá analizar y cuantificar la similitud entre los artículos científicos, considerando los tres atributos seleccionados. Además, facilitará la exploración y el descubrimiento de patrones y relaciones dentro del *dataset*.

En la etapa tres del proyecto, se procedió al desarrollo del *frontend* de la aplicación, implementado mediante el *framework* Angular versión 14.2.7, que por su estructura, permite crear interfaces de usuario dinámicas. Las librerías de PrimeNG y PrimeFlex se incorporaron para estilizar la interfaz y agregarle responsividad. La representación visual de los resultados se logró mediante la integración de la librería *amCharts5* para la creación de gráficos dinámicos.

En la interfaz web se cargará el archivo CSV que será enviado, por medio de la API, hacia el *backend* donde se realizará el proceso de NLP y cálculo de la matriz de similitudes devolviendo los resultados al *frontend* para su posterior representación gráfica mediante un mapa de calor, un grafo y un diagrama de dispersión utilizando la librería amCharts5.

Para llevar a cabo la clasificación de los documentos, de cada de área de conocimiento, se utilizó un algoritmo de *clustering* aglomerativa que es una técnica de aprendizaje no supervisado que recibe como entrada la matriz de similitudes ponderada y devuelve como salida a cada articulo científico agrupado en un *cluster*. Las etiquetas de las instancias representan la pertenencia de cada documento a uno de los grupos que se identificaron, i.e., cuatro grupos, uno por cada área de conocimiento.

Con los grupos previamente obtenidos del *clustering* aglomerativo y utilizando la matriz MDT se extrajeron las coordenadas de cada elemento de la matriz, valor de la fila y valor de la columna, junto con el valor de la celda que representa la distancia entre cada documento. El proceso etiqueta a cada documento a un grupo, lo que permite visualizar las similitudes entre los documentos mediante colores donde los más intensos indican mayor similitud, mientras que tonalidades más tenues resaltan las diferencias. Esta combinación de agrupamiento y representación gráfica del mapa de calor facilita la identificación de patrones y relaciones dentro del conjunto de datos.

Adicionalmente, se generó un diagrama bidimensional (2D) mediante un Escalamiento Multidimensional (MDS) que sigue Algoritmo 2. Esta técnica permite la identificación de agrupamientos naturales, patrones de distribución, *outliers* y relaciones entre los documentos. Los puntos cercanos en el diagrama indican documentos con mayores similitudes, mientras que aquellos más distantes representan a los menos relacionados.

Algoritmo 2 Escalamiento Multidimensional de los Doc_i

Entrada: MDT
Salida: $distanceData$
Paso 1: Instanciar algoritmo
$hc \leftarrow AgglomerativeClustering()$;
Paso 2: Obtener etiquetado
$y_hc \leftarrow hc.predict(MDT)$;
Paso 3: Formato para el gráfico
$mds \leftarrow MDS()$;
Paso 4: Coordenadas
$coor \leftarrow mds.predict(MDT)$;
Paso 5: Formato para el gráfico
for $i \leftarrow y_hc.length$ **do**
    $distanceData \leftarrow formatData(coor(i), y_hc(i))$;
end for

Finalmente, se elaboró un grafo que sigue el pseudocódigo del Algoritmo 3. El resultado obtenido representa las conexiones entre los documentos agrupados, donde cada nodo del grafo representa un documento y las aristas indican las relaciones de similitud entre ellos.

Algoritmo 3 Diagrama de Grafo con Clustering

Entrada: MDT
Salida: $clusters$
Paso 1: Instanciar algoritmo
$hc \leftarrow AgglomerativeClustering()$;
Paso 2: Obtener etiquetado
$y_hc \leftarrow hc.predict(MDT)$;
Paso 3: Formato para el gráfico
$y_unique \leftarrow get_unique(y_hc)$;
for $label \leftarrow cluster, y_unique$ **do**
    $Cluster_{label} \leftarrow formatData(cluster, y_hc)$;
end for
$Clusters \leftarrow Cluster_{label}$;

C. Descripción Funcional de la Herramienta Web

La aplicación web está compuesta por cinco páginas. En la página principal se encuentra información detallada del proyecto. La página denominada *papers* es el lugar donde los usuarios cargarán la información de sus datos para obtener los resultados del sistema web. Aquí se puede subir el conjunto de datos recopilado en formato CSV y se debe seleccionar el separador correspondiente al *dataset*. Una vez que el archivo se ha cargado y se ha elegido el parámetro de separador, se debe hacer clic en el botón *Subir*.

Es necesario tener en cuenta que para que el sistema web acepte el archivo, el *dataset* recopilado debe contener los tres atributos T_i, PC_i y R_i. Caso contrario, la aplicación notificará al usuario sobre las inconsistencias en el archivo. Una vez que la información se ha cargado con éxito en el sistema web, se presentará de manera resumida en una tabla toda la información del conjunto de datos.

Además, se ha implementado un controlador que permite al usuario seleccionar los atributos a visualizar, e.g., si el usuario hace clic en *Títulos*, la tabla presentará exclusivamente los títulos correspondientes; de manera similar, al seleccionar *Resúmenes* o *Palabras Clave*, la tabla se ajustará para mostrar, únicamente, la información seleccionada por el usuario, tal como se muestra en la Figura 4.

Fig. 4: Previsualización de la información cargada por el usuario en el sistema web

Al seleccionar la opción *Subir* toda la información del archivo se envía al *backend* en formato JSON mediante la API, para que ejecute el proceso de NLP y cálculo de matrices de similitudes. Cuando este proceso termina se devuelve todas las matrices de similitudes calculadas al *frontend*, por cada atributo del artículo científico, así como la matriz ponderada total. Los resultados del cálculo de similitud entre artículos científicos se presentan de forma gráfica en las siguientes tres páginas.

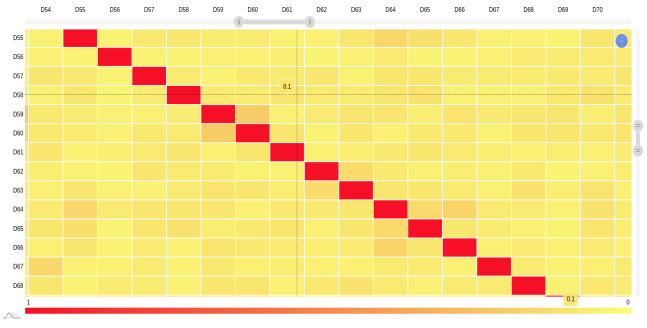

Fig. 5: Mapa de calor para representar similitud entre los documentos cargados en la interfaz web

En la sección *Heat Map* se muestra un mapa de calor, como se ilustra en la Figura 5, donde se representan visualmente los valores de similitud entre los artículos científicos. Este mapa de calor cuenta con dos controladores que permiten realizar un acercamiento a la imagen, facilitando una mejor visualización de las similitudes entre los *papers*. Además, al seleccionar uno de los cuadros dentro del diagrama, se presenta el valor de similitud entre dos artículos, contenido en un rango entre 0 y 1, donde los valores más cercanos a 1 representan mayor similitud.

En la sección *MDS*, se presenta un diagrama de dispersión bidimensional que representa la categorización de documentos según su similitud, como se observa en la Figura 6. Este diagrama, también permite el acercamiento mediante dos controladores y al seleccionar un punto dentro del diagrama, se muestra la etiqueta del artículo correspondiente.

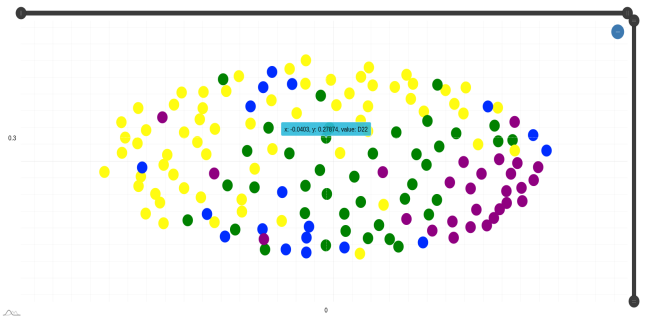

Fig. 6: Diagrama MDS para representar las relaciones de similitud entre los documentos

En la sección *Graph*, los documentos son mostrados en forma de un grafo, como se ilustra en la Figura 7. Las relaciones entre los *papers*, cargados por el usuario, se agrupan de acuerdo al resultado del algoritmo de *clustering* que segmentó en cuatro grupos a los artículos científicos.

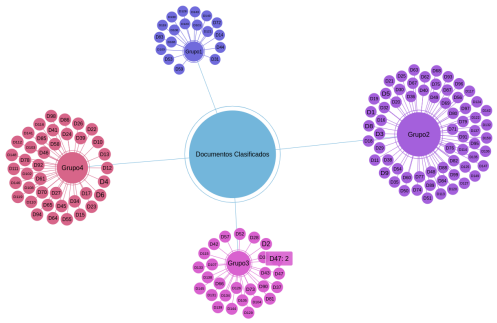

Fig. 7: Grafo de agrupamiento de los Doc_i aplicando el algoritmo de *clustering*

El sistema web desarrollado como parte de esta investigación está disponible en el siguiente enlace https://scientificrelevancedetector.onrender.com. Adicionalmente, el código fuente completo se encuentra alojado en el repositorio público Github: https://github.com/nicasop/ScientificRelevanceDetector.git. Los usuarios interesados pueden acceder al código fuente

para revisar, contribuir o utilizar el sistema de acuerdo con las licencias y políticas establecidas en el repositorio.

IV. Experimentos y Resultados

El proceso experimental, se dividió en tres etapas, las cuales utilizan el *dataset* disponible en el repositorio: https://raw.githubusercontent.com/sebas979/archicosCSV/main/DataSet.csv. Este conjunto de datos contiene artículos científicos recopilados manualmente, cumpliendo con todas las reglas, clarificadas en la metodología, para su indexación como un *dataset* válido.

En la primera etapa de la experimentación, se generó un mapa de calor y un diagrama de caja y bigotes para cada subconjunto de datos correspondiente a las cuatro áreas de conocimiento. El objetivo es evaluar la precisión del algoritmo implementado en el cálculo de la matriz de similitudes en el *backend*, para identificar los artículos científicos similares recopilados en el *dataset*. Para lograr esto, se dividió el conjunto de datos, asignando cada Doc_i a su respectiva área de estudio, resultando en la creación de cuatro subconjuntos de datos, cada uno compuesto por 48 documentos.

Durante el análisis, se observó que el algoritmo para el cálculo de las matrices de similitudes obtenía valores cercanos entre documentos, indicando una relación fuerte entre varios artículos científicos. Para observar los resultados obtenidos, en cada una de las áreas analizadas, se tomaron en cuenta dos valores estadísticos: la media de similitud y su desviación estándar, como se muestra en la Tabla I. En cada área se observó que los valores se encuentran en un rango normalizado de 0 a 1, donde 1 representa que los documentos son idénticos y 0 que son distintos.

Grupo	**Área de Estudio**	**Media** ($\mu \pm \sigma$)
1	Ciencias Exactas	0.11 $\pm$0.14
2	Ciencias de la Computación	0.15 $\pm$0.14
3	Medicina	0.1 $\pm$0.14
4	Ciencias Sociales	0.1 $\pm$0.15

Tabla I: Media y desviación estándar de cada área de estudio

En la Figura 8, que muestra los mapas de calor generados por cada área de conocimiento, se observa que en Medicina y Ciencias Sociales existe una baja similitud dentro de su respectivo conjunto de datos, esto concuerda con los valores de media y desviación estándar obtenidos. Por otro lado, en Ciencias de la Computación y Ciencias Exactas, se observó que existe mayor similitud entre los Doc_i de sus respectivas áreas.

Además, en la representación gráfica que se muestra en la Figura 9, se observa que la media de las distancias presenta asimetría en las similitudes entre los documentos, indicando que la tendencia central de los datos no se encuentra en un punto medio. Se observa la presencia de valores atípicos en el diagrama de caja y bigotes, señalando artículos científicos cuyas distancias difieren de la mayoría. Esta variación puede atribuirse a distintas características en los *tokens* de los documentos recopilados.

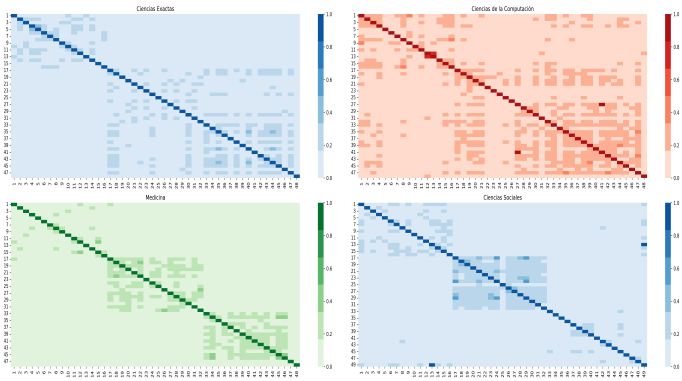

Fig. 8: Mapas de calor distribuidas por cada área de estudio

La presencia de valores atípicos sugiere que algunos documentos tienen similitudes bajas en comparación con el resto de instancias, mientras que la ubicación no centrada de la media refuerza la idea de que ciertos documentos presentan similitudes más notables que otros, como se evidencia también en el mapa de calor. Esta combinación de valores atípicos y asimetría en la media de las distancias revela una diversidad en las relaciones entre los documentos, demostrando la complejidad y variabilidad en la similitud entre los artículos científicos del conjunto de datos.

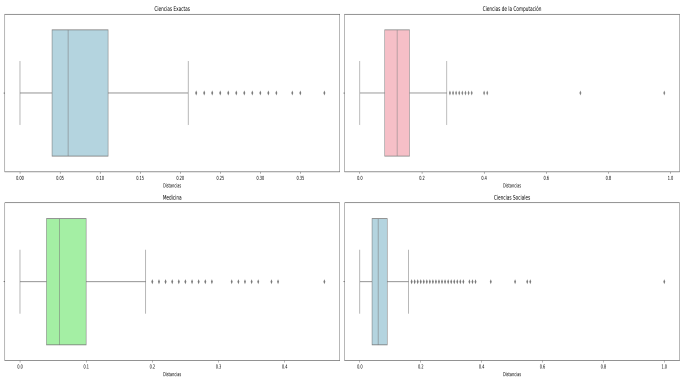

Fig. 9: Diagrama de caja y bigotes, distribuidas por cada área de estudio

La segunda etapa de experimentación analiza las medias obtenidas de las matrices de similitud por cada subconjunto de datos. Para esto se han planteado las hipótesis nula y alternativa, mostradas a continuación:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

Mediante un Análisis de Varianza (ANOVA) se comprobarán o rechazarán las hipótesis planteadas. Este análisis requiere agrupar los valores de la media y la desviación estándar como se muestra en la Tabla I. El análisis de las medias fue realizado mediante una *ANOVA* de dos factores los cuales fueron: el grupo al que pertenece la media y su

desviación estándar. Además se realizó la prueba de Tukey, que es utilizada para comparar las posibles combinaciones de las medias de cada subconjunto para identificar cuáles son significativamente diferentes entre sí.

El nivel de confianza determinado para la aceptación o rechazo de una hipótesis fue del 97%, esto implica que el nivel de significancia sea de 0.03. Al analizar los resultados obtenidos, en el lenguaje de programación estadístico R, se obtuvo un $p-value$ asociado de 2×10^{-16}, el cual es inferior a 0.03, lo que significa que se puede rechazar la hipótesis nula y aceptar la alternativa, por lo tanto los valores de las medias son diferentes entre sí. El valor del estadístico *F* fue de 4.528×10^{30}, que al ser un valor alto, implica que al menos una de las clases muestra diferencias significativas entre si. Para conocer las diferencias entre clases se utilizó la prueba de Tukey, los resultados se muestran en la siguiente Tabla II.

Grupo	Dif	Lim inferior	Lim Superior	Valor P
2-1	0.04	3.9×10^{-2}	4×10^{-2}	2.7×10^{-14}
3-1	-0.01	-1×10^{-2}	-9.9×10^{-3}	2.7×10^{-14}
4-1	-0.01	-1×10^{-2}	-9.9×10^{-3}	2.7×10^{-14}
3-2	-0.05	-5×10^{-2}	-4.9×10^{-2}	2.7×10^{-14}
4-2	-0.05	-5×10^{-2}	-4.9×10^{-2}	2.7×10^{-14}
4-3	0.00	-5.2×10^{-17}	5.2×10^{-17}	1

Tabla II: Resultado de la prueba de Tukey entre los valores de las medias de cada área

De acuerdo a los resultados mostrados en la Tabla II, se observa que entre los grupos 3 y 4 no existe una diferencia significativa, i.e. que estadísticamente estos dos grupos son iguales. Para el resto de combinaciones, entre áreas de estudio, se puede apreciar que existe una diferencia estadísticamente significativa. Esto permite concluir que los documentos seleccionados para los grupos 3 y 4 son similares a pesar de que pertenecen a distintas áreas del conocimiento.

Para la última etapa de experimentación, se usó el conjunto de datos completo con el objetivo de generar un mapa de calor que represente la relación entre todos los documentos, permitiendo así visualizar su similitud. Además, se procedió a calcular el tiempo de ejecución para evaluar la eficiencia de los algoritmos implementados en el proceso de NLP, así como en el cálculo de matrices de distancia. En el mapa de calor de la Figura 10 se muestran los 192 documentos y se observa la diferencia entre los grupos de cada área.

Esta diferenciación entre grupos demuestra que la mayoría de los documentos comparten similitudes más significativas con otros dentro de la misma área. También se aprecia que la relación entre artículos científicos de diferentes áreas es mínima. La media para la matriz ponderada fue de 0.07 con un valor de 0.08 para la desviación estándar. Este análisis refleja la capacidad de los algoritmos para discernir la relevancia y coherencia temática de cada documento, proporcionando una herramienta capaz de identificar la importancia relativa de los documentos de un conjunto de datos.

Por último, en la tercera etapa se evaluó el rendimiento de los algoritmos implementados en el *backend* del sistema web, para lo cual se llevó a cabo el proceso de NLP y el cálculo

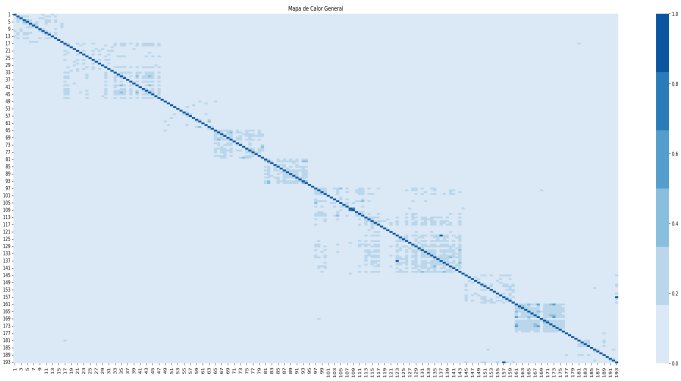

Fig. 10: Mapa de calor de todas las instancias del conjunto de datos

de matrices de similitudes en dos computadoras con características distintas. La primera, mencionada en la metodología del desarrollo de software (PC_1), y la segunda, equipada con un procesador Intel Core i7 de décima generación, 16 GB de memoria RAM, 1 TB de almacenamiento y 8 núcleos (PC_2). Los resultados indican que el tiempo de ejecución para la PC_1 y PC_2 fueron de 1.3555 y 1.2222 segundos respectivamente, mostrando una reducción de tiempo en la computadora con mejores características.

V. Conclusiones y Trabajos Futuros

En este trabajo, se abordó el desarrollo de un sistema web que tiene como objetivo identificar los documentos con mayor relevancia mediante el cálculo de matrices de similitudes. La finalidad del sistema es optimizar una etapa del proceso de SLR, reduciendo el tiempo y esfuerzo para seleccionar artículos científicos. Según los valores estadísticos obtenidos durante la experimentación, se observa que el algoritmo diseñado para calcular la similitud presenta una mayor precisión en las áreas de Ciencias Exactas y Ciencias de la Computación. Sin embargo, se debe considerar la posibilidad de que la eficacia del algoritmo para estas áreas se deba a un sesgo generado al momento de seleccionar de manera subjetiva los artículos científicos.

El prototipo inicial presentando en este articulo científico, se limitó a la comparativa en cuatro áreas, con el objetivo de evaluar la eficacia al identificar documentos relevantes. En trabajos futuros, se aspira eliminar esta restricción, permitiendo a los usuarios definir el número de grupos a analizar. En trabajos futuros, se sugiere considerar que para alcanzar una mejor precisión, es necesario optar por un *dataset* compuesto por documentos más homogéneos y heterogéneos dentro de cada grupo.

References

[1] B. Foster, "Transmission of knowledge," *A Companion to the Ancient near East*, pp. 261–272, 2020.

[2] M. D. Lombard, "Professional writing, technology, and the rhizomatic transmission of knowledge," Ph.D. dissertation, Purdue University, 2008.

[3] R. Jáimez, "Manual de redacción académica e investigativa: cómo escribir, evaluar y publicar artículos," *Letras*, vol. 53, no. 84, pp. 131–135, 2011.

[4] A. Martinson, *Guía para la redacción de artículos científicos destinados a la publicación*. Unesco París, 1983.

[5] W. I. B. Beveridge *et al.*, "The art of scientific investigation." *The art of scientific investigation.*, 1950.

[6] D. Curcic, "Number of academic papers published per year – WordsRated," https://wordsrated.com/number-of-academic-papers-published-per-year/, jun 2023, accessed: 2023-10-2.

[7] S. Schneegans, J. Lewis, and T. Straza, "Informe de la unesco sobre la ciencia: la carrera contra el reloj para un desarrollo más inteligente–resumen ejecutivo [internet]," 2021.

[8] W. To and B. T. Yu, "Rise in higher education researchers and academic publications," *Emerald Open Research*, vol. 2, p. 3, 2020.

[9] E. Araujo-Bilmonte, L. Huertas-Tulcanaza, and K. Párraga-Stead, "Análisis de la producción científica del ecuador a través de la plataforma web of science," *Cátedra*, vol. 3, no. 2, pp. 150–165, 2020.

[10] L. Moreira-Mieles, J. C. Morales-Intriago, S. Crespo-Gascón, and J. Guerrero-Casado, "Caracterización de la producción científica de ecuador en el periodo 2007-2017 en scopus," *Investigación bibliotecológica*, vol. 34, no. 82, pp. 141–157, 2020.

[11] A. P. Cardoso Ermel, D. P. Lacerda, M. I. W. M. Morandi, and L. Gauss, *Systematic Literature Review*. Cham: Springer International Publishing, 2021, pp. 19–30. [Online]. Available: https://doi.org/10.1007/978-3-030-75722-9_3

[12] R. van Dinter, B. Tekinerdogan, and C. Catal, "Automation of systematic literature reviews: A systematic literature review," *Information and Software Technology*, vol. 136, p. 106589, 2021.

[13] A. Dresch, D. P. Lacerda, and J. A. V. Antunes, *Systematic Literature Review*. Cham: Springer International Publishing, 2015, pp. 129–158. [Online]. Available: https://doi.org/10.1007/978-3-319-07374-3_7

[14] R. Alchokr, M. Borkar, S. Thotadarya, G. Saake, and T. Leich, "Supporting systematic literature reviews using deep-learning-based language models," in *Proceedings of the 1st International Workshop on Natural Language-based Software Engineering*, 2022, pp. 67–74.

[15] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*. Prentice Hall, 2010.

[16] O. G. Yalçın, *Natural Language Processing*. Berkeley, CA: Apress, 2021, pp. 187–213. [Online]. Available: https://doi.org/10.1007/978-1-4842-6513-0_9

[17] A. M. Nair, J. P. George, and S. M. Gaikwad, "Similarity analysis for citation recommendation system using binary encoded data," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–5.

[18] X. Cai, B. Xiao, C. Wang, and R. Zhang, "Quadratic-chi similarity metric learning for histogram feature," in *The First Asian Conference on Pattern Recognition*. IEEE, 2011, pp. 47–51.

[19] D. M. Shawky and A. F. Ali, "An approach for assessing similarity metrics used in metric-based clone detection techniques," in *2010 3rd international conference on computer science and information technology*, vol. 1. IEEE, 2010, pp. 580–584.

[20] J. Almas and U. Qamar, "Affect of data filter on performance of latent semantic analysis based research paper recommender system," in *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*. IEEE, 2020, pp. 50–54.

[21] D. Moya, L. Tapia, M. Albán, and G. Rodríguez, "Un enfoque de machine learning en el desarrollo de sistema recomendadores para procesos de investigación," *Revista Ibérica de Sistemas e Tecnologias de Informação*, no. E28, pp. 816–827, 2020.

[22] M. B. Magara, S. O. Ojo, and T. Zuva, "A comparative analysis of text similarity measures and algorithms in research paper recommender systems," in *2018 conference on information communications technology and society (ICTAS)*. IEEE, 2018, pp. 1–5.

[23] J. Chen, Y. Liu, S. Zhao, and Y. Zhang, "Citation recommendation based on weighted heterogeneous information network containing semantic linking," in *2019 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2019, pp. 31–36.

[24] W. Waheed, M. Imran, B. Raza, A. K. Malik, and H. A. Khattak, "A hybrid approach toward research paper recommendation using centrality measures and author ranking," *IEEE access*, vol. 7, pp. 33 145–33 158, 2019.