13 DE JULIO DE 2024

SISTEMA DE RECOMENDACIÓN

BÚSQUEDA DE ARTÍCULOS CIENTÍFICOS

LUIS CARLOS ORTEGA, DANIEL VALENCIA, LUIS ADOLFO BOTERO
UNIVERSIDAD DE CALDAS
Facultad de Ingenierías

Contenido

Modelo Propuesto

El modelo propuesto es un sistema de recomendación de artículos de investigación que considera el año de publicación y el número de citas, utiliza las bases de datos Web of Science (WOS), Scopus y Google Scholar. El sistema consta de tres componentes: entradas, procesos y salidas. El investigador ingresa una consulta de búsqueda, el sistema procesa la información en las bases de datos y analiza los resultados según los criterios establecidos. Luego, entrega una lista de 50 artículos y prioriza los más relevantes. Para utilizar las recomendaciones, el usuario debe ser miembro de la Universidad de Caldas e ingresar con el correo electrónico institucional, para asegurar que solo los estudiantes de universidades con acceso a estas bases de datos puedan acceder a los artículos completos.

El modelo clasifica los artículos obtenidos según el número de citas y el año de publicación, lo que permite filtrar artículos relevantes y recientes. Genera una lista de cincuenta documentos que el usuario puede revisar y seleccionar según sus necesidades. El sistema no personaliza las recomendaciones según la información del usuario, pero se propone incorporar esta funcionalidad en el futuro. La interfaz de usuario es intuitiva y permite ingresar el área de conocimiento deseada, para luego presentar los resultados en una tabla con detalles como título, autor, fecha de publicación, número de citaciones, resumen, URL y una medida de distancia que evalúa la relevancia de los artículos según la dispersión en el rango de fechas de publicación y el número de citaciones. Esta métrica ayuda a visualizar de manera numérica y gráfica el impacto relativo de los artículos recomendados.

Modelo genérico

El modelo genérico de sistema de recomendación de artículos de investigación se centra en proporcionar a los investigadores una herramienta eficaz para descubrir y seleccionar artículos relevantes en su campo de estudio. Este sistema opera en tres componentes principales: entradas, procesos y salidas. En la etapa de entradas, el investigador introduce una consulta específica relacionada con su área de interés. A continuación, el sistema procesa esta información mediante la búsqueda y recopilación de datos de múltiples bases de datos académicas, tales como *WOS*, *Scopus* y *Google Scholar*.

Durante el proceso, se extraen características clave de los artículos, como título, autores, año de publicación, resumen y número de citas. Luego, el sistema aplica algoritmos de análisis y clasificación, que utiliza métricas como la distancia euclidiana para comparar la similitud entre los artículos basados en criterios como el año de publicación y el número de citas. Por último, el componente de salidas presenta al investigador una lista de artículos recomendados y destaca los más relevantes según los parámetros establecidos. Además, proporciona visualizaciones gráficas que facilitan la comprensión de la distribución de similitudes y la comparación de características entre los artículos recomendados y el artículo de referencia.

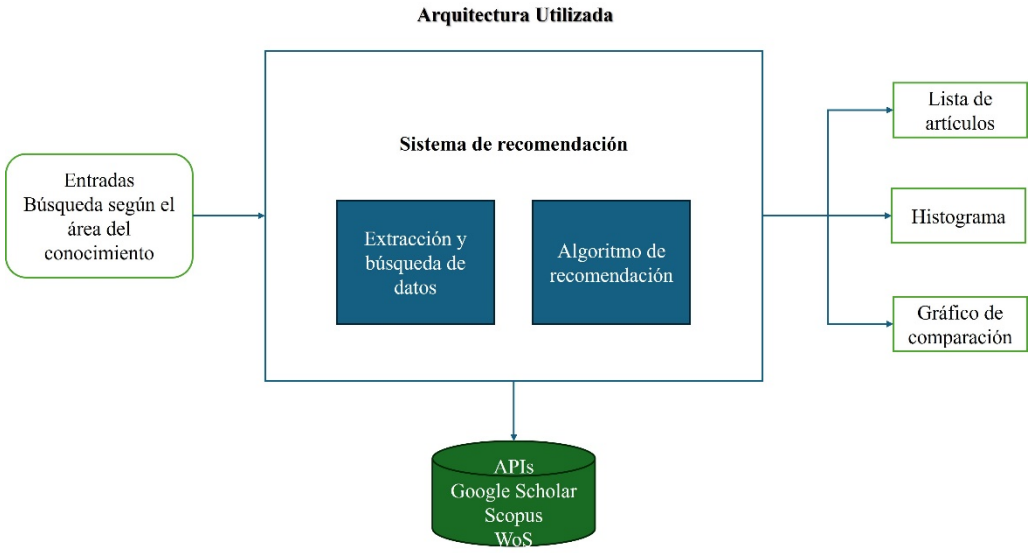

Figura 1: Modelo genérico para la recomendación de artículos
Fuente: Elaboración propia

Componente 1: Entradas.

Las entradas al sistema de recomendación se refieren a toda la información necesaria para que el sistema funcione de manera adecuada. Los sistemas de recomendación necesitan información tanto de los usuarios como de los ítems para entregar resultados precisos:

- Usuarios: Los investigadores que ingresan consultas específicas en el sistema.
- Ítems: En este modelo genérico, la información de los ítems se refiere al contenido del área de conocimiento que el investigador desea consultar, e incluye detalles como el tema de investigación, palabras clave, y otros parámetros relevantes

Componente 2: Proceso de recomendación.

El proceso de recomendación del sistema se basa en la búsqueda y análisis de artículos de investigación en varias bases de datos:

1. Búsqueda de artículos: El sistema utiliza la consulta ingresada por el investigador para buscar artículos relacionados en las bases de datos *WOS*, *Scopus* y *Google Scholar*.
2. Obtención de características de los artículos: Para cada artículo encontrado, se extraen las siguientes características: título, autores, año de publicación, resumen, número de citas y URL.
3. Creación del *DataFrame*: Se compila un *DataFrame* con los metadatos de los artículos encontrados y valida que las columnas *year* y *num_citations* sean de tipo numérico.
4. Selección de un artículo de referencia: Se selecciona el primer artículo del *DataFrame* como artículo de referencia y se extraen sus características relevantes (año de publicación y número de citas).
5. Cálculo de distancias: Se calcula la distancia – similitud – entre el artículo de referencia y cada uno de los demás artículos mediante la norma Euclidiana. Las distancias se almacenan en una nueva columna del *DataFrame*.
6. Ordenamiento y selección de recomendaciones: Los artículos se ordenan por distancia y prioriza aquellos con menor distancia al artículo de referencia. Se seleccionan los 50 artículos más relevantes como recomendaciones.

7. Presentación de resultados: Se imprimen los detalles de los artículos recomendados, como el título, año de publicación, número de citas, autores, resumen y URL. Se visualizan los artículos recomendados en un formato de tabla y se generan gráficos para mostrar la distribución de distancias y la comparación de características entre el artículo de referencia y los artículos recomendados.

Componente 3: Salidas

El componente de salidas se refiere al dominio de aplicación del modelo y está diseñado para presentar los resultados de manera clara y accesible al investigador. Este componente incluye la entrega de una lista de 50 artículos de investigación recomendados y muestra para cada uno el título, autores, año de publicación, resumen, número de citas y URL para acceder al artículo. Además, se proporciona una visualización tabular de los artículos recomendados y se generan gráficos para mostrar la distribución de distancias entre los artículos evaluados y el artículo de referencia, así como una comparación de características clave (año de publicación y número de citas) entre el artículo de referencia y los artículos recomendados.

Esto permite a los investigadores comprender mejor la relevancia y similitud de los artículos recomendados para facilitar la toma de decisiones informadas en su área de investigación.

Estrategias de aplicación del modelo propuesto.

El modelo propuesto hace uso de una estrategia de búsqueda basada en *k-Nearest Neighbors* (k-NN) que utiliza la distancia euclidiana. Este algoritmo mide la similitud entre artículos de investigación y considera dos características principales: el año de publicación y el número de citas. Los artículos más cercanos al artículo de referencia en términos de estas características se consideran los más similares y, por lo tanto, se recomiendan.

Estrategia de Selección

La estrategia de selección emplea el algoritmo k-NN [1][2]. Este algoritmo es un método de *Machine Learning* que clasifica los datos en función de su proximidad a otros datos. En el contexto del sistema de recomendación, se calcula la distancia euclidiana entre el artículo de referencia y otros artículos en la base de datos. Los artículos con la menor distancia son seleccionados como los más relevantes.

Estrategia de Integración

La estrategia de integración se centra en cómo se combinan los datos de diferentes fuentes para generar las recomendaciones. El sistema recopila información de las bases de datos académicas mencionadas y las integra en un *DataFrame* que contiene metadatos de los artículos encontrados, como título, autores, año de publicación, resumen, número de citas y URL.

El proceso de integración sigue los siguientes pasos:

1. Recopilación de Datos: El sistema recoge los datos de las bases de datos mencionadas de acuerdo con la consulta ingresada por el usuario.
2. Unificación de Formatos: Los datos recopilados se unifican en un formato común y asegura que las columnas clave (como año de publicación y número de citas) sean de tipo numérico.
3. Análisis y Comparación: Se utiliza la distancia euclidiana para comparar la similitud entre los artículos. Esta distancia se calcula en base al año de publicación y el número de citas.

4. Visualización de Resultados: Los artículos recomendados se presentan en una tabla que incluye detalles como título, autores, año de publicación, número de citas, resumen y URL. Además, se generan gráficos para visualizar la distribución de distancias y la comparación de características entre los artículos recomendados y el artículo de referencia.

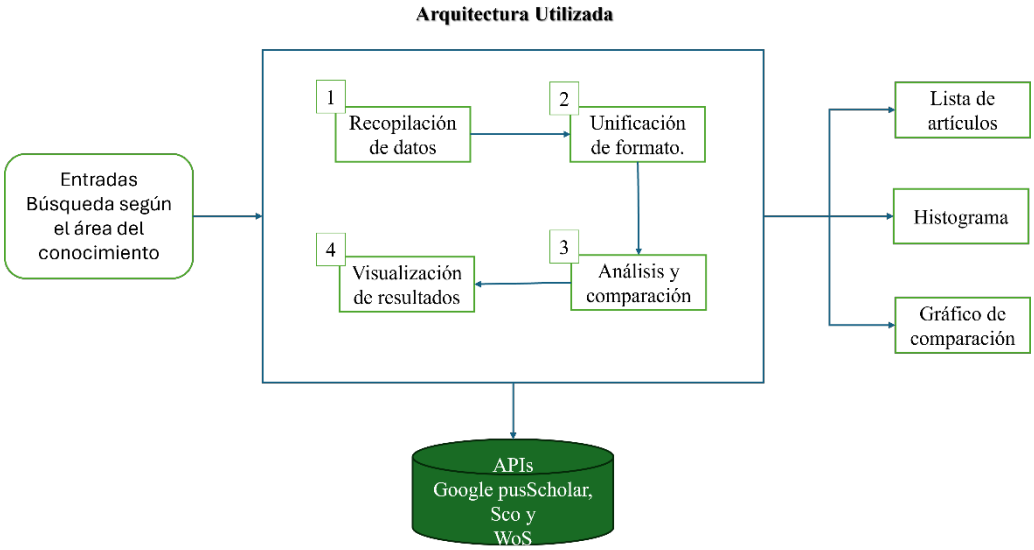

Figura 2: Modelo de la estrategia de integración
Fuente: Elaboración propia

Conclusiones

Basado en la implementación y los resultados del sistema de recomendación de artículos de investigación, se concluye que el modelo propuesto ha demostrado ser efectivo y valioso para facilitar la búsqueda y selección de literatura académica relevante. La integración de diversas bases de datos académicas como *WOS*, *Scopus* y *Google Scholar* aseguró una cobertura exhaustiva de información. Además, el uso de algoritmos de comparación basados en características clave como el año de publicación y el número de citas garantiza recomendaciones precisas y fundamentadas.

La presentación clara de los resultados mediante listados detallados de artículos recomendados y visualizaciones gráficas proporciona una interpretación accesible y comprensible para los investigadores. Aunque el sistema muestra robustez y utilidad práctica, futuras mejoras podrían incluir la integración de más bases de datos como *IEEE* y *Science Direct*, así como la expansión de características y criterios de relevancia. Además, se podría optimizar la experiencia del usuario para adaptarse mejor a diferentes contextos y necesidades específicas de la investigación académica.

Referencias

[1] J. Mora-Florez, G. Morales-España, and R. Barrera-Cárdenas, “Evaluación del clasificador basado en los k vecinos más cercanos para la localización de la zona en falla en los sistemas de potencia ,” *Ingeniería e Investigación* , vol. 28. scieloco , pp. 81–86, 2008.

[2] A.-J. Gallego, J. R. Rico-Juan, and J. J. Valero-Mas, “Efficient k-nearest neighbor search based on clustering and adaptive k values”, doi: 10.1016/j.patcog.2021.108356.