

Universidad de La Habana  
Facultad de Matemática y Computación



# **Recuperación Semántica de Información: Un enfoque integrado.**

Autor:

**Laura Victoria Riera Pérez**

Tutores:

**Dra. C. Lucina García Hernández**

**Lic. Carlos León González**

Trabajo de Diploma  
presentado en opción al título de  
Licenciado en Ciencia de la Computación

27 de noviembre de 2023

<https://github.com/LRiera24/semantic-information-retrieval>

Dedicación

# Agradecimientos

Agradecimientos

# Opinión del tutor

Opiniones de los tutores

# Resumen

Resumen en español

# Abstract

Resumen en inglés

# Índice general

Introducción	1
1. Estado del Arte	5
2. Propuesta	6
3. Detalles de Implementación y Experimentos	7
Conclusiones	8
Recomendaciones	9

# Índice de figuras



## Ejemplos de código

# Introducción

A lo largo de los siglos, mentes ilustres como las de Descartes, Newton y Bacon han tejido la noción cautivadora de que *el conocimiento es poder*. En la contrucción de este conocimiento, la información desempeña un papel fundamental, siendo el material primario del cual se extraen ideas, conceptos y comprensiones profundas. En la sociedad contemporánea, este papel adquiere una relevancia sin precedentes, consolidándose la información como una fuerza motriz esencial que impulsa los engranajes del progreso y facilita la toma de decisiones cruciales.

referencias

El origen de la Recuperación de Información, ese arte y ciencia de extraer conocimiento de vastos conjuntos de datos, estuvo impulsado por la necesidad de superar desafíos en el acceso a información. La gestión manual de registros, especialmente en el ámbito de las publicaciones científicas y los archivos de bibliotecas, implicaba una labor intensiva y propensa a errores.

referencia  
al Intro-  
duction

A medida que las computadoras comenzaron a desarrollarse y evolucionar, se reconocieron sus capacidades para manejar grandes volúmenes de datos y facilitar la recuperación de información. El uso de computadoras con este propósito se remonta a mediados del siglo XX. Durante la Segunda Guerra Mundial, se utilizaron computadoras electromecánicas como la IBM Harvard Mark I para realizar cálculos y procesar información relacionada con la guerra. En la década de 1950, se desarrollaron las primeras bases de datos electrónicas y sistemas de procesamiento de información, como CAIRS (Cranfield Automatic Indexing System). Más tarde, en 1964, se desarrolló el sistema SMART (System for the Mechanical Analysis and Retrieval of Text), que utilizó algoritmos para la indexación y recuperación de información textual. Los años 1970s vieron avances en estándares como el Modelo de Recuperación de Información Probabilística, que permitió hacer búsquedas más precisas. Con la introducción de las computadoras personales en los 1980s y la irrupción de la World Wide Web en 1991, se transformó radicalmente el panorama, con hitos como Yahoo! en 1994 y el algoritmo PageRank de Google en 1996, marcando un nuevo paradigma orientado a la web. La década de 2010 presenció una transición hacia la personalización y búsqueda semántica, con eventos destacados como el lanzamiento del algoritmo Hummingbird por Google en 2013.

ref

ref

ref

ref

ref

ref

ref

ref

ref

A partir de 2020, las tendencias fundamentales en la evolución de la Recuperación

de Información han impulsado una exploración profunda en áreas como la inteligencia artificial, el aprendizaje de máquinas y el procesamiento del lenguaje natural, con el objetivo de perfeccionar la precisión de los resultados de búsqueda. Este período ha sido testigo del surgimiento de enfoques innovadores como la búsqueda conversacional y la personalización y recomendación de contenido, mejorando directamente la experiencia del usuario; así como la representación semántica y relacional, integrando la esencial comprensión contextual. Además, se han explorado métodos para la extracción de patrones y relaciones complejas, y la clasificación y organización temática.

La transformación significativa de la Recuperación de Información, que ha pasado de ser un ámbito exclusivo de profesionales a involucrar a cientos de millones de personas en la búsqueda diaria de información, ha tenido un impacto profundo en diversas esferas. En el ámbito académico, los Sistemas de Recuperación de Información (SRI) no solo agilizan la investigación, sino que también fomentan la colaboración entre investigadores y facilitan el acceso a recursos compartidos, promoviendo así la difusión eficiente de conocimientos. Paralelamente, en el entorno empresarial, estos sistemas contribuyen a la productividad simplificando la búsqueda de información esencial y respaldando decisiones estratégicas. Asimismo, en redes sociales, personalizan la experiencia del usuario y fortalecen la conexión con recomendaciones adaptadas, mientras que en la industria médica, agilizan diagnósticos y tratamientos. Además, en el ámbito cultural, contribuyen a la preservación y acceso a archivos históricos y museos virtuales.

poner referencias

A pesar de estos avances notables, persisten desafíos significativos en los SRI. La ambigüedad semántica, consultas vagas que dificultan la comprensión precisa de la intención del usuario, junto con problemas de relevancia, como el ruido de información y la sensibilidad al contexto, plantean obstáculos a la eficacia de la recuperación de información. Limitaciones tecnológicas, como la dificultad para manejar información multimedia y desafíos en el procesamiento del lenguaje natural, también presentan retos. Además, las preocupaciones éticas, la privacidad del usuario y la adaptación a nuevos contenidos emergentes son áreas críticas a abordar.

Este trabajo se centrará en uno de los enfoques fundamentales de la clasificación y organización temática: los modelos de tópicos. Estos modelos matemáticos, propuestos en y consolidados con el modelo Latent Dirichlet Allocation (LDA) , poseen la capacidad de descubrir patrones temáticos subyacentes y organizar documentos de manera automática según su tema, proporcionando información sobre las palabras que componen cada uno. Sin embargo, enfrentan desafíos actuales, como la adaptación a contextos cambiantes, ya que dependen de hiperparámetros cruciales, cuya configuración se basa en información extraída del corpus y que impacta directamente en la deducción de las temáticas. Este ajuste no es automático y depende del análisis experto para su determinación. En entornos dinámicos, donde la cantidad y naturaleza de los tópicos pueden cambiar con el tiempo, esta adaptación constante

ref

ref

puede representar un problema. Otro desafío es la subjetividad en la interpretación de temas, ya que no proporcionan un nombre específico para los tópicos y requieren la intervención de expertos para su identificación. Presentan también problemas en la gestión efectiva de la polisemia, la cual implica distinguir entre los diversos significados de una palabra para una interpretación precisa del tópico. La mejora de la robustez y la capacidad para manejar la variabilidad del lenguaje son áreas cruciales de investigación para perfeccionar los modelos de tópicos.

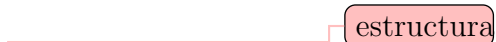
En combinación, se trabajará con los ampliamente utilizados motores de búsqueda, los cuales facilitan el acceso a vastas cantidades de información a millones de personas. A pesar de sus beneficios, enfrentan limitaciones notables, especialmente en la precisión de los resultados de búsqueda. La ambigüedad del lenguaje humano hace que la común búsqueda por palabras clave ya no sea suficiente, generando resultados no deseados o excluyendo información relevante. La capacidad limitada para comprender el contexto y la intención del usuario, subrayan la necesidad constante de mejoras en las herramientas de búsqueda web. La desorganización en la visualización de los resultados es otro desafío, ya que la creciente cantidad de información en línea dificulta la identificación rápida de datos pertinentes. La eficacia en la consulta del corpus como parte del proceso de recuperación de resultados relevantes se vuelve una limitación palpable, ya que examinar la totalidad de la base de datos puede desencadenar procedimientos considerablemente menos eficientes. Esta circunstancia resalta la importancia de implementar estrategias más selectivas y ágiles en la búsqueda de información.

Considerando los desafíos anteriormente mencionados, ¿puede la implementación de mejoras en los modelos de tópicos, específicamente en la adaptación a contextos cambiantes y la gestión de la subjetividad en la interpretación de temas, fortalecer la eficiencia de los Sistemas de Recuperación de Información? ¿Puede la aplicación de estrategias más selectivas y ágiles, junto con una comprensión mejorada del contexto, mejorar la eficiencia de los motores de búsqueda en la era de la información abundante y diversa? ¿Puede la visualización descriptiva o clasificada de los resultados mejorar la experiencia del usuario en la interacción con los sistemas de búsqueda de información?

El objetivo general consiste en concebir, diseñar e implementar una solución computacional basada en el paradigma de un motor de búsqueda que automatice los procesos de los modelos de tópicos, con énfasis en la adaptación a contextos cambiantes y la gestión de la subjetividad en la interpretación de temas, durante la etapa de preprocesamiento. Con esto se quiere desarrollar un sistema más flexible y eficiente. Además, se busca que esta solución, durante la etapa de recuperación, optimice la obtención de resultados relevantes despreciando parte del corpus. Se incorporará también una mejora en la visualización de los resultados de búsqueda para potenciar la experiencia del usuario en la interacción con el sistema de Recuperación de Información.

Para alcanzar el cumplimiento del objetivo general, se proponen los siguientes objetivos específicos:

1. Estimar la cantidad de tópicos de forma automática en un corpus dado.
2. Asignar nombres significativos a los tópicos identificados.
3. Incorporar medidas de similitud para manejar la incertidumbre en el dominio de aplicación.
4. Reducir el espacio de búsqueda y mejorar la eficiencia en la recuperación de información.

El resto del presente documento se ha estructurado en ...  estructura

# Capítulo 1

## Estado del Arte

# Capítulo 2

## Propuesta

## Capítulo 3

# Detalles de Implementación y Experimentos



# Conclusiones

Conclusiones

# Recomendaciones

Recomendaciones