Universidad de La Habana
Facultad de Matemática y Computación

Recuperación Semántica de Información: Un enfoque integrado.

Autor:
Laura Victoria Riera Pérez

Tutores:

Lic. Carlos León González
Dra. C. Lucina García Hernández

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

26 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

Índice de figuras

Ejemplos de código

Introducción

Según Sir Francis Bacon, “el conocimiento es poder”, y 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.

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 Introduction

A medida que las computadoras comenzaron a desarrollarse y evolucionar, se reconocieron sus capacidades para manejar grandes cantidades 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) . En la década de 1960, 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 . La década de 1970 vio 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 la década de 1980 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

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 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, facilitando así el acceso al patrimonio cultural.

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 estudio se centrará en uno de los enfoques fundamentales de la clasificación y organización temática: los modelos de tópicos. Estos modelos poseen la capacidad de descubrir patrones temáticos subyacentes y organizar documentos de manera automática según su tema, proporcionando información detallada sobre las palabras que componen cada uno. Sin embargo, enfrentan desafíos actuales, como la adaptación a contextos cambiantes, ya que la entrada necesaria es la cantidad de tópicos presentes en el corpus. Este ajuste no es automático y depende del análisis experto para determinar la cantidad correcta de tópicos. En entornos dinámicos, donde la cantidad y naturaleza de los tópicos pueden cambiar con el tiempo, esta adaptación constante puede representa 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. Además, presentan problemas en

la gestión efectiva de la polisemia, lo 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 estos modelos.

El problema científico abordado en este trabajo se centra en la adaptación a contextos cambiantes, la subjetividad en la interpretación de temas y la ineficiencia en la comparación de consultas con cada tópico presente en el corpus.

La pregunta científica que se plantea es: ¿Será posible desarrollar una herramienta computacional que incorpore técnicas de procesamiento semántico, permitiendo la automatización de los procesos de estimación de la cantidad de tópicos en un corpus y la asignación de nombres a dichos tópicos? ¿Será factible despreciar parte del corpus al recibir una consulta y, al mismo tiempo, obtener todos los resultados relevantes a la misma?

El objetivo general consiste en concebir, diseñar e implementar una solución computacional, enmarcada en el paradigma de un motor de búsqueda, que aborde la automatización de los procesos de los modelos de tópicos planteados en el problema en la etapa de clasificación, logrando un sistema más flexible.. Además, se busca que esta solución sea capaz de despreciar parte del corpus en la obtención de resultados relevantes a una consulta, con el fin de aumentar la eficiencia en la etapa de recuperación.

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 ...

creo que aquí podria hablarse un poco de motores de busqueda, como se realizan las mismas y que problemas existen

los??

las??

hay que ponerlo más conciso?

estructura

Capítulo 1

Estado del Arte

Capítulo 2

Propuesta

Capítulo 3

Detalles de Implementación y Experimentos

Conclusiones

Conclusiones

Recomendaciones

Recomendaciones