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

7 de diciembre de 2023

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

Dedicatoria

Agradecimientos

Agradecimientos

Opinión de los tutores

Opiniones de los tutores

Resumen

Resumen en español

Abstract

Resumen en inglés

Índice general

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 construcción 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.

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, Alan Turing y las computadoras británicas Colossus fueron fundamentales para descifrar mensajes encriptados nazis. También se utilizaron computadoras electromecánicas, como la IBM Harvard Mark I, para cálculos y procesamiento de información bélica. . En los 1950s, se implementaron sistemas como el General Electric, que buscaba más de 30,000 resúmenes de documentos, representando un hito inicial en el uso de computadoras para gestionar grandes conjuntos de información. Durante la década de 1960, se destacaron avances en la formalización de algoritmos para clasificar documentos en relación con una consulta. Un enfoque destacado consideraba documentos y consultas como vectores en un espacio N-dimensional. . Durante los aõs 1970, se produjeron avances significativos, como la complementación de los pesos de frecuencia de término (tf) de Luhn, basados en la ocurrencia de palabras dentro de un documento, con el trabajo de Spärck Jones sobre la ocurrencia de palabras en el conjunto de documentos de una colección. . Con el auge 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

referencias

referencia al Introduction

ref

ref

ref

ref

ref

ref

ref

paradigma orientado a la web. En los 2000 se presenció una transición hacia la personalización y la búsqueda semántica , con eventos destacados como el lanzamiento del algoritmo Hummingbird por Google en 2013 .

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 generalización a la gran mayoría de aplicaciones de 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, ha pasado de ser un ámbito exclusivo de profesionales a involucrar a cientos de millones de personas en la búsqueda diaria de información, provocando 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. También, 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.

En el colectivo de Sistemas de Información de la Facultad de Matemática y Computación (MATCOM) de la Universidad de La Habana, se evidencia una rica trayectoria de investigación y logros en diversas temáticas. Entre los antecedentes, se encuentran trabajos como el presentado por Quintana Wong, García Hernández, Guillot Jiménez y Amable Ambrós en COMPUMAT 2019, que abordó recomendaciones para la promoción de la salud mediante el uso de bases de datos NoSQL. También,

añadir referencias

se resalta la investigación de Quintana Wong, García Garrido y García Hernández en IWOR 2019, donde integraron modelos de vecindario y factores latentes para obtener recomendaciones precisas. La participación en eventos internacionales, como la Escuela Latinoamericana de Verano en Investigación Operativa (ELAVIO) 2019 en Lleida, España, refleja el compromiso del colectivo con trabajos que exploraron la combinación de filtrado colaborativo y modelación de tópicos para la recomendación de información. En ese mismo año también se destaca el trabajo de Leon González y García Garrido, el cual abordó modelos de generación de tópicos con word embedding, explorando modelos probabilistas como LDA y presentando el algoritmo lda2vec. Además, Quintana-Wong, García Hernández, y colaboradores han contribuido con artículos en revistas especializadas, como el estudio sobre sistemas de recomendación en soluciones analíticas transaccionales para la atención médica. Prado Romero y colaboradores, por su parte, han destacado en la predicción de la popularidad de temas en proveedores de noticias, presentando sus avances en ICOR 2020 y con un artículo aceptado en el "Intelligent Data Analysis Journal". Asimismo, se destaca la colaboración internacional en un proyecto conjunto con profesores de la Universidad de L'Aquila, Italia, para desarrollar un Sistema de Recuperación de Información relacionada con la COVID-19, llevando a cabo esta iniciativa de manera efectiva durante los meses de mayo a julio de 2020.

El presente trabajo se centrará en uno de los enfoques fundamentales de la clasificación y organización temática: los modelos de tópicos. Estos modelos matemáticos 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 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 simple búsqueda por palabras clave ya no sea suficiente, al generarse resultados no deseados o excluirse 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 en la 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 eficiencia 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 más lentos. Esta circunstancia resalta la importancia de implementar estrategias más selectivas y ágiles en la búsqueda de información.

Por tanto, el problema científico a abordar es la dificultad de los modelos de tópicos para resolver eficazmente el ajuste automatico de hiperparámetros en entornos dinámicos, así como la limitación para la precisión semántica de los tópicos.

Considerando los desafíos anteriormente mencionados las preguntas científicas a responder en esta tesis son, ¿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, incrementar 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 ampliar la experiencia del usuario en la interacción con los sistemas de búsqueda de información?

El objetivo general de este trabajo consiste en concebir, diseñar e implementar una solución computacional, mediante la creación de un motor de búsqueda que automatice los procesos de los modelos de tópicos, contribuyendo a la adaptación a contextos cambiantes, la gestión de la subjetividad en la interpretación de temas, y la optimización de los resultados.

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

1. Profundizar en el marco teórico y conceptual de los SRI, dando prioridad a la comprensión de modelos de tópicos, motores de búsqueda y ontologías.

2. Realizar un estudio exhaustivo del estado actual en entornos dinámicos y en la literatura académica sobre SRI, identificando tendencias clave.

3. Concebir y diseñar estrategias para potenciar la adaptabilidad de los modelos de tópicos en entornos dinámicos y perfeccionar la interpretación semántica de temas.

4. Implementar y evaluar las estrategias concebidas, integrándolas en SRI para medir su eficiencia y efectividad sinérgica en el contexto de adaptabilidad y precisión semántica.

El contenido restante de esta tesis se organiza en cuatro capítulos, abarcando las distintas etapas que constituyen el desarrollo del trabajo. En el Capítulo 2, "Marco Teórico-Conceptual", se proporciona un análisis detallado del estado actual de la ciencia y tecnología en las áreas relevantes, sirviendo como fundamento esencial para la investigación y los resultados obtenidos. El Capítulo 3, "Concepción y Diseño de la Solución", aborda la caracterización general de la propuesta computacional, la arquitectura del sistema, la estructura del modelo analítico y los procesos vinculados a la precisión semántica. Detalles técnicos de la implementación del sistema se presentan en el Capítulo 4, "Implementación y Experimentación" donde se explora cualitativa y experimentalmente la validez de la solución implementada, aprovechando las herramientas disponibles. En la parte del desenlace, se exponen las conclusiones, destacando los logros clave en relación con los objetivos planteados, así como las recomendaciones que señalan futuras direcciones de investigación. La bibliografía utilizada para respaldar la base científica de la solución propuesta y los anexos complementarios se incluyen para facilitar la exploración de temas relacionados.

Capítulo 1

Marco Teórico - Conceptual

1.1. Modelos de Tópicos

1.2. Ontologías

Capítulo 2

Concepción y diseño de la solución

2.1. Estado del Arte

2.1.1. Estimación automática del número de tópicos

2.1.2. Asignar etiqueta a un tópico

2.1.3. Despreciar parte del corpus

2.1.4. Visualización de resultados

Capítulo 3

Implementación y Experimentación

Conclusiones

Conclusiones

Recomendaciones

Recomendaciones