Comparación de Sistemas de Recuperación de Información: BM25 vs. DPR

Eisler Francisco Valles Rodriguez, Rafael Acosta Márquez, and Jorge Alejandro Pichardo

Universida de La Habana, La Habana, Cuba

Abstract. Este informe presenta una comparación entre dos sistemas de recuperación de información: BM25, un modelo tradicional sin el uso de inteligencia artificial, y Dense Passage Retrieval (DPR), un modelo moderno que integra técnicas de inteligencia artificial. Se evaluaron ambos sistemas en términos de precisión y eficiencia, utilizando un conjunto de datos de prueba. BM25 se destacó por su simplicidad y velocidad, mientras que DPR mostró una mayor precisión en la recuperación de documentos semánticamente relevantes, aunque a un costo computacional más alto. Se concluye con una discusión sobre las limitaciones de ambos enfoques y propuestas de mejora futuras.

Keywords: Recuperación de información, BM25, DPR, Inteligencia artificial, Similitud semántica

1 Introducción

La recuperación de información es un campo fundamental en la ciencia de la computación, con aplicaciones que van desde motores de búsqueda hasta sistemas de recomendación. Este trabajo compara dos enfoques de recuperación de información: BM25, un método basado en la relevancia de palabras clave, y Dense Passage Retrieval (DPR), un modelo que utiliza representaciones densas de consultas y documentos para mejorar la precisión. El objetivo fue evaluar la efectividad de ambos métodos en términos de precisión y costo computacional.

2 URL del Proyecto y Autores

El proyecto se encuentra alojado en GitHub y es accesible a través del siguiente enlace: SRI-Project.

3 Descripción del Tema

La recuperación de información ha evolucionado desde métodos simples basados en coincidencias de palabras clave hasta enfoques más sofisticados que emplean técnicas de inteligencia artificial para capturar la semántica de las consultas.

BM25 es un método probabilístico tradicional que ha sido ampliamente utilizado en sistemas de búsqueda. En contraste, DPR utiliza redes neuronales para generar embeddings densos que capturan el significado contextual de las consultas y los documentos, ofreciendo una recuperación más precisa en situaciones complejas.

4 Antecedentes

BM25 ha sido el estándar en recuperación de información durante varias décadas, siendo parte integral de muchos motores de búsqueda debido a su efectividad y simplicidad. Sin embargo, con el auge de la inteligencia artificial, han surgido métodos como DPR, que utilizan redes neuronales profundas para mejorar la precisión en la recuperación de información, especialmente en dominios donde la semántica es crucial. La capacidad de DPR para entender el contexto y el significado de las palabras lo diferencia de métodos tradicionales como BM25.

5 Soluciones Implementadas

En este trabajo se implementaron dos sistemas:

- **BM25:** Utilizando la librería `rank_bm25`, se implementó un sistema que calcula la relevancia de los documentos en función de la presencia de términos de consulta, ajustados por su frecuencia en el documento y en el corpus.
- **DPR:** Se utilizó la librería `transformers` de Hugging Face para cargar un modelo preentrenado de DPR, que genera embeddings densos de consultas y documentos. La similitud entre estos embeddings se utilizó para determinar la relevancia de los documentos.

6 Consideraciones Implementadas

Al implementar estos sistemas, se tomaron en cuenta las siguientes consideraciones:

- **Complejidad vs. Precisión:** BM25 fue seleccionado por su simplicidad y bajo costo computacional, mientras que DPR fue elegido por su capacidad para capturar relaciones semánticas complejas, aunque con un costo computacional más alto.
- **Escalabilidad:** Se evaluó la escalabilidad de ambos métodos, considerando el tiempo de respuesta y el uso de recursos, especialmente en grandes volúmenes de datos.
- **Contexto de la Consulta:** DPR se priorizó en consultas donde la interpretación del contexto era crucial, mientras que BM25 se utilizó en consultas más simples y directas.

7 Evaluación Cuantitativa y Cualitativa

7.1 Evaluación Cuantitativa

La precisión de ambos sistemas se evaluó utilizando la métrica de Exact Match Ratio en un conjunto de pruebas compuesto por 100 consultas. Los resultados fueron los siguientes:

- **BM25:** 0.57
- **DPR:** 0.60

7.2 Evaluación Cualitativa

Cualitativamente, se observó que DPR tenía un mejor desempeño en consultas complejas donde la semántica del contexto era crucial. BM25, aunque menos preciso en estos casos, se destacó en consultas simples y fue significativamente más rápido. Si bien es cierto que ambos alcanzan un porcentaje de precisión algo parecido, es notable cómo solo DPR es capaz de captar el contexto en el siguiente ejemplo:

- **Consulta:** Why whenever I get in the shower my girlfriend want to join?
- **Resultados de DPR:** They would have hot rough sex, or kill each other. This is a really dumb question.
- **Resultados de BM25:** Don't let apps that are liers put adds on your site. Like ones that say they have free age verification, but try charging your card. There all fucking liers, just dont understand why sites promote them and let them post there lies on your site. Also if you want your site to be better, when I click on mature anal lover's. It does not go to that it goes to a lying BS hook up site, that charges you for age verification. Like mature women just a bummer I can't get to the site. Big waist of my time really. Sure your making money from the bastards, if not making money from them. Your getting screwed then.

8 Declaración Autocrítica y Propuestas de Mejora

A pesar de los resultados obtenidos, se reconocen varias limitaciones en la implementación:

- **Costo Computacional de DPR:** Aunque DPR fue más preciso, su alto costo computacional lo hace menos práctico en entornos con recursos limitados. Se sugiere explorar técnicas de optimización como la reducción de la dimensionalidad o el uso de modelos más ligeros.
- **Limitaciones de BM25:** BM25 no es capaz de capturar relaciones semánticas complejas, lo que limita su efectividad en ciertas consultas. Un enfoque híbrido podría combinar lo mejor de ambos métodos, utilizando BM25 para consultas simples y DPR para las más complejas.

– **Generalización en Diferentes Dominios:** Los experimentos se realizaron en un conjunto de datos específico. Se propone realizar pruebas en diferentes dominios y con otros tipos de consultas para evaluar la generalización de los resultados.

9 Conclusión

Este trabajo comparó dos enfoques de recuperación de información: BM25, un método tradicional, y DPR, un modelo basado en inteligencia artificial. Aunque DPR mostró una mayor precisión en la recuperación de información semánticamente compleja, su alto costo computacional es una desventaja significativa. Un enfoque híbrido podría combinar la simplicidad y rapidez de BM25 con la precisión de DPR para lograr un equilibrio óptimo entre ambos.

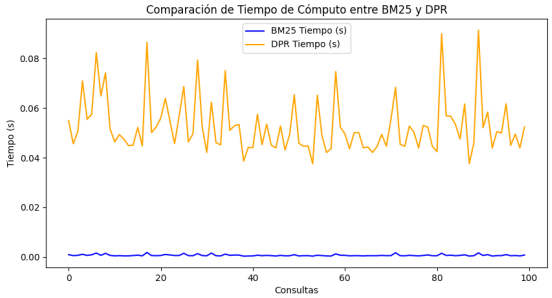

Fig. 1. Uso de CPU

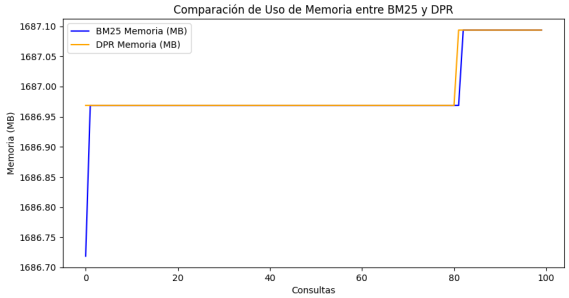

Fig. 2. Uso de memoria

10 Bibliografía

References

1. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, **3**(4), 333–389 (2009).
2. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al.: Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781 (2020).
3. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al.: Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (2020).