Memoria de Mejoras Implementadas en el Sistema de Recuperación de Información

Introducción

El objetivo de estas mejoras es potenciar la efectividad y robustez del sistema de recuperación de información original. Se han incorporado siete mejoras que permiten optimizar tanto la recuperación de documentos como la relevancia de los resultados. Estas mejoras buscan abordar desafíos comunes en sistemas IR, como la escasez de términos en consultas breves, la ambigüedad semántica, y la necesidad de un ranking más robusto.

Descripción de las Mejoras y Justificación

1. Pseudo-realimentación por Relevancia (PRF)

Descripción:

Se ejecuta una primera recuperación utilizando la consulta original y se seleccionan los cinco documentos más relevantes. A partir de ellos, se extraen los cinco términos más frecuentes (totalizando 25 términos) que se añaden a la consulta para una segunda recuperación.

Justificación:

Esta técnica mejora la cobertura de la consulta al incluir términos contextualmente relevantes que podrían haber sido omitidos inicialmente, lo que incrementa la probabilidad de recuperar documentos pertinentes.

2. Expansión de Consulta

Descripción:

Se implementa la expansión de la consulta mediante un pequeño diccionario de sinónimos. Por ejemplo, términos como "ciencia" se amplían a "conocimiento" y "estudio".

Justificación:

La expansión de consulta ayuda a superar la limitación de vocabulario (vocabulary mismatch) entre la consulta y los documentos, mejorando así la recuperación de información relevante.

3. Filtrado por Categorías

Descripción:

Una vez obtenidos los documentos relevantes, se extraen las categorías (por ejemplo, keywords presentes en el XML) y se permite al usuario filtrar los resultados según la categoría de interés.

Justificación:

El filtrado por categorías incrementa la precisión del sistema, permitiendo al usuario refinar los resultados y centrar la búsqueda en un ámbito temático específico.

4. Uso de BM25 para el Cálculo de Similitud

Descripción:

Se incorpora la métrica BM25, que ajusta la puntuación de relevancia basándose en la frecuencia del término, la longitud del documento y parámetros específicos (k1 y b).

Justificación:

BM25 es ampliamente reconocido por su rendimiento superior en comparación con la métrica TF-IDF tradicional. Su capacidad para manejar la variabilidad en la longitud de documentos y la distribución de términos lo hace ideal para sistemas IR modernos.

5. Indexación con Semántica Latente (LSI)

Descripción:

Se aplica la descomposición en valores singulares (SVD) a la matriz documento-término para reducir la dimensionalidad, extrayendo la “semántica” subyacente en los textos.

Justificación:

LSI permite capturar relaciones latentes entre términos y documentos, reduciendo el ruido y mejorando la recuperación en casos donde las coincidencias léxicas directas son insuficientes.

6. Uso de Otros Tokenizadores (Tokenización a Nivel de Sub-palabra)

Descripción:

Se incorpora la opción de utilizar tokenizadores preentrenados (por ejemplo, los basados en modelos BERT) que segmentan el texto a nivel de sub-palabra.

Justificación:

Esta estrategia mejora el manejo de palabras fuera de vocabulario y variaciones morfológicas, lo que resulta especialmente útil en lenguajes con alta inflexión como el español.

7. Codificación de Documentos con Modelos del Lenguaje (Embeddings)

Descripción:

Se integra una vía de búsqueda en la que tanto documentos como consultas se codifican utilizando modelos de sentence transformers. La similitud se mide mediante la distancia del coseno entre los vectores generados.

Justificación:

El uso de embeddings permite capturar la similitud semántica de forma más robusta que los métodos tradicionales basados en frecuencias, facilitando la identificación de documentos relevantes incluso cuando la coincidencia lexical es limitada.

Conclusión

La integración de estas mejoras responde a la necesidad de contar con un sistema de recuperación de información más inteligente y adaptable. Cada técnica ha sido seleccionada por su capacidad para abordar problemas específicos:

PRF y expansión de consulta mejoran la cobertura y relevancia al enriquecer la consulta original.

Filtrado por categorías y BM25 optimizan la precisión y el ranking de los resultados.

LSI y modelos de embeddings permiten capturar relaciones semánticas complejas que los enfoques tradicionales no abordan.

Finalmente, la opción de tokenización a nivel de sub-palabra mejora el procesamiento del lenguaje, manejando eficazmente la variabilidad lingüística.

Estas mejoras en conjunto elevan el rendimiento del sistema y lo acercan a estándares modernos en recuperación de información, ofreciendo resultados más precisos y relevantes para el usuario.