Mateo Trejo, Hugo; NIP 816678 & Bielsa Uche, Jaime; NIP 819033
6 de noviembre de 2023 - Ingeniería Informática - EINA UNIZAR

Recuperación de Información - Práctica 3

Evaluación de sistemas de recuperación de información

Índice

Programa de análisis de métricas

Nuestro programa utiliza un fichero con juicios de relevancia y otro con los resultados, sobre una o más búsquedas, con el fin de obtener métricas que expliquen cómo de certera ha sido la búsqueda (precisión, exhaustividad, F_β, precisión media por necesidad de información y las mismas como media de todas las búsquedas, incluyendo el MAP, además de puntos para formar una curva precision-recall).

Al comprobar tanto las curvas precisión recall como los valores de las métricas, se observa que los resultados son idénticos a los propuestos en el enunciado para los resultados de las búsquedas de los sistemas A y B al utilizar los primeros cincuenta documentos recogidos para cada necesidad de información. Esto indica que el programa está realizando bien el cálculo de las métricas (o al menos, igual que las propuestas en el enunciado).

Eficacia de nuestro sistema

Utilizamos nuestro programa de análisis para, sobre las necesidades de información, obtener resultados de búsqueda. Al igual que para los sistemas A y B, trabajamos con los 50 primeros resultados de cada búsqueda para obtener las métricas y la curva precision-recall, y compararla con dichos sistemas.

A nivel general, como se puede ver en la **figura 1** el sistema A es el mejor de los tres. Tiene una precisión muy alta en los primeros documentos, y se mantiene relativamente bien al aumentar el recall.

El sistema B tiene una curva similar a la del A, pero con una precisión menor. En los primeros documentos tiene una precisión menor, y decae más o menos en la misma proporción que el A. Es notable que en torno al 0.3% de recall la curva B se queda horizontal hasta el 0.6. Eso implica que en esta zona el sistema B ha recuperado documentos relevantes, y el A no.

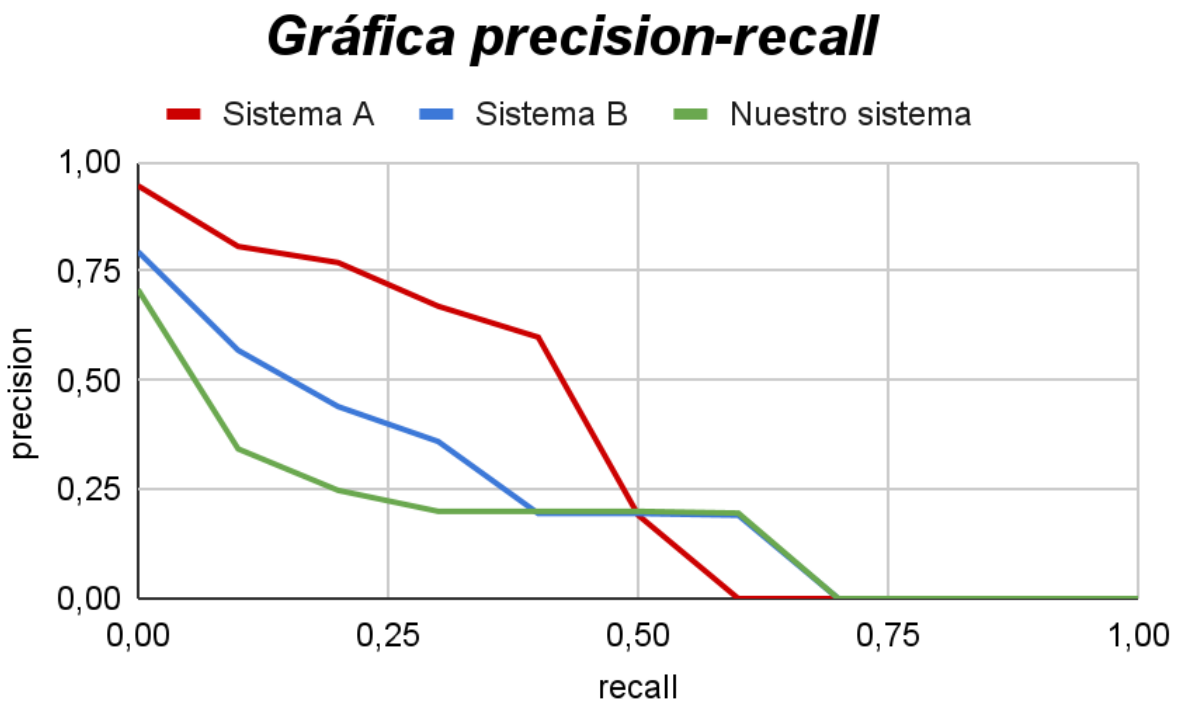

Figura 1: gráfica precision-recall de los sistemas A, B y el nuestro propio.

Nuestro sistema es muy similar al B, aunque los documentos recuperados al principio son un poco menos relevantes que los del B.

Tanto nuestro sistema como el B tiene una capacidad de recall ligeramente superior a la del A, dado que en las zonas más altas de recall lo superan en precisión. Ambos sistemas recuperan documentos menos relevantes que el A, pero una mayor cantidad de ellos. Por eso sus curvas superan a la del A en la zona anteriormente mencionada.

Comparando las métricas que se encuentran en la **figura 2**, se observa algo similar a las gráficas. La precisión y el recall son inferiores en el sistema B y en el nuestro, igual que la métrica F_1. El sistema A tiene un F_1 de 0.5, que indica que la precisión y el recall están balanceadas. Esto cuadra con que la curva tenga un decaimiento más o menos lineal.

Respecto a la precisión a 10, deducimos que el sistema A ha recuperado 7 documentos relevantes (de los 10), el B 5 y el nuestro 4. Nuestro sistema y el B quedan muy cerca, y el A de nuevo los supera a los dos notablemente.

El MAP es similar a lo anteriormente mencionado. Todas las consultas tienen una precisión de entorno a 0.7 en el A, y del 0.5 en el B y el nuestro.

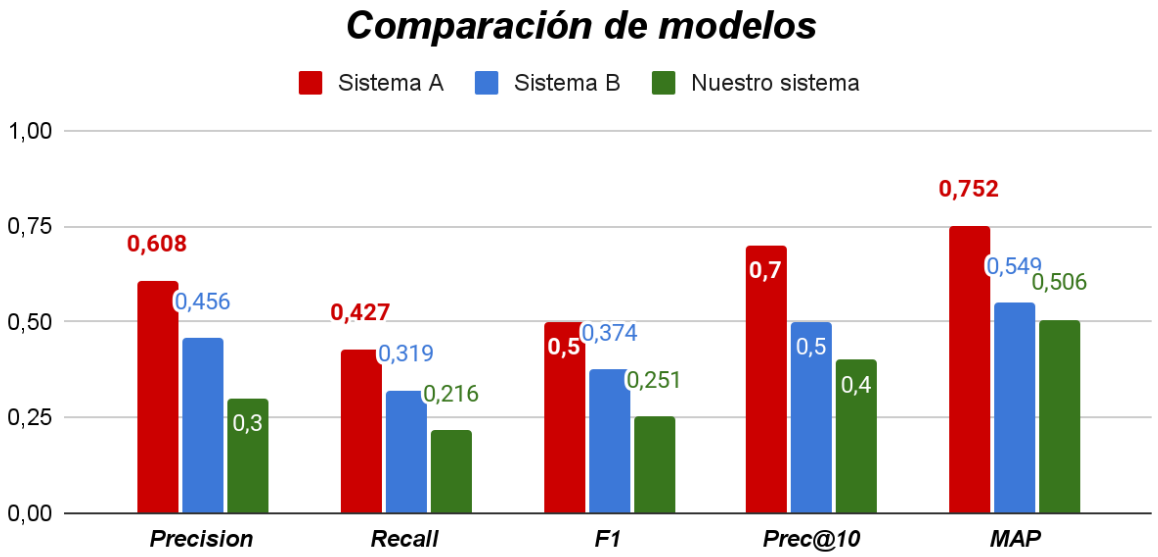

Figura 2: métricas para la eficacia de los sistemas A, B y el nuestro propio.

Análisis de las métricas para las búsquedas

Todas las métricas han aportado resultados similares. Realmente, en este caso lo más interesante habría sido comparar solo el F1 y el MAP, que dan una visión más general de todos los resultados.

Estas medidas son las más relevantes en la evaluación dado que combinan las medidas anteriores para dar resultados más generales, pero aun así el resto de métricas siempre son interesantes a la hora de hacer un análisis más en profundidad del rendimiento.

En la **figura 3** se observan las distintas métricas por búsqueda y los resultados medios. Las necesidades de información más relevantes son la 101-4 y la 106-4, las dos primeras. La primera tiene un rendimiento pésimo, en todos los sentidos. Concretamente, la query menciona que quiere que sean trabajos de fin de grado o máster. Muchos documentos tienen decenas de menciones al trabajo de fin de grado o máster, por lo que se recuperan como mucho más relevantes que los realmente relevantes (que tienen menos menciones a grado).

Además especifica que quiere que sean de los últimos 20 años. El sistema no soporta esta estructura, solo especificar fechas e intervalos del tipo “entre 2000 y 2023”.

La segunda consulta es más sencilla de procesar: especifica con claridad los términos relevantes y por ello el sistema es capaz de recuperar los mejores documentos.

La tercera consulta de nuevo menciona que sean trabajos de fin de grado, pero pese a ello consigue un resultado mejor ya que utiliza términos más específicos que la primera y estructuras menos complejas.

La cuarta consulta también tiene un mal resultado. De nuevo, utiliza una estructura temporal que el sistema no soporta: *hasta el siglo XIX*. El sistema no procesa los números romanos para traducirlos a cifras, aunque sería un cambio sencillo.

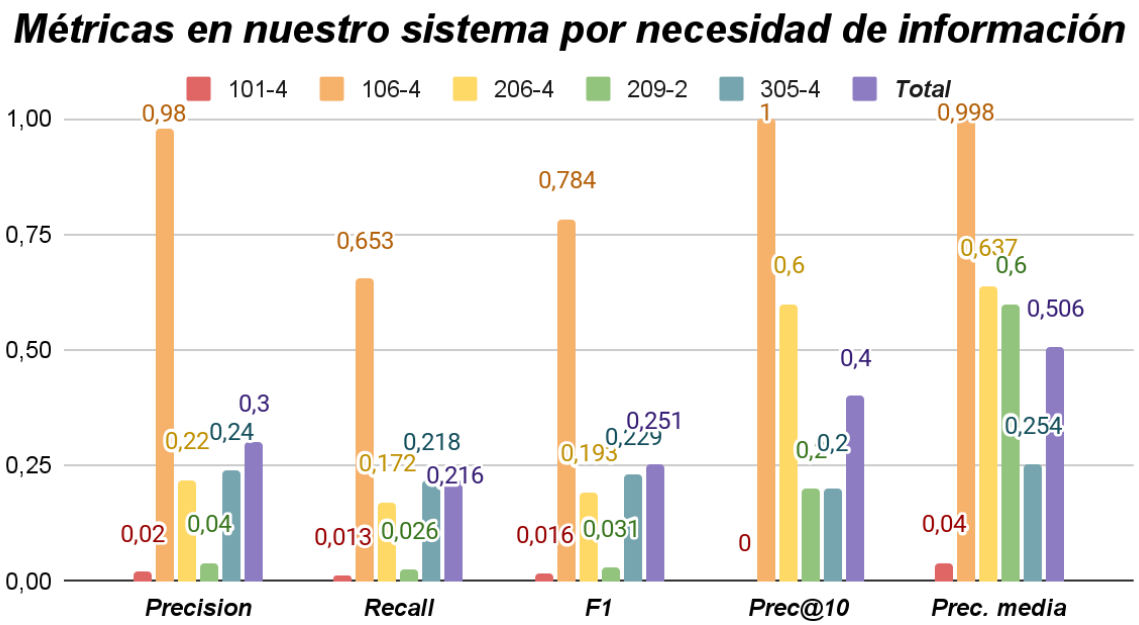

Figura 3: métricas para nuestro sistema, desglosadas por búsqueda.

La quinta consulta tiene un resultado intermedio. Da ejemplos de cosas que quiere ver, lo cual no cuadra del todo bien con el procesamiento del sistema, que interpreta que es altamente recomendable que esos elementos aparezcan. Aun así, el resultado no es del todo malo.

Conclusión

En general el mal resultado del sistema proviene del procesamiento de estructuras complejas del lenguaje natural como las muchas formas de especificar fechas e intervalos, y algún problema con la relevancia de los términos. Además, conviene ser selectivos con los campos en los que se buscan las palabras que obtenemos, en vez de buscar los términos en todos los campos del documento. Por ejemplo, para los campos identificador o tipo sería interesante una búsqueda más especializada. En una siguiente versión del sistema, ampliar esta capacidad de análisis sería el principal objetivo.