Recuperación de información

Jesús Germán Ortiz Barajas

08 de abril de 2021

1 Introducción

La recuperación de información aborda el problema de encontrar aquellos documentos cuyo contenido coincida con la solicitud de un usuario entre una gran colección de documentos. Los sistemas informáticos que devuelven documentos cuyo contenido coincide con una necesidad de información declarada se han denominado históricamente sistemas de recuperación de información. Los sistemas de recuperación de información buscan en una colección de documentos en lenguaje natural con el objetivo de recuperar exactamente el conjunto de documentos que pertenecen a la pregunta de un usuario [3].

A continuación se describen algunos modelos de recuperación de información:

1.1 Modelo de recuperación booleano

Es un modelo de recuperación basado en la teoría de conjuntos y el álgebra booleana de gran simplicidad. Su principal algoritmo de recuperación está fundamentado en un criterio de decisión binario sin ninguna noción de escala de medida ni ningún emparejamiento parcial en las condiciones de la consulta. En este modelo el método de representación, consiste en especificar los documentos como un conjunto de términos de indexación o *keywords*[2].

1.2 Modelo de recuperación estadístico

Este modelo utiliza un espacio vectorial basado en estadísticas, en el cual, un documento se documento se representa conceptualmente mediante un vector de palabras clave extraídas del documento, con pesos asociados que representan la importancia de las palabras clave en el documento y dentro de toda la colección de documentos.De la misma forma, una consulta se modela como una lista de palabras clave con pesos asociados que representan la importancia de las palabras clave en la consulta. Un enfoque común para determinar el peso de las palabras es el *tf-idf*, el cual se basa en dos factores: que tan frecuente aparece un término en un documento dado, y que tan frecuentemente aparece en toda la colección de documentos. Una vez que se determinan los pesos de los términos, es necesaria una función para medir la similitud entre los vectores de consulta y de documento. Una medida de similitud común, conocida como medida del

coseno, determina el ángulo entre los vectores del documento y el vector de consulta cuando se representan en un espacio euclidiano de dimensión V, donde V es el tamaño del vocabulario [1].

En las secciones siguientes se describen los sistemas desarrollados, para recuperar documentos relevantes al tema de sexualidad humana en un corpus de 15 documentos. Todo el código desarrollado está disponible de manera publica en Github [1]

2 Sistemas de recuperación booleana

2.1 Mediante matriz de incidencias

Para este primer sistema se construyó una matriz de incidencias con los 15 documentos que componen el corpus, en la cual se tienen 15 renglones, cada uno correspondiente a un documento del corpus, y 230 columnas, cada una correspondiente a una palabra del vocabulario. La posición de la matriz (i, j) tiene un valor de 1, si la palabra j, aparece en el documento i, en caso contrario su valor es cero. Para facilitar el acceso y las consultas, una vez obtenida la matriz de incidencias, se construyó un *dataframe* con la biblioteca *Pandas* a partir de dicha matriz. Las consultas corresponden a una lista de palabras de interés, las cuales se relacionan mediante operadores booleanos. Para realizar las consultas se utilizó la función *query* de la biblioteca *Pandas*, la cual utilizar operadores booleanos a parir de una cadena de texto ingresada. Para probar el funcionamiento de este primer sistema, se realizó la consulta:

$$q = (preservativo\, OR\, (\, sexualidad\, AND\, humana\,)\, OR\, orgasmo\, OR\, sexo\, OR\, fecundidad\, OR\, pelo)$$

Los documentos recuperados con esta consulta fueron: [1, 2, 3, 4, 5, 7, 12]

La tabla 1 muestra el rendimiento de este sistema con base en las métricas *precision*, *recall* y *f1-score*.

Precision	Recall	F1-score
0.8571	0.75	0.79

Table 1: Resultados del sistema booleano y matriz de incidencia

2.2 Mediante índice invertido

El índice invertido es una estructura que tiene la intención de mapear de forma eficiente los términos y los documentos en los que aparecen. Se forma de un diccionario de términos, y cada entrada del diccionario tiene una lista de postings, los cuales son los documentos en los que aparece dicho término. Para construir el índice invertido del corpus proporcionado, se partió de la matriz de incidencias construida en el sistema anterior, realizando un recorrido por todas

[1] https://github.com/jgermanob/TextMining/tree/master/Tarea%205

las columnas correspondientes a cada uno de los términos del vocabulario del corpus. Para realizar las consultas se pasa de notación infija a notación posfija y se utiliza una pila para respetar la precedencia de los operadores booleanos.

Para probar el rendimiento del sistema se utilizó la misma consulta q del sistema anterior:

$$q = (preservativo\, OR\, (\, sexualidad\, AND\, humana\,)\, OR\, orgasmo\, OR\, sexo\, OR\, fecundidad\, OR\, pelo)$$

Los documentos recuperados con esta consulta fueron: [1, 2, 3, 4, 5, 7, 12]

La tabla 2 muestra el rendimiento de este sistema con base en las métricas *precision*, *recall* y *f1-score*.

Precision	Recall	F1-score
0.8571	0.75	0.79

Table 2: Resultados del sistema booleano e índice invertido

3 Sistema de recuperación estadístico

Finalmente, el último sistema desarrollado emplea el algoritmo *tf-idf* para obtener una representación vectorial de los documentos del corpus. Posteriormente se obtuvo un índice invertido con base en los valores de tf-idf de cada término, en el cual, la lista de postings corresponde a los documentos en los que el valor de *tf-idf* para el término es distinto de cero. Una vez obtenida la lista de postings, esta se ordeno de forma descendente, de forma que el primer elemento de la lista es que tiene mayor valor *tf-idf*.

Para la recuperación de información, se obtuvo el vector de la consulta mediante el método *tf-idf* y se comparo con los vectores de los documentos que contienen dicha palabra mediante la similitud coseno. Para determinar si es un documento relevante con base esta medida, se creó una función que realiza esta operación y un umbral, si la similitud es mayor al umbral, el documento se agrega a la lista de documentos recuperados, en caso contrario no se agrega. Para este corpus, el valor del umbral fue de 0.20.

La consulta q realizada fue:

sexo animales orgasmo sexualidad humana planificación familiar

Los documentos recuperados con esta consulta son: [7, 15]

La tabla 3 muestra el rendimiento de este sistema con base en las métricas *precision*, *recall* y *f1-score*.

Precision	Recall	F1-score
1.0	0.25	0.4

Table 3: Resultados del sistema estadístico mediante *tf-idf*

4 Recuperación manual de documentos

Finalmente, se realizó una recuperación manual de los documentos relevantes al tema de sexualidad humana con base en la lectura de los 15 documentos.

Los documentos recuperados mediante este método son: [1, 2, 4, 7, 12, 13, 15]

La tabla 4 muestra el rendimiento de este método con base en las métricas *precision*, *recall* y *f1-score*.

Precision	Recall	F1-score
1.0	0.875	0.93

Table 4: Resultados del sistema de recuperación manual

5 Conclusiones

Es posible observar que los sistemas de recuperación booleana obtienen los mismos resultados si se emplea la misma consulta, sin embargo, la recuperación mediante índice invertido es más eficiente, ya que no necesita recorrer todos los documentos, sino que solo debe comparar en los que ya se sabe que aparecen las palabras de la consulta. Por otra parte, el sistema estadístico que utiliza el algoritmo *tf-idf* obtiene los peores resultados debido a la forma en que se realizó la consulta, en la que solo se agregaron las palabras relevantes al tema de sexualidad humana sin seguir una estructura.

References

[1] D. L. Lee, Huei Chuang, and K. Seamons. Document ranking and the vector-space model. *IEEE Software*, 14(2):67–75, 1997.

[2] Sandra Rodríguez Correa Paula Andrea Benavides Cañón. Procesamiento del lenguaje natural en la recuperación de información. 2007.

[3] Ellen M Voorhees. Natural language processing and information retrieval. In *International summer school on information extraction*, pages 32–48. Springer, 1999.