Proyecto Final de Sistema de Recuperación de Información

Roxana Peña[1], Marié del Valle[1], Josué Reodríguez[1]

Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba

1. Modelo de Recuperación de Información Booleano

1.1. Definición

[1] El Modelo de Recuperación de Información (MRI) Booleano es un cuádruplo $(D, Q, F, R(q, d_j))$:

D Conjunto de términos indexados.
Q Expresiones booleanas sobre términos indexados utilizando operaciones lógicas: *not*, *and*, *or*.
F Álgebra booleana sobre conjuntos de términos y conjuntos de documentos.
R Función booleana que indica si el documento d_j es relevante a la consulta q.

$$sim(d_j, q) = \begin{cases} 1 \text{ si } \exists \overrightarrow{q_{cc}} : (\overrightarrow{q_{cc}}^1 \in \overrightarrow{q_{fnd}}^2) \wedge (\forall {k_i}^3, g_i(\overrightarrow{d_j}) = g_i(\overrightarrow{q_{cc}})). \\ 0 \text{ en otro caso.} \end{cases}$$

1.2. Implementación

Clases

En esta primera implementación se conformaron cuatro clases:

- `Document`
  Representa la un documento de la colección. Tiene como atributos:id, título, autor y cuerpo.

- `Collection`
  Es una clase abstracta que representa una lista de `Document` pertenecientes a una colección específica. En la construcción de un objeto `Collection` se pasa como argumento la dirección donde se encuentra el archivo con el nombre de dicha colección. Tiene un método `parse` cuya función es leer los documentos del archivo y trasformarlos en un objeto `Document`, identificando en cada uno el título, autor y cuerpo. El cuerpo es una lista de los términos relevantes

[1] Una de las componentes conjuntivas de $\overrightarrow{q_{fnd}}$.
[2] Forma Normal Disyuntiva de la consulta q.
[3] Término i-ésimo de la consulta q.

del documento. Este método retorna una lista de `Document` y un conjunto de los términos relevantes que aparecen en todos los documentos.

- `BooleanIRM`
  Representa el Modelo de Recuperación de Información Booleano. En la construcción de un objetos `BooleanIRM` se pasa como argumento: la lista de términos relevantes, la lista de `Document` y la cadena consulta, elementos a los cuales se les aplica este modelo. Entre los métodos que implementa la clase están:
  - `indexing_terms`: Crea un diccionario de términos indexados, donde las llaves son los términos y el valor es un arreglo binario del tamaño de la cantidad de documentos en la colección.
  - `process_query`: Procesa la consulta, teniendo en cuenta los operadores lógicos. Considera el término *but* como operador *and*.
  - `process_boolean_op`: Tiene como parámetros el operador lógico, arreglo binario anterior al operador y arreglo binario posterior al operador. Retorna el arreglo binario resultante de aplicar la operación lógica a los dos arreglos pasados como argumento.
  - `retrieve_documents`: Retorna los documentos relevantes resultantes de procesar la consulta.

- `Newsgroups`: Clase que hereda de `Collection` e implementa el método `parse`, para el tipo específico de documentos de esta colección.

Métodos

El único método implementado, fuera de las clases es el método `start`. Los parámetros que recibe son: colección y MRI a utilizar, y la consulta. Se encarga de crear los objetos y ejecutar los métodos necesarios para la recuperación de la información. Retorna una lista con el título y autor todos los documentos relevantes.

1.3. Ejecución

En esta primera entrega, solo se tienen implementados el modelo booleano y la colección newsgroup. Antes de ejecutar el código, se debe escribir en la varia `DIR`, la dirección del archivo donde está la colección Newsgroup, en particular el archivo que tiene otros archivos, los cuales separan los documentos por temas comunes.

Referencias

1. Fleitas Aparicio, Carlos. Sánchez Aguilar, Marcel. Conferencia de Sistemas de Recuperación de Información: Modelos de recuperación de Información Booleano y Vectorial. 3er Año Licenciatura de Ciencia de la Computación. Departamento de Programación. Facultad de Matemática y Computación. Universidad de La Habana. 2022.