Proyecto de Sistemas de Información

Alexander Antonio González Fertel C-512 a.fertel@estudiantes.matcom.uh.cu
Hieu Do Ngoc C-511 a.fertel@estudiantes.matcom.uh.cu
Joel David Hernández Cruz C-511 j.cruz@estudiantes.matcom.uh.cu

Intro

Se presenta como solución una aplicación de escritorio basada en *eel*, que presenta un sistema de recuperación de información sobre un directorio a través de una interfaz amena e intuitiva. Para la implementación de la aplicación se utilizaron disímiles bibliotecas, listadas en *requirements.txt*.

Interfaz de Usuario

El sistema de recuperación de información implementado consta de una aplicación de escritorio como medio de comunicación con los modelos de recuperación de información implementados. Como se puede ver en la imagen 1, en la interfaz se presentan 2 zonas de interacción, al inicio (Configuración), se escoge el modelo a usar y se selecciona (provee) un directorio a indexar, los cuales se guardan para realizar las consultas a través del botón *Submit*. Debajo se presenta una entrada de texto para formular la consulta, lo cual deviene en un ranking de los documentos que se muestra luego al presionar el botón de búsqueda.

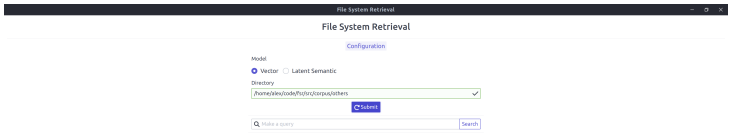

Figura 1: File System Retrieval

Procesamiento de Texto

El SRI es capaz de recibir documentos de texto plano (txt) y también documentos de formato portable (pdf). En el caso de los documentos pdf, debido a la naturaleza de su codificación y la inexistencia de un éstandar, es posible que para algunos documentos su decodificación funcione de manera correcta y sus textos son extraídos en su totalidad, y para otros este no devuelva el texto correcto de los documentos.

Una vez extraídos los textos de los documentos del directorio selecionado, se le aplica los siguientes procesamientos:

- Análisis lexicográfico
- Eliminación de stopwords

- Lematización

Se ha optado por utilizar la técnica de *lematización* en vez de *stemming* porque este es un algoritmo más complejo que produce resultados con mejor precisión debido a que utiliza conocimientos linguísticos, a diferencia de *stemming*

Modelo

El modelo de recuperación de información principal implementado por el SRI es el modelo de Indexación Semántica Latente (LSI).

El modelo LSI es una variación del Modelo Vectorial, en la que los documentos se representan a partir de vectores de pesos no binarios, al igual que las consultas, la función de similitud es el coseno del ángulo entre el vector del documento y el de la consulta.

La idea principal de este modelo consiste en hacer un mapa entre cada documento y el vector consulta a un espacio de dimensiónalidad reducida el cual está asociado a conceptos.

Definición 1. *Sea t la cantidad de términos índice en la colección y N el total de documentos. Se define* $M = (m_{i,j})$ *como la matriz de asociación de término-documento con t filas y N columnas. Cada elemento* $m_{i,j}$ *de la matriz M es el peso asociado a la pareja término i - documento j.*

Existen diferentes formas de generar estos valores $m_{i,j}$, ya sea usando la frecuencia término-documento o usando la matriz de tf-idf. En el SRI implementado, se utiliza la matriz tf-idf como la matriz de asociación término-documento, al ser este una de las técnicas de *"term-weighting"* más populares que existen en la actualidad. Además, su implementación resulta bastante sencillo y eficiente haciendo uso de la librería $scikit - learn$.

Para lograr la reducción de dimensiones de la matriz M, el modelo LSI propone utilizar la técnica de descomposición SVD para descomponer la matriz M en 3 componentes de la manera siguiente.

$$M = KSD^t \tag{1}$$

La matriz S es una matriz diagonal $r * r$ de valores singulares (ordenados de mayor a menor) donde $r = min(t, N)$ es el rango de la matriz M.

Si conservamos sólamente los s mayores valores singulares de S y sus correspondientes columnas en K y D. La matriz resultante es la matriz de rango r que mayor aproxima a la matriz original M usando como métrica la norma de *Frobenius*. Esta matriz esta dada por

$$M_s = K_s S_s D_s^t \tag{2}$$

donde s, $s < r$, es la dimensión del espacio de conceptos reducidos. La selección de un valor para s se realiza para balancear 2 efectos opuestos. Primero, s debe ser suficientemente grande para representar todas las propiedades de los datos reales. Y segundo, s debe ser suficientemente pequeño para permitir el filtrado de detalles irrelevantes en la representación.

En el modelo LSI implementado, se utiliza la librería $scikit - learn$ para realizar la reducción de dimensionalidad de la matriz M. La dimensionalidad del espacio de conceptos reducidos esta dado por

$$s = min(100, t, N) \tag{3}$$

debido a que esta librería recomienda a sus usuarios utilizar $s = 100$ cuando se trabaja en análisis de sémantica latente.

Para rankear los documentos con respecto a una consulta, simplemente se modela la consulta como un *pseudo-documento* en la matriz término-documento M. Asumiendo que la consulta es el documento número 0, entonces la primera fila de la matriz $M^t M$ proporciona los valores de similaridad de los documentos con respecto a la consulta.

Además, la aplicación provee una implementacíon del modelo clásico de espacio vectorial para que el usuario pueda realizar comparaciones de los resultados obtenidos por los dos modelos.

Creación de Índices

Para mejorar la eficiencia del SRI cuando el usuario haga varias consultas sobre el mismo directorio, se realiza la creación de índice cuando el usuario escoge un directorio. Los índices son almacenados solamente en la memoria RAM, pero la implementación del sistema está pensada para que este sea extensible y pueda ser fácilmente modificado para guardar los índices en el sistema de ficheros local.

En el modelo LSI, se almacena en memoria la matriz de pesos tf-idf calculado para la colección de documentos,que es usado posteriormente para agilizar la generación de la matriz M, y además se almacena el vocabulario y la tabla de valores idf, para poder transformar la consulta en un *pseudo-documento* y agregarlo a la matriz tf-idf guardada.

Evaluación

Para realizar la evaluación del modelo LSI, se escogieron 30 consultas previamente *tagueadas* sobre un corpus de 1033 documentos; el resultado de las consultas viene dado por la lista de los documentos con coeficiente de similaridad con respecto a la consulta mayor que 0,5, ordenado de más similar a menos similar.

A continuación los promedios de precisión, recobrado, f-medida y r-precisión obtenidos, usando un R=10:

1. Precisión Promedio = 0.22
2. Recobrado Promedio = 0.77
3. F-medida Promedio = 4.83
4. F1-medida Promedio = 3.10
5. R-Precisión Promedio = 0.71