Departamento de
Lenguajes y Computación

Universidad de Almería

Práctica 2
Informática Documental

Profesor: Antonio Becerra Terón

Aula de Prácticas: Aula 4, CITE III

Enero, 2008

PRÁCTICA 3. Implementación de un sistema de recuperación espacio-vectorial II. Proceso de consulta y generación de resultados — 6 horas

Objetivo general de la práctica

OBJETIVO

El objetivo básico de esta práctica es implementar el proceso de consulta en el modelo de recuperación espacio-vectorial, permitiendo definir y ordenar según un ranking los documentos más relevantes con respecto a una consulta.

Descripción de la práctica

DESCRIPCIÓN

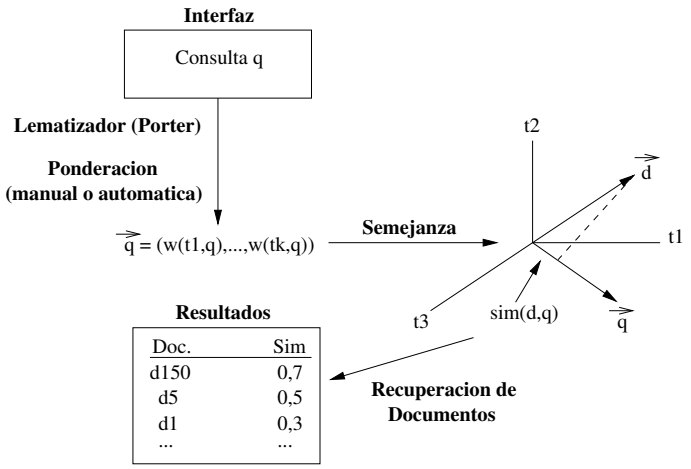
El proceso de consulta implica dos tareas fundamentales:

(a) Traducir la consulta a un vector de términos de indización

(b) Recuperar los documentos con mayor semejanza con respecto al documento de la consulta

La figura 1 muestra, de forma gráfica, el contenido de esta última práctica. A continuación, tal y como hemos realizado en la práctica 2, desglosamos cada una de estas tareas:

Figura 1: Proceso de la práctica 2 de Informática Documental



(a) En la traducción de la consulta, pediremos al alumno que implemente una interfaz sencilla que permita al usuario expresar consultas en el sistema de recuperacin de información.

Estas consultas estarán expresadas en lenguaje natural, y pueden verse como un documento cuyo contenido son las palabras de la consulta. Este apunte es muy importante porque nos permitir utilizar el programa desarrollado en la práctica 2 para eliminar las palabras vacías y lematizar el contenido de la consulta. Una vez lematizada la consulta, debemos plantear el proceso de ponderación de la consulta, que se puede implementar considerando dos tipos de perfiles de usuario.

- Usuario normal: en este caso, la ponderación se realiza de forma automática utilizando la misma fórmula que la ponderación del fondo documental de la práctica 1;
- Usuario avanzado: ahora, permitiremos la introducción manual de los pesos, normalizados entre 0 y 1, de forma que el usuario experto podrá indicar qué términos tienen mayor importancia dentro de la propia consulta;

Por último, generaremos el correspondiente vector asociado a la consulta de la forma

$$q \rightarrow \overrightarrow{q} = (w(t_1, q), \ldots, w(t_k, q))$$

donde k es el nmero de términos de indización, y el peso $w(t_i, q)$ indica la relevancia, manual o automática, del término de indización t_i en la consulta q. Un apunte más; lógicamente la consulta no tiene por qué tener todos los términos de indización $t_1, \ldots, t_k$. La razón de generar el vector con dimensión k es realizar un proceso de normalizaciń de vectores, de forma que todos los vectores tengan el mismo módulo, y poder definir una relación de semejanza donde los términos estén equilibrados y son comparables.

(b) Una vez generado el vector que representa la consulta, también denominado ecuación de búsqueda, debemos implementar el proceso de consulta o recuperación de documentos, definiendo una función de semejanza. Esta función debe permitir recuperar aquellos documentos que sean relevantes a la consulta. En la recuperación de documentos relevantes, también interesa considerar dos tipos de perfiles de usuario:

- Usuario normal: el proceso de consulta recuperar los documentos relevantes ordenados por relevancia y agrupados en 5 documentos
- Usuario avanzado: en este caso, el usuario decide el umbral de semejanza que desea establecer en la recuperación; el usuario establece el umbral y el proceso de consulta debe recuperar todos los documentos que estén por encima del umbral de relevancia ordenados, lógicamente, por dicha relevancia

A continuación, especificamos, muy brevemente, el proceso de consulta que debe implementar el alumno para el modelo de recuperación espacio-vectorial. En este proceso, $F_{i,j}$ representa la frecuencia del término i en el

documento j e Idf_i la inversa de la frecuencia de documentos del término i.

> Los documentos de cada término deben almacenarse en orden decreciente segn F_i
> La idea es mantener un ranking de los R documentos d_i con mayor relevancia
> Comenzamos analizando el término de la consulta con mayor Idf
> (i.e. lista de direccionamiento más corta o término en muy pocos documentos)
> seleccionamos los R primeros documentos de su lista (en ese orden)
> si no coleccionamos R documentos, entonces
> seguimos con el segundo trmino de mayor Idf y así sucesivamente

Una vez implementado el proceso de consulta, el alumno debe testear el sistema planteando diferentes tipos de consultas y comprobando que la recuperación de documentos es relevante con respecto a la consulta planteada. Además, debe comprobar que los documentos recuperados estén ordenados según la relevancia de los mismos.

Tenemos, por tanto, que el alumno debe implementar el sistema de recuperación permitiendo al usuario expresar una consulta en lenguaje natural, tratar esta consulta como un vector de documentos, y recuperar documentos relevantes a la consulta expresada como vector. La salida debe ser una lista de documentos ordenados por relevancia con respecto a la consulta.

Recursos:

RECURSOS

- Recursos software.
  - Compilador de Borland C++
  - Java Development Kit