Evaluación del Módulo 5.2

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Marzo 2020

1. Defina la tarea de Recuperación de Información (2 puntos).

La Recuperación de Información es la tarea que trata de satisfacer las necesidades de información de los usuarios, es decir, ante una consulta de un usuario, un sistema de recuperación de información busca en una colección de documentos aquel documento o fragmento de texto existente en una base de datos documental, que resuelve la consulta planteada por el usuario.

Las necesidades de información de los usuarios, que se representan a través de consultas en lenguaje natural, se resuelven con los sistemas de RI, que seleccionan los documentos o textos que contienen estos términos de las consultas, para ayudar y permitir a los usuarios a acceder a gran cantidad de información textual disponible en Internet.

Ejemplos de sistemas de recuperación de información son los buscadores de Google o Bing.

El objetivo de los sistemas de RI es determinar si los términos seleccionados que contienen los documentos o textos son relevantes, y deben incorporar el concepto de normalización u orden de relevancia. Y su funcionamiento es:

- Se parte de una colección de documentos para representar otros documentos o fragmentos de texto y consultas (lenguaje de representación).
- Un usuario tiene una necesidad de información y plantea una consulta al sistema de RI para procesar documentos (indexación) y consultas.
- El sistema de RI devuelve los documentos relevantes, que satisfacen la necesidad de información del usuario, seguramente normalizando los documentos para obtener una aproximación de la relevancia de los estos en la consulta (calculo de relevancia o similitud).

2. Describa el modelo espacio vectorial en Recuperación de Información (2 puntos).

El Modelo Espacio Vectorial es el modelo de RI más utilizado por su eficiencia y facilidad de implementación, con un enfoque fundamentado matemáticamente.

Características del modelo:

- Representa los documentos y las consultas de los usuarios como vectores ponderados n-dimensionales.
- Utiliza técnicas de comparación de vectores: coseno.
- Se basa en álgebra vectorial.
- Permite el ajuste parcial y el ranking.
- Representación de expresión en lenguaje natural como vector de pesos de términos.

- Considera las frecuencias de aparición de la palabra local (tf) y global (idf)
- Proporciona resultados ordenados.
- Trabaja bastante bien en la práctica a pesar de debilidades obvias.
- Permite la implementación eficiente para grandes colecciones de documentos.

3. Defina las medidas de evaluación Precisión y Recall en el contexto de la Recuperación de Información (2 puntos).

La medidas de evaluación

Recall proporcion entre documentos relevantes recuperados y documentos relevantes

Precision proporcion entre documentos relevantes recuperados y documentos recuperados

La precisión es una medida para evaluar los sistemas de RI

La tasa de recuperación o cobertura, también llamada Recall, es otra medida para evaluar los sistemas de RI

Medidas de evaluación de un sistema de recuperación de la información.

Generalmente, recall y precision son inversamente proporcionales.

Se suele buscar un equilibrio entre ellas o primar la preferida por el usuario tipo

4. ¿Qué es Lucene? ¿Y Solr? (2 puntos).

Son herramientas para el desarrollo de sistemas de recuperación de información.

Existen diversas librerias que facilitan el desarrollo de motores de busqueda. Ejemplos de librerías son Lucence y Sorl

5. Tras tres lecciones sobre Recuperación de Información, si le dijeran que tiene que diseñar un buscador de documentos ¿cuál sería la arquitectura de su sistema? No se limite al sistema de Recuperación de Información, piense que al menos necesita de una fuente de documentos, y que debe presentar los resultados de la búsqueda. (2 puntos).