Recuperación de Información en Internet
Tema 3: Principios de Recuperación de Información
Mestrado Universitario *Língua e usos profesionais*

Miguel A. Alonso Jesús Vilares

Departamento de Computación
Facultad de Informática
Universidade da Coruña

Índice

¿Qué es la Recuperación de Información?

- **Recuperación de información** (RI), *Information Retrieval*, es un área de la ciencia y la tecnología que trata de la **representación**, **almacenamiento**, **organización** y **acceso** a elementos de información
- Idealmente, un proceso de RI produce como salida un conjunto de documentos cuyo contenido satisface la necesidad de información de un usuario
- Problemas:
  - La necesidad de información del usuario debe ser expresada en forma de una consulta en lenguaje natural
  - No es fácil concretar en un texto los pensamientos o ideas que han dado lugar a la necesidad de información
  - Cada persona elegirá una terminología y unas construcciones diferentes para construir la consulta, determinadas por su uso habitual del lenguaje y su familiaridad con el ámbito de búsqueda de documentos

Sistema de Recuperación de Información

- Un sistema de RI trata de determinar el grado de semejanza de cada uno de los documentos disponibles con la consulta creada por el usuario. Asunto clave: discernir entre documentos relevantes y no relevantes para la consulta
- Realizan una **indexación** previa de los documentos, extrayendo los términos que representan mejor su contenido.
- La salida es una lista con los documentos relevantes, ordenada por el grado de relevancia.

Sistema de Recuperación de Información

- Problemas que debe enfrentar un sistema de RI:
  - Ambigüedad del lenguaje natural
  - La forma de expresar conceptos utilizada en la consulta puede no corresponderse con la utilizada en todos o parte de los documentos

- Caso peor: la existencia de un conjunto pequeño de documentos con información relevante pero expresada de forma distinta a como aparece en la consulta, a la vez que existe otro conjunto más numeroso de documentos con información no relevante pero que contienen todos o parte de los términos empleados en la consulta.

Otras tareas de RI

Aparte de la recuperación **ad-hoc**, existen otras tareas

- Categorización o clasificación de cocumentos
- Routing de documentos
- Clustering de documentos
- Segmentación de documentos
- ...

Medidas de rendimiento

Conceptos previos:

- D, conjunto de **documentos**
- R, conjunto de **documentos relevantes**
- $\bar{R} = D - R$, conjunto de **documentos no relevantes**
- A, conjunto de **documentos recuperados**
- $A \cap R$, conjunto de **documentos relevantes recuperados**
- $|X|$, **tamaño** del conjunto X

Precisión y cobertura

- La **precisión**, *Precision*, mide la porción de documentos recuperados que son relevantes

$$precision = \frac{A \cap R}{|A|}$$

- La **cobertura**, *recall*, mide la porción de documentos relevantes que son recuperados

$$recall = \frac{|A \cap R|}{|R|}$$

Medida-F y Fall-out

- La Medida-F, *F-measure*, combina la preción y la cobertura según un parámetro $\alpha \in (0, +\infty)$

$$F_\alpha = \frac{(1+\alpha) \times recall \times precision}{recall + \alpha\ precision)}$$

$$F_1 = \frac{2 \times recall \times precision}{recall + precision)}$$

- El **fall-out**, la porción de documentos no relevantes que son recuperados

$$fallout = \frac{\left|A \cap \bar{R}\right|}{\left|\bar{R}\right|}$$

Medidas que dependen de la ordenación de los documentos devueltos

- **Precisión a los *n* documentos recuperados**, mide la precisión obtenida cuando se han recuprado los 4, 10, 30, 100, 500 primeros documentos
- **R-precisión**, mide la precisión obtenida cuando se ha recuperado un número de documentos igual al número de documentos relevantes
- **Precisión media no interpolada**, calcula la media de las precisiones de todos los puntos en que se encuentran documentos relevantes
- **Precisión media interpolada en 11 puntos**, calcula la media de las precisiones en los puntos en que se alcanza el 0%, 10%, 20%, . . . , 100% de los documentos relevantes

Ejemplo

	Ranking 1	Ranking 2	Ranking 3
	d1	d10	d6
	d2	d9	d1
	d3	d8	d2
	d4	d7	d7
	d5	d6	d8
	d6	d5	d3
	d7	d4	d4
	d8	d3	d5
	d9	d2	d9
	d10	d1	d10
precisión	0,5	0,5	0,5
precisión R	1	0	0,4
precisión no interpolada	1	0,3544	0,5726
precisión interpolada 11 pt.	1	0,5	0,6440

Modelos de Recuperación de Información

Un modelo define:

- La manera en que se representan las consultas
- La manera en que se representan los documentos
- La forma en que se realiza el emparejamiento de consultas y documentos

Modelo Booleano

- Un término sólo puede tener dos estados:
  - verdadero ("aparece")
  - falso ("no aparece")
- Representa las consultas como una expresión booleana de términos
- Representa documentos como el conjunto de términos que aparecen en ellos
- Un documento es relevante si al evaluar la consulta sobre el documento, se obtiene el valor "verdadero"

- Problema: es difícil hacer un ranking con sólo dos valores

Modelo vectorial

- Representa consultas y documentos como un **vector** (una lista) de términos
- Cada término tiene un peso de acuerdo a:
  - Al número de veces que aparece en la consulta/documento
  - A lo "raro" que es, i.e., el número de documentos en los que aparece
- El emparejameinto de realiza "midiendo" cuán lejos está el vector de la consulta del vector de cada uno de lso documentos
- La medida más utilizada es el ángulo entre vectores (realmente el **coseno** del ángulo entre vectores, una medida entre 1 (coincidencia total) y 0 (ninguna coincidencia, vectores ortogonales)
- Es fácil realizar un ranking de acuerdo a las medidas obtenidas

Otros modelos

- Modelos probabilísticos
- Modelos basados en redes de inferencia
- Modelos basados en lógica difusa
- Modelos basados en semántica latente
- ...

Modelos para la web: PageRank

- Los modelos para IR "pura" no tienen en cuenta la estructura de hipertexto de la red
- Un buen modelo de IR para web debe tene en cuenta:
  - La estructura de las páginas wed
  - El texto de los hiperenlaces (que se asocia a la página de destino)
  - La persistencia en el tiempo de una página
  - La **popularidad** de una página web
- Una manera de medir la popularidad es contar el número de enlaces que apuntan a una página web. Se trata de un concepto relacionado con la medida de la importancia de una publicación según las **citas** que obtiene de otras publicaciones.
- El algoritmo **PageRank** de Google permite calcular la popularidad de una página web.

PageRank de Google

- Desarrollado en la Universidad de Stanford for Larry Page y Sergei Brin, fue la base del éxito de Google
- Es un algoritmo de análisis de grafos que asigna a cada página web un número en función de su "popularidad"
- Un enlace la página B a la página A se interpreta como un "voto" de la página B a la A
  - La página con más votos en la más popular o importante
  - Un voto desde una página muy popular vale más que un voto desde una página poco popular: es un algoritmo retroalimentado
- Matemáticamente, el pagerank de una página es un valor del autovector principal de la matriz de adyacencia de la web...

PageRank de Google

- . . . lo que se traduce en la fórmula

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B_1)}{L(B_1)} + \frac{PR(B_2)}{L(B_2)} + \ldots + \frac{PR(B_m)}{L(B_m)}\right)$$

donde

  - A es la página cuyo pagerank vamos a calcular
  - B_i son las páginas que contienen enlaces a A
  - $PR(X)$ es el pagerank de la página web X
  - $L(X)$ es el número de enlaces diferentes de la página X
  - d es el probabilidad de que una persona siga los enlaces de cualquier página (alrededor de 0,85), por lo que $1-d$ es la probabilidad de que salte arbitrariamente a cualquier otra página
  - N es el número total de páginas web

- Esta fórmula se recalcula iterativamente hasta que los valores tienden a estabilizarse