UNIVERSIDAD COMPLUTENSE MADRID

Recuperación de información (introducción)

Sistemas de Gestión de Datos y de la Información
Enrique Martín - emartinm@ucm.es
Máster en Ingeniería Informática
Fac. Informática

Introducción

- Según Gerard Salton, la recuperación de información *(information retrieval)* es:

“la disciplina que se preocupa por la estructura, el análisis, la organización, almacenamiento, búsqueda y recuperación de la información”

Introducción

- Tradicionalmente siempre ha tratado con **documentos de texto**:
  - Páginas web: Google, Yahoo!, Bing.
  - E-mails: búsqueda en Thunderbird o en Outlook.
  - Artículos de investigación: catálogos como ACM Digital Library o SpringerLink.
  - Libros: catálogos de biblioteca.
  - Noticias de prensa.

Introducción

- Un documento de texto tiene algunos atributos concretos (p.ej. autor, fecha, título) pero lo **principal** está en su **contenido.**
- Esto hace que todo el manejo que se hace sea diferente a lo que se realiza con una base de datos tradicional.

Introducción

- Base de datos tradicional, p.ej. cuentas en un banco. Las consultas son del tipo:
  - Datos de la cuenta 1234-5678-90-009042987.
  - Cuentas con más de 10.000€ de saldo.
  - Cuentas creadas en el mes de enero.
- Sistema de recuperación de información: buscador Web. Consultas del tipo:
  - World Cup Qatar 2022
  - “Type inference” “declarative programming”
  - Constitución España 1812

Introducción

- Además de documentos de texto, la recuperación de información puede utilizarse para tratar otros documentos: imágenes, vídeo, audio, ficheros escaneados...
- En estos casos suele utilizarse descripciones textuales de los documentos:
  - Realizadas por el autor: *keywords*
  - Contenidas en el propio documento: subtítulos
  - Obtenidas a partir del documento: reconocedores de voz y caras, OCR

Aspectos importantes

- **Relevancia**: los resultados que se obtienen tienen que tener importancia con respecto a la consulta realizada.
- Hay dos tipos de relevancia:
  - Relevancia temática
  - Relevancia de usuario

Aspectos importantes

- Relevancia **temática**: los resultados que se obtienen tratan sobre el mismo **tema** que la consulta.
- P.ej: al buscar “World Cup 2022”, un documento sobre cómo cocinar tortilla de patata no tiene relevancia temática

Aspectos importantes

- Relevancia de **usuario**: ¿los resultados obtenidos aportan algo al usuario concreto?
  - El usuario puede conocer de antemano un documento
  - El documento es demasiado antiguo
  - El documento está en un registro lingüístico que no entiende bien

Aspectos importantes

- Para resolver el asunto de la relevancia temática se utilizan los **modelos de recuperación** (*retrieval models*).
- Un modelo de recuperación formaliza el proceso de encajar un documento y una consulta.
- Son la base de las funciones de *ranking*, que sirven para ordenar los resultados.

Aspectos importantes

- Otro aspecto importante son las **necesidades de información del usuario**.
- Las búsquedas deben devolver resultados relevantes incluso **aunque la consulta no haya sido muy buena**.
  - Sugerencias de otras consultas (“Quizá quiso decir...”)
  - Expansión de consultas con términos relacionados

Bibliografía

Bibliografía

- **Information Retrieval: Implementing and Evaluating Searching Engines.** Stefan Bütcher, Charles L. A. Clarke, Gordon V. Cormak. The MIT Press (2016).
- **Modern Information Retrieval, the concepts and technology behind search**, second edition. *Ricardo Baeza-Yates y Berthier Ribeiro-Neto*. Pearson Education Limited (2011).

Bibliografía

- **Introduction to Information Retrieval.** *Christopher D. Manning, Prabhakar Raghavan y Hinrich Schütze*. Cambridge University Press. 2008. http://www-nlp.stanford.edu/IR-book/
- **Search Engines: Information Retrieval in Practice** (International Edition). *W. Bruce Croft, Donald Metzler y Trevor Strohman*. Person Education. 2010.