Gestión de Información en la Web
Máster en Ingeniería Informática
Parte II: Recuperación de información y Sistemas de recomendación

Tema 4: Recuperación de información
4.1. Introducción a la RI

Juan Manuel Fernández Luna
Dpto. Ciencias de la Computación e Inteligencia Artificial
jmfluna@decsai.ugr.es

OBJETIVOS

1. Entender el problema de la búsqueda de información.
2. Comprender el concepto de recuperación de información.
3. Conocer los elementos principales de la recuperación de información.

ÍNDICE

INTRODUCCIÓN

Analicemos la siguiente situación:

Una *persona* reconoce que su conocimiento es inadecuado para resolver un problema o alcanzar alguna meta.

Con objeto de resolver esa situación problemática, tiene que hacer uso de una *fuente de conocimiento* externa.

INTRODUCCIÓN

El usuario interactúa con la fuente de conocimiento a través de un intermediario.

Los tres componentes (usuario, fuente de conocimiento e intermediario) forman un *Sistema de Recuperación de Información* (S.R.I.).

El objetivo de un S.R.I. es que la situación problemática de un usuario se solvente.

Se busca facilitar la interacción efectiva del usuario con los objetos apropiados de información (elementos de la fuente de conocimiento).

INTRODUCCIÓN

Relevancia:

Indicador o medida de lo apropiado de un objeto de información con respecto a la situación problemática del usuario.

Dicho objeto trata sobre la misma materia que la situación problemática.

Y es útil para resolverla.

INTRODUCCIÓN

Objetivo de un S.R.I.:

Predecir, con un conocimiento previo sobre el usuario y la propia fuente de recursos, qué objetos son los más adecuados para que el usuario pueda resolver su problema.

Pasos:

1. Representar el problema de necesidad de información del usuario (consulta).
2. Representar y organizar el contenido de la fuente de conocimiento.
3. Comparar la consulta con los componentes del contenido.
4. Presentar los resultados al usuario para que interactúe o los juzgue.

INTRODUCCIÓN

¿Qué es **Recuperación Información**?

Estamos acostumbrados a hacer uso de herramientas de R.I.
Google, Yahoo, Bing!,...

Entradas:
Conjunto de documentos (texto en lenguaje natural).
Consulta de un usuario (también en lenguaje natural).

Salida:
Conjunto ordenado de documentos que satisfacen mejor la consulta.

Características deseables:

Útiles: Obtenemos documentos relevantes a nuestra consulta.
y eficiente: Los obtenemos en un corto periodo de tiempo.

ÍNDICE

EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN

Definiciones de Recuperación de Información:

(C.N. Mooers, 1959) "*el ámbito que comprende los aspectos intelectuales de la descripción de la información y su especificación para buscar, así como cualquier sistema, técnica o máquina que se emplee para desarrollar la aplicación*".

(G. Salton,1968) RI es la disciplina encargada de la representación, almacenamiento y organización de la información, y su posterior acceso y recuperación para responder a las necesidades de un usuario.

EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN

Definiciones de Recuperación de Información:

(Manning, Raghavan, 2008) “RI es la disciplina que trata de encontrar material (típicamente documentos) de una naturaleza desestructurada (típicamente texto) que satisface una necesidad de información en una colección grande (típicamente almacenada en ordenadores)”.
(Wikipedia) “RI es la ciencia que se encarga de la búsqueda de información en documentos, búsqueda de los mismos documentos, búsqueda de metadatos que describen un documento, o la búsqueda en base de datos de textos, sonidos, imágenes etc.”

Es un campo interdisciplinar, que cubre entre otras las áreas de

informática, documentación, lingüística, procesamiento de lenguaje natural, inteligencia artificial, …

EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN

Definiciones de Recuperación de Información:

(C.N. Mooers, 1959) "*el ámbito que comprende los aspectos intelectuales de la descripción de la información y su especificación para buscar, así como cualquier sistema, técnica o máquina que se emplee para desarrollar la aplicación*".

(G. Salton,1968) RI es la disciplina encargada de la representación, almacenamiento y organización de la información, y su posterior acceso y recuperación para responder a las necesidades de un usuario.

EL CONCEPTO DE RECUPERACIÓN DE INFORMACIÓN

SIMILITUD es el concepto básico en RI:
los documentos que utilizan vocabulario similar tienden a ser relevantes a las mismas consultas.

La RI está plagada de incertidumbre:

- En la representación del contenido de los documentos mediante términos o palabras clave (indexación).
- En la descripción de la necesidad de información por parte del usuario (expresada mediante una consulta).

La similaridad puede medirse de múltiples formas:

- Comparación de cadenas,
- Uso del mismo vocabulario,
- probabilidad de que el documento provenga de un mismo modelo,
- igual significado del texto.

ÍNDICE

RECUPERACIÓN DE INFORMACIÓN y BASES DE DATOS

De James Allan

	Bases de Datos	Recuperación de Información
Datos	Estructurados	No estructurados
Campos	Semántica clara (Teléfono, edad)	No hay campos (sólo texto)
Consultas	Definida (Algebra relacional, SQL)	Texto libre (lenguaje natural, booleano)
Recuperación	Crítica (control de concurrencia, operaciones atómicas)	Minimizar
Comparación	Exacta (resultados son siempre “correctos”)	Imprecisa (se debe considerar la efectividad)

ÍNDICE

MODELOS DE RECUPERCIÓN DE INFORMACIÓN

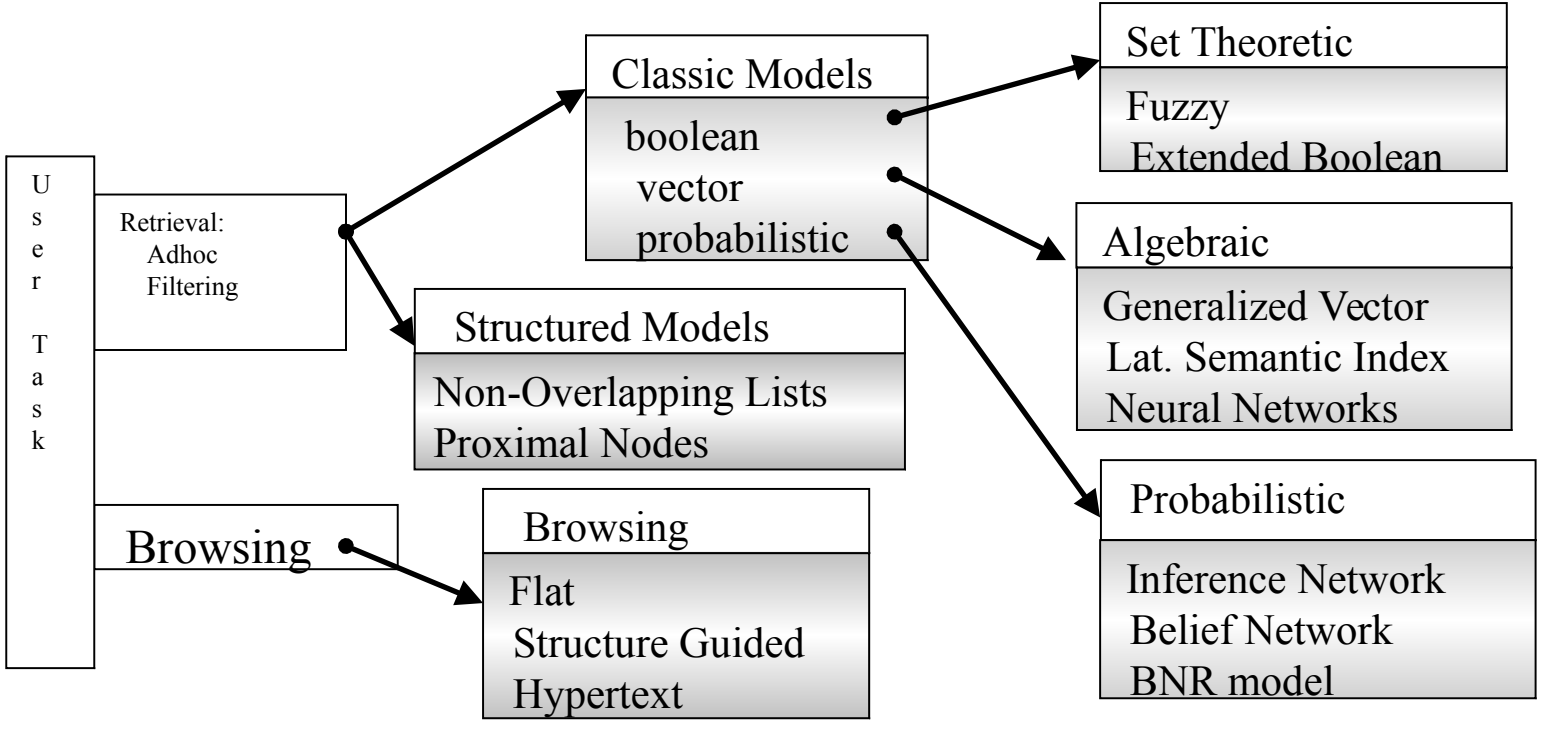
Un **Modelo de RI:**

Especificación sobre cómo representar
documentos y
consultas,
y cómo comparar unos con otras.

De Baeza-Yates

MODELOS DE RECUPERCIÓN DE INFORMACIÓN

De Baeza-Yates, Ribeiro

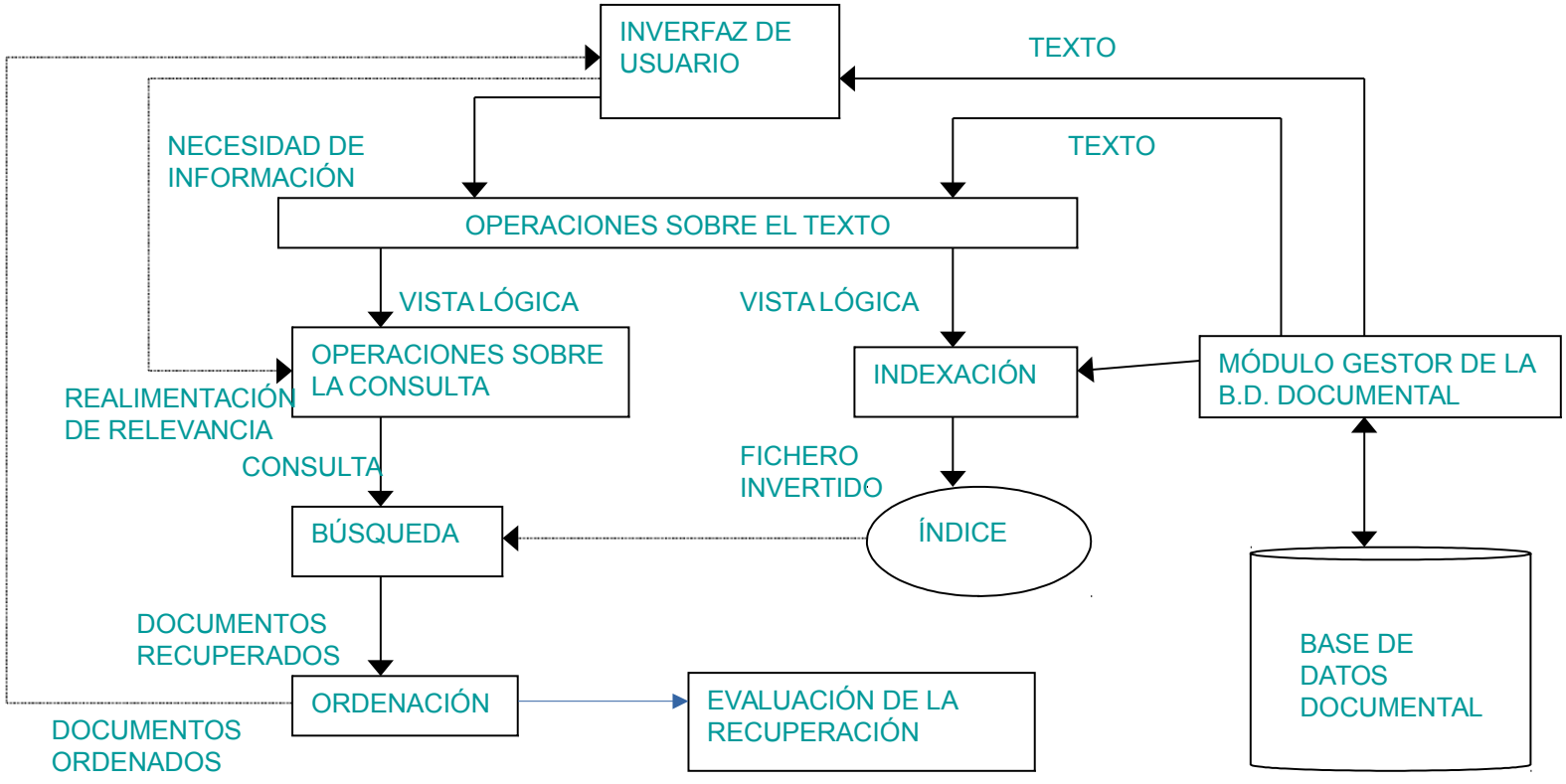


ÍNDICE

Sistema de Recuperación de Información (SRI)

El software que implementa un modelo de Recuperación de información.

SISTEMAS DE RECUPERCIÓN DE INFORMACIÓN



ÍNDICE

BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN

Décadas de los 60 y 70:

- Desarrollo de sistemas de recuperación para pequeñas colecciones (resúmenes de artículos científicos).
- Desarrollo de los modelos booleano y vectorial.
- Salton, en la Universidad de Cornell, es el líder de la investigación en esta época.

Década de los 80:

Se construyen sistemas para gestionar grandes colecciones documentales, fundamentalmente por compañías:

- Lexis-Nexis
- Dialog
- MEDLINE

BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN

Década de los 90:

Primeras búsquedas por Internet (FTP):

- Archie
- WAIS

Búsquedas en la WWW:

- Lycos
- Yahoo!
- Altavista

BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN

Década de los 90:

- Organización de pruebas:
- NIST TREC
- Sistemas de recomendación:
- Ringo
- Amazon
- NetPerceptions
- Clasificación automática y agrupamiento.

BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN

En el nuevo siglo:

- Análisis de enlaces en la Web:
- Google
- Extracción automática de información.
- Question Answering: TREC Q/A track.
- R.I. Multimedia:
  - Imagen,
  - vídeo, y
  - audio.
- R.I. Multilíngüe.
- Generación automática de resúmenes.
- Detección de novedad y de redundancia.
- Documentos estructurados
- ...

BREVE HISTORIA DE LA RECUPERACIÓN DE INFORMACIÓN

Dentro de la disciplina de la R.I. podemos encontrar tareas como:

- Question answering
- Agents (filtering, routing)
- Recommender systems
- Automatic organization (e.g., clustering)
- Cross-language retrieval
- Leveraging XML and other Metadata
- Data and information mining
- Knowledge management
- Meta-search (multi-database searching)
- Summarization
- Multimedia retrieval.
- …

ÍNDICE

EN ESTE TEMA...

1. Esta introducción.
2. Procesado e indexación de documentos.
3. Modelos de recuperación de información.
4. Evaluación de la RI.
5. RI en la Web.
6. Motores de búsqueda de código abierto.
7. Técnicas avanzadas de RI.

Referencias

Para diseñar los materiales de este tema, he hecho uso de material desarrollado por expertos en el área:

C.D. Manning, P. Raghavan, H. Schuütze. Introduction to Information Retrival. http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

F. Cacheda, J.M. Fernández Luna, J.F. Huete. Recuperación de Información. Un enfoque práctico y multidisciplinar. Ra-Ma. 2011. http://www.ra-ma.es/libros/RECUPERACION-DE-INFORMACION-UN-ENFOQUE-PRACTICO-Y-MULTIDISCIPLINAR/57515/978-84-9964-112-6