Information Retrieval Techniques and Applications

LIS 114

Introduction to Information Retrieval Systems

Introduction

Information Retrieval deals with representation, storage and access of information

1. It is concerned with the organization and retrieval of information from large database collections
2. IR is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge, discovery as a response to a user request (query)

Introduction

3. The process involves various stages returning relevant information to the user

Introduction

- The main goal of IRS is:
  - Finding relevant information or a document that satisfies user information needs

Introduction

To achieve this goal, IRSs usually implement the following processes:

- In indexing process the documents are represented in summarized content form
- In filtering process all the stop words and common words are removed
- Searching is the core process of IRS. There are various techniques that match with users need

Information Retrieval: Key Concepts and Components

- Information
- Information representation
- Information retrieval
- Database
- Search mechanism
- Browsing
- Language
- Interface
- Search strategy
- Retrieval performance

Information Retrieval

Synonymous with:

- Information access
- Information seeking
- Information searching
- Data mining
- Resource discovery

Information Representation

- Extraction of some elements from a document (keywords or phrases) or the assignment of terms (descriptors or subject headings) to a document so that its essence can be characterized and presented
- Can be done via abstracting, indexing, categorization, summarization, and extraction
- Information processing and information management are the synonyms

Database

- Record
- Field
  - Title, Author, etc.
- Sequential file
- Inverted file

Search Mechanism

- Information in a database can only be searched and retrieved when a corresponding search mechanism is provided
- Is defined by search algorithms and procedures
- Basic
- Advanced

Browsing

- Seeking and selecting information by skimming, scanning, and other similar activities
- Is needed when
  - Topic is not clear
  - Information is hard to be specified explicitly
  - Obtaining an overview of information
  - Discovering new information
  - Differentiating between relevant and irrelevant items

Language

- Natural language
- Controlled vocabulary

Interface

- What the user sees, hears, and touches while interacting with a computer system
- The quality of an interface is decided by
- Interaction mode (menu selection) and
- Display features (screen layout and font type)

Search Strategies

The approaches people take when conducting searches.

1. Identification of the information need of the user and the keywords that express the main concepts of the information need.
2. Identification and selection of the most suitable database.
3. Selection and execution of the query (basic or advanced search).
4. Review of the references in the database results page.
5. Occasionally, refinement of the query and obtaining new keywords to repeat the process.

Retrieval Performance

- Recall = $\frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in the collection}} \times 100$

- Precision = $\frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} \times 100$

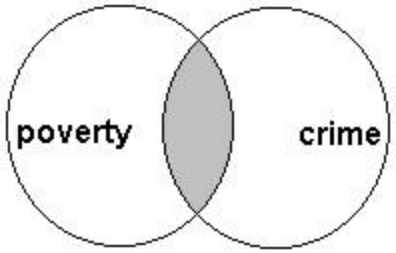
Retrieval Techniques

- Basic
  - Boolean operators
  - Phrase searching
  - Truncation / Wildcard searching
  - Proximity searching
  - Focusing / Limiting a search
- Advanced
  - Fuzzy searching
  - Weighted searching
  - Query expansion
  - Multiple database searching

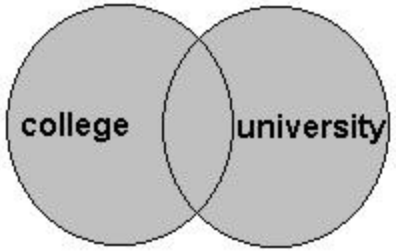
Boolean Operators

Boolean operators allow you to join terms together, widen a search or exclude terms from your search results. This means you can be more precise in locating your information.

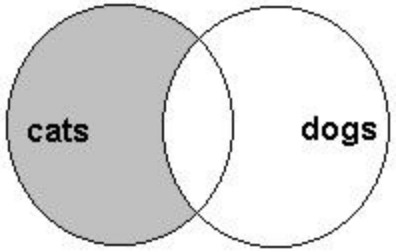
AND



OR



NOT



Phrase Searching

It narrows your search down by searching for an exact phrase or sentence. It is particularly useful when searching for a title or a quotation. Usually quotation marks are used to connect the words together.

For example

“Towards a healthier Scotland”

Truncation / Wildcard

These search techniques retrieve information on similar words by replacing part of the word with a symbol usually a * or ?. However, different databases use different symbols, so check what is used.

- In truncation the end of the word is replaced.
  - For example physiother* will retrieve physiotherapy, physiotherapeutic, physiotherapist and so on.
- In wildcard searching, letters from inside the word are replaced.
  - For example wom*n will retrieve the terms woman and women.

Proximity Searching

It looks for documents where two or more separately matching term occurrences are within a specified distance, where distance is the number of intermediate words or characters

- For example
  - Term A NEAR Term B
  - Term A ADJ Term B

Focusing / Limiting a Search

There are many ways to focus your search and all search tools offer different ways of doing this. Some of the ways of limiting your search are as follows:

- Date
- Language
- Place
- Publication type
- Age groups
- Type of material e.g. you could just need to find case studies

Fuzzy Searching

- It is designed to find terms that are spelled incorrectly at data entry or query input
- For example, the term *computer* could be misspelled as *compter, compiter, comptuer,* or *compyter*
- “Did you mean” at Google is an example

Weighted Searching

- Puts different emphasis or significance on each term in a query
- Prerequisite is that weighting has been applied at indexing stage
- A weighting scale can be used
- For example
  - *filtering3* AND *controversy6*

Query Expansion

- Allows the end-user to improve retrieval performance by revising search queries based on results already retrieved

Multiple Database Searching

- Searching in more than one IR systems simultaneously