1985

جامعة محمد بوضياف - المسيلة
Université Mohamed Boudiaf - M'sila

Information Retrieval (IR)

By: **Dr. LOUNNAS Bilal**

Recall

Recall

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Also

> Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources

Terminology

- General: Information Retrieval, Information Need, Query, Retrieval Model,Retrieval Engine, Search Engine, Relevance, Relevance Feedback, Evalua-tion, Information Seeking, Human-Computer-Interaction, Browsing, Inter-faces, Ad-hoc Retrieval, Filtering
- Related: Document Management, Knowledge Engineering
- **Expert: term frequency, document frequency, inverse document frequency,vector-space model, probabilistic model, BM25, DFR, page rank, stemming, precision**

Objectives

Objectives of an IR

Information retrieval deals with the representation, storage, organization of, and access to information items

The general objective of an information retrieval system is to minimize the overhead of a user locating needed information.

IR and DB

IR and database system

Information retrieval

- Process stored documents.
- Search documents relevant to user queries.
- No standard of how queries should be.
- Query results are permissive to errors or inaccurate items.

Database system

- Normally no processing of data.
- Search records matching queries.
- Standard: SQL language.
- Query results should have 100% accuracy. Zero tolerant to errors.

IR and DM

IR and Data mining

Information retrieval

- User target: Existing relevant data entries.

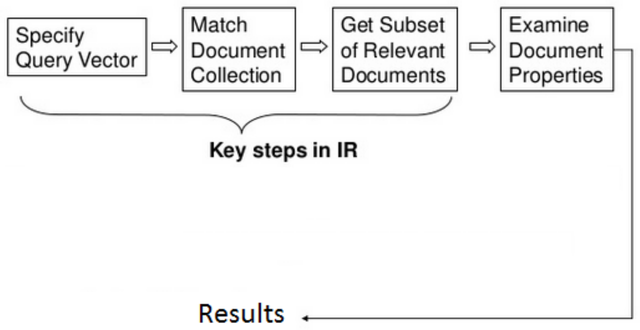
Data mining

- User target: Knowledge (rules, etc.) implied by data (not the individual data entries themselves).

Many techniques and models are shared and related. Example: classification of documents.

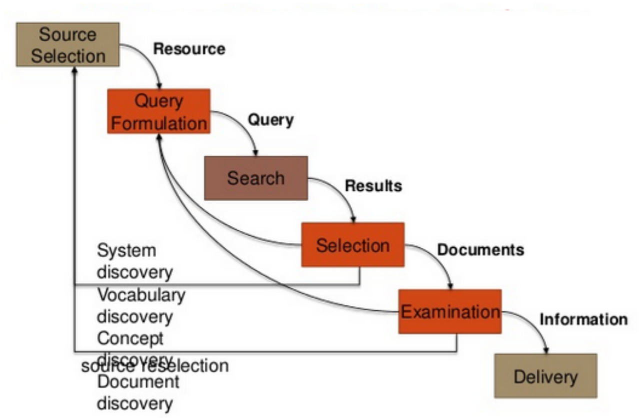
Keysteps in IR

Keysteps in information retrieval


Specify Query Vector
Match Document Collection
Get Subset of Relevant Documents
Examine Document Properties
Key steps in IR
Results

Cycle of IR

Information retrieval cycle



IR architecture

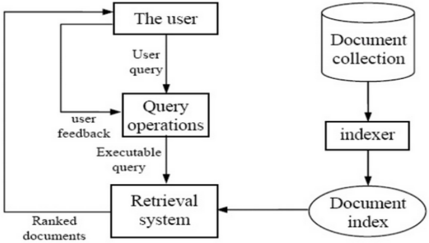
Architecture of IR system

The architecture of an information retrieval system can be presented in a very simple way, it just need to accumplish the primary goal of an IR system which is:

> Retrieve all the documents which are relevant to a user query, while retrieving as few non-relevant documents as possible.

IR architecture

Architecture of IR system

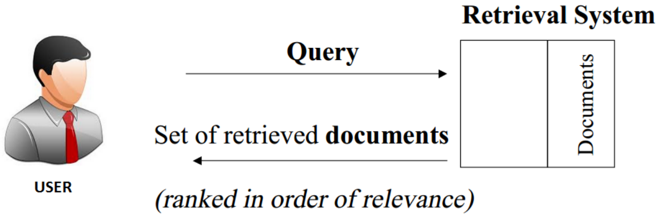


Elements of information retrieval architecture

- Processing of documents.
- Acceptance and processing of queries from users.
- Modelling, searching and ranking of documents.
- Presenting the search result

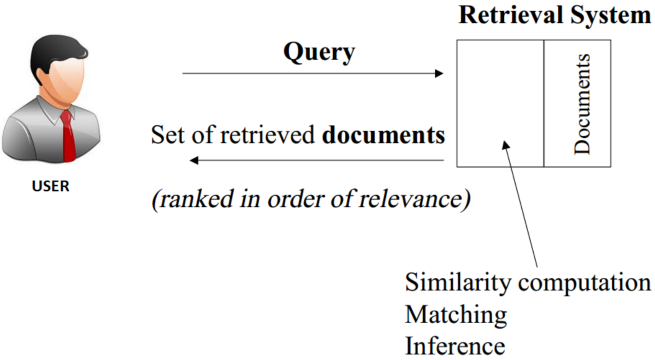
IR architecture

Top level view (User view)



IR architecture

Within the system level view (Model view)



IR models

What is a model

A model is an abstract representation of a process used to study properties, draw conclusions, make predictions. The quality of the conclusions depends upon how closely the model represents reality.

The IR model is a way to represent the contents of a document and a query, also to compare a document representation to a query representation, in order to produce relative documents based on this comparison (document ranking, score function).

IR models

Components of IR model

A retrieval model consists of the following items:

1. **D:** representation for documents.
2. **Q:** representation for queries.
3. **F:** a modeling framework for D, Q, and the relationships among them.
4. **R(q, di):** a ranking or similarity function which orders the documents with respect to a query.

Types of IR models

We have two type of IR models

Exact match

- Query specifies precise retrieval criteria
- Every document either matches or fails to match query
- Result is a set of documents
  - Usually in no particular order.
  - Often in chronological order.

Best match

- Query describes retrieval criteria for desired documents.
- Every document matches a query to some degree.
- Result is a ranked list of documents, best first.

Boolean model

Definition

The Boolean model of information retrieval (BIR) is a classical information retrieval model and, at the same time, the first and most adopted one. It is used by many IR systems to this day.

1. Simple model based on Boolean algebra
2. Intuitive concept
3. Precise semantics
4. Clear formal basis
5. Widely adopted by early information systems

Boolean model

Representation

The BIR is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query.

1. Use Boolean algebra and Set theory
2. Document are logical conjunction of terms.
3. Term weights are binary
   - $w_{i,j} \in \{0,1\}$
   - $w_{i,j} = 1$ - term present
   - $w_{i,j} = 0$ - term not present
4. Queries are Boolean expressions
5. Documents are considered **relevant** if the query evaluates to 1 (true)

Boolean model

Example

- Which plays of Shakespeare contain the words **Brutus** AND **Caesar** but **NOT Calpurnia**?
- Could **grep** all of Shakespeare's plays for **Brutus** and **Caesar** then strip out lines containing **Calpurnia**?
  1. Slow (for large corpora)
  2. NOT is hard to do
  3. Other operations (e.g., find the Romans NEAR countrymen) not feasible

Boolean model

Example

Term-document incidence

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

Boolean model

Example

Term-document incidence

1. So we have a 0/1 vector for each term.
2. To answer query: take the vectors for Brutus, Caesar and Calpurnia (complemented) and then we applied bitwise AND operation.

Boolean model

Example

Take the vectors for Brutus, Caesar and Calpurnia (complemented)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	**1**	**1**	**0**	**1**	**0**	**0**
Caesar	**1**	**1**	**0**	**1**	**1**	**1**
Calpurnia	**0**	**1**	**0**	**0**	**0**	**0**
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
Bitwise AND						

Boolean model

Example

Take the vectors for Brutus, Caesar and Calpurnia (complemented)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	**1**	**1**	**0**	**1**	**0**	**0**
Caesar	**1**	**1**	**0**	**1**	**1**	**1**
¬**Calpurnia**	**1**	**0**	**1**	**1**	**1**	**1**
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
Bitwise AND	1	0	0	1	0	0

1. Antony and Cleopatra
2. Hamlet

Boolean model

Advantages and Disadvantages

Advantages

1. Clean formalism
2. Easy to implement
3. Intuitive concept

Disadvantages

1. Exact matching may retrieve too few or too many documents.
2. Hard to translate a query into a Boolean expression
3. All terms are equally weighted