Information Retrieval (Part 1)

Fabio Aiolli

http://www.math.unipd.it/~aiolli

Dipartimento di Matematica Pura ed Applicata
Università di Padova

Anno Accademico 2007/2008

Bibliographic References

- Copies of slides
- Selected scientific papers
- Suggested Textbooks
  - C.D. Manning, P. Raghavan and H. Schütze, ***Introduction to Information Retrieval***, Cambridge University Press. 2007 *(Preliminary Draft available in the course Web page)*
  - Other interesting books in the course Web page

A definition of IR [Manning et.al. 2007]

Information retrieval (IR) is finding material (usually documents) of unstructured nature (usually text) that satisfy an information need from within large collections (usually on local computer servers or on the internet).

1. Finding material (documents) in *large collections*
2. *Unstructured nature* as opposed to structured data: data do not have a fixed semantic/structure
3. Satisfy an *information need*

1. Large document base

- CD-ROM and distributed technology have given rise to larger and larger document bases (e.g. from $O(10^6)$ to $O(10^9)$ documents). This is the size at which the most of IR problems arise;
- The document base can be static (e.g. CD-ROM) or dynamic (e.g. digital libraries, the Web), centralised or distributed;

2. Unstructured documents

Free-form expressions, usually having an information content (in DB terminology: semi-structured data)

- Scientific papers, letters, e-mails, newspaper articles, image captions, audio transcript -> Textual IR
- Images, graphics, audio (spoken or not spoken), video, … , stored in digital form -> Multimedia IR

3. Information Need

- A desire (possibly specified in an imprecise way) of information useful to the solution of the problem, or resources useful to a given goal;
- Useful (Relevant), according to the subjective opinion of the user.

Tipical tasks covered in IR

- Search ('ad hoc' retrieval)
  - Static document collection,
  - Dynamic queries
- Filtering
  - Static query,
  - Dynamic document feeds
- Categorization
- Clustering
- Collaborative Filtering or Recommendation
- Browsing
- Summarization
- Question Answering
- …

Media involved in IR

- Text
  - monolingual o multilingual text
  - Structured text (XML)
  - OCR->Text
  - Spoken text
- Hypertext
- Music
- Graphics
- Images
- Video / Animation
- …

The nature of IR

- Information Retrieval is difficult because of the indeterminancy of relevance:
  - The system might interpret differently from the user the meaning of the documents and/or the query, due to inherent ambiguities in natural language;
  - The user might not know exactly what she wants (vague or imprecise information need);
  - It is not clear what 'degree of relevance' the user is happy with (i.e. it might not be clear to the system whether the user is 'recall-oriented' or 'precision-oriented').

What IR is not!

- **Data Retrieval**, as in DBs
  Information need cannot directly be expressed as a simple query and documents have not a precise semantic. A translation of them into logical representations is needed.
  - In IR the set of objects to be retrieved are not clearly determined -> Slightly different retrieved sets should not be necessarily considered as a 'fatal' error of the system
  - User satisfaction is the issue of IR
- **Knowledge Retrieval**, as in AI
  - In AI a fact α is inferred from a knowledge base Γ of facts expressed in a certain formalism
- **Question Answering**
  - In QA a query an answer is returned generated from a semantic analysis of documents.
  - Huge amount of domain knowledge needed
- **Information Browsing**, as in Hypermedia
  - Relevant documents are retrieved by an active intervention of the user and not by a search routine
  - The goal of a browsing task is less clear in the mind of the user

Evolution of IR

- In the past, IR systems were used only by expert librarians as reference retrieval systems in batch modality.
  - Many libraries still use categorization hierarchies to classify their volumes
- The advent of novel computers and the Web have brought to
  - efficient indexes, capable to index and displaying entire documents
  - processing of user queries with high performance
  - ranking algorithms which improve the quality of the answers
  - methods to deal with multimedia
  - interaction with the user
  - methods to deal with distributed document collections (e.g. WWW)

A formal characterization

- An IR model can be defined by M=[D,Q,R] where
  - D is a representation for the documents in the collection
  - Q is a representation for the user information needs (queries)
  - $R(d_i,q_j)$ is a ranking function which associates a real number with a query $q_j \in Q$ and a document representation $d_i \in D$.

    N.B. It defines an ordering among the documents with regard to a given query q_j.

- Unlike query satisfaction in DBs, the relationship of *relevance* R between documents D and information needs Q is not formally defined, but is subjective, i.e. determined by the user. Therefore,
  - unlike in DBs, effectiveness is an issue
  - a degree of effectiveness (user satisfaction) can be defined
- In an IR system or model, it is necessary to choose whether to treat relevance as
  1. Boolean-valued $R : D \rightarrow Q \in \{0,1\}$
  2. Finite-valued $R : D \rightarrow Q \in \{1,..,N\}$
  3. Infinite-valued $R : D \rightarrow Q \in \mathbb{R}$
- Approaches 2. and 3. are definitely the most plausible, but approach 1. is the most popular, as it greatly simplifies both relevance feedback and experimentation.

Factors which influence the relevance

- Relevance is an ineffable notion, in particular it is:
  - *Subjective*: two users may pose the same query and give different judgments on the same retrieved document;
  - *Dynamic*:
    - A user may judge relevant a given document at a given retrieval pass, and later judge the same document as irrelevant, or viceversa;
    - The documents retrieved and displayed to the user may influence her relevance judgment on the documents that will be displayed to her later;
  - *Multifaceted*: it is determined not just by topicality (i.e. thematic affinity), but also by authoritativeness, credibility, specificity, exhaustivity, recency, clarity,…

- R is not known to the system prior to user judgment!
- The system may only 'take a guess' by computing a Retrieval Status Value $RSV(d_i,q_j)$ which depends on the model used e.g.
  - Logical consequence in Boolean Models
  - Similarity between vectors in Vector Space Models
  - Probability of relevance in Probabilistic Models
- The range of RSV should be a totally ordered set to allow for ranking the documents according to the scores.

Binary and Ranked Retrieval

- Binary Retrieval $RSV(d_i,q_j) \in \{0,1\}$
  - Does not allow the user to control the magnitude of the output. In fact, for a given query, the system may return
    - under-dimensioned output
    - over-dimensioned output
- Ranked Retrieval (ordering induced by $RSV(d_i,q_j) \in \mathbb{R}$)
  - Allows the user to start from the top of the ranked list and explore down until she sees fit. This caters for the need of different types of users, those that want just few highly relevant documents, and those that want many more.