Information Retrieval (Intro)

Fabio Aiolli

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

Dipartimento di Matematica Pura ed Applicata
Università di Padova

Anno Accademico 2009/2010

Class schedule

Monday, 11:30 - 13:30
Tuesday, 15:30 – 17:30
Thursday, 11:30 – 13:30

Office Hours:
Monday, 14:30 – 15:30 or by appointment

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. 2008 *(Draft available in the course Web page)*
  - Other interesting books in the course Web page http://www.math.unipd.it/~aiolli/corsi/0910/IR/IR.html

Examination Modalities

- Oral Exam
  - Info about dates will be given soon
- Project / Talk
  - A small project related to a course topic
  - A 40 min talk on a topic not (fully) covered in the course

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) and/or finds difficult to formalize it;
  - 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 (e.g. Dewey Decimal Classification DDC)
- 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 $R : D \times Q \rightarrow \mathbb{R}$ as
  1. Boolean-valued $R \in \{0,1\}$
  2. Finite-valued $R \in \{1,..,N\}$
  3. Infinite-valued $R \in \mathbb{R}$
- 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.

How a modern IR system operates

- (off-line)
  - building document representations and loading them into an internal index structure (ad-hoc retrieval)
  - Or building query representations and loading them into an internal index structure (filtering)
- (on-line) [ad-hoc case]
  - reading a query and returning the user documents that the system considers relevant to the information need expressed by this query
    - Binary-retrieval or Ranked-retrieval
  - Peraphs reading user's feedback on the documents retrieved and using it for performing an improved retrieval pass.

Representing textual documents and information needs

- String searching is inappropriate for IR (best suited for data retrieval)
  - Computationally hard
  - The occurrence of string x in a document d in neither necessary nor sufficient condition for relevance of d to an information need about x
- It is thus usual to produce an internal representation (IREP) of the 'meaning' of documents (off-line) and of queries (on-line), and to determine their match at retrieval time

Indexing

Indexing: The process by which IREPs of documents and queries are produced

- It typically generates a set of (possibly weighted) index terms (or features) as the IREPs of a document (or query)
- Underlying assumption: the meaning of this set well approximates the meaning of the document (or query)

Indexing in textual IR

- Index terms in textual IR can be:
  - Words (e.g. classification) automatically extracted
  - Stems of words (e.g. class-) automatically extracted
  - N-grams automatically extracted
  - Noun-phrases (e.g. classification if industrial processes) automatically extracted
  - Words (or noun phrases) from a controlled vocabulary (e.g. categorization)
  - …

Incidence Matrix

The result of the indexing process: the ***incidence matrix.***

	d_1	…	d_i	…	d_m
t_1	w_{11}	…	w_{1i}	…	w_{1m}
…	…	…	…	…	…
t_k	w_{k1}	…	w_{ki}	…	w_{km}
…	…	…	…	…	…
t_n	w_{n1}	…	w_{ni}	…	w_{nm}

N.B. W_{ki} can either be binary or real.

$T = \{t_1,..,t_n\}$ is the **dictionary** of the document base

Indexing - Weights

- Weights can be assigned
  1. **Manually** (typically binary weights are used) by trained human indexers or intermediaries who are familiar with
     - The discipline the documents deal with
     - The indexing technique (e.g. the optimum number of terms for an IREP, the controlled vocabulary,..)
     - The contents of the collection (e.g. topic distribution)
  2. **Automatically** (either binary or real weights are used): by indexing processes based on a statistical analysis of word occurrence in the documents, in the query and in the collection.
- Approach 2. is nowadays the only one left in *text retrieval* (cheaper and more effective).
- Approach 1. and 2. has been used in conjunction until recently because of the difficulty in producing effective automatic indexing techniques for non-textual media.

Indexing - considerations

- The use of the same indexing technique for documents and query alike tends to guarantee a correct matching process
- There is indeterminacy in the indexing process: different indexers (human or automatic) do not produce in general the same IREP for the same document!
- Unlike what happened in reference retrieval systems, the on-line availability of the entire document allows the use of the entire document also for indexing

Evaluation and Experimentation

- The evaluation of an IR technique or system is in general accomplished experimentally (instead of analytically)

  1. Effectiveness
     How well it separates/ranks docs w.r.t. the degree of relevance for the user
  2. Efficiency
     How long it takes for indexing, for searching,.. How expressive is the query language. How large is the doc base
  3. Utility
     Quality w.r.t. costs (of design, development, update, use, etc.) paid by involved parties (developers, users, etc.)
  4. Coverage (e.g. for Web search engines)
     The subset of the available docs that the search engine 'covers' (i.e. indexes, provides access to)

- Criteria 1. and 4. are widely used, 3. is hardly quantifiable.

Effectiveness for Binary Retrieval: Precision and Recall

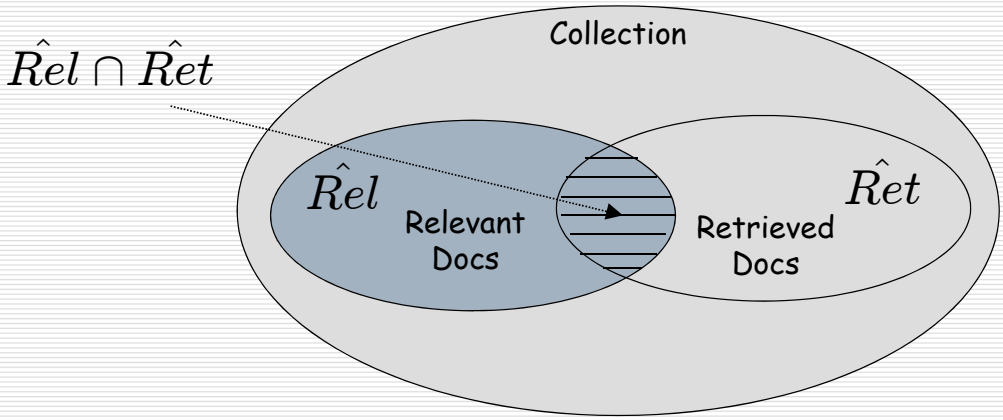
If relevance is assumed to be binary-valued, effectiveness is typically measured as a combination of

- Precision: the "degree of soundness" of the system

$$\pi = Pr(Rel|Ret) = \frac{|\hat{Rel} \cap \hat{Ret}|}{|\hat{Ret}|}$$

- Recall: the "degree of completeness" of the system

$$\rho = Pr(Ret|Rel) = \frac{|\hat{Rel} \cap \hat{Ret}|}{|\hat{Rel}|}$$



Contingency Table

	Relevant	Not Relevant
Retrieved	True positives (**tp**)	False positives (**fp**)
Not Retrieved	False negatives (**fn**)	True negatives (**tn**)

$$\pi = \frac{tp}{tp+fp} \qquad \rho = \frac{tp}{tp+fn}$$

Why NOT using the accuracy $\alpha = \frac{tp+tn}{tp+fp+tn+fn}$?

Evaluation for Ranked retrieval Precision at Recall

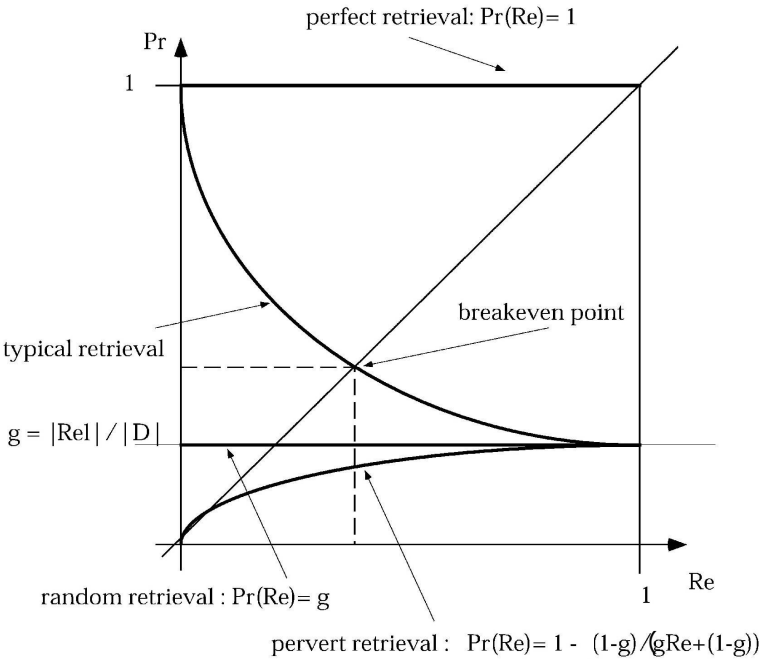
- In a ranked retrieval system, precision and recall are values relative to a rank position r.
- These systems can be evaluated by computing *precision as a function of recall*, i.e. $\pi(\rho)$
  - *What is the precision* $\pi(r(\rho))$ *at the first rank position* $r(\rho)$ *for which recall has a value of* ρ *?*
- We compute this function at each rank position in which a relevant document has been retrieved, and the resulting values are interpolated yielding a *precision/recall plot*
- A unique numerical value of the effectiveness can be obtained by computing e.g. the integral of precision as a function of recall

Collection D, |D| = 100, a query q with |Rel|=20

Rank Pos. q	Rel?	ρ	$\pi(\rho)$
1	Y	1/20=0.05	1/1=1.00
2	Y	2/20=0.10	2/2=1.00
3	N		
4	Y	3/20=0.15	3/4=0.75
5	N		
6	N		
7	Y	4/20=0.20	4/7=0.57
...	...	...	...

Note that

- The effectiveness of a system is typically evaluated by *averaging* over different queries (*macroaveraging*)
  - Different searchers are equally important
  - Partial view of a problem: different methods may work best for different types of queries
- A *typical* precision/recall plot
  - Is monotonically decreasing
  - For ρ=1 it takes the value π=g, where g=|Rel|/D is the *generality* (*frequency*) of the query
- When the document base is big, it is very important to have high precisions for small recall values.
  - Measures such as *precision at 10* (P@10) are often used in place of $\pi(\rho)$
- 'Typical' values are not beyond .4 precision at .4 recall



Precision and Recall Drawbacks

1. The proper estimation of maximum recall requires detailed knowledge of the entire (possibly very large) collection.
2. Precision and recall capture different aspects of the set of retrieved documents. A single measure would be more appropriate.
3. Do not fit the *interactive* retrieval process settings
4. Inadequate for systems which require weak orderings

F measure

- Precision-oriented users
  - Web surfers
- Recall-oriented users
  - Professional searchers, paralegals, intelligence analysts
- A measure that trades-off precision versus recall?

  F-measure (weighted harmonic mean of the precision and recall)

$$F = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho} \qquad F_{\beta=1} = \frac{2\pi\rho}{\pi+\rho}$$

β < 1 emphasizes precision!

Benchmark Collections

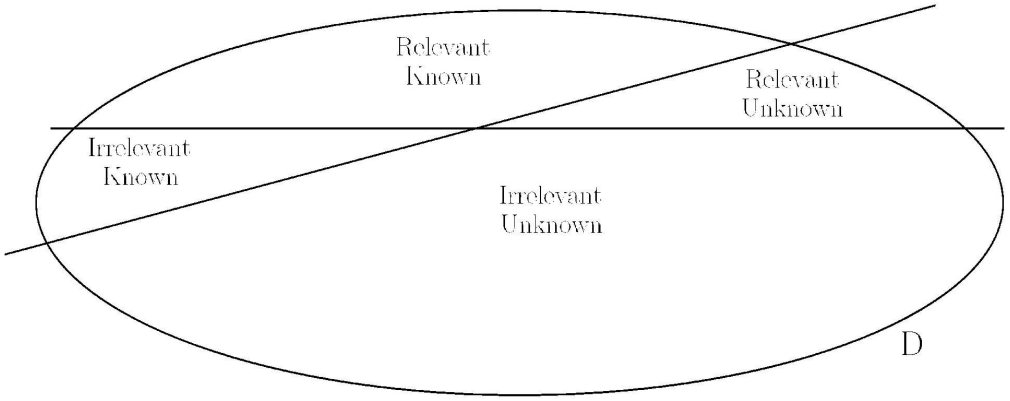
- A set of (up to $O(10^7)$) documents
  $D = \{d_1,\ldots,d_m\}$
- A set of queries (topics in TREC terminology)
  $Q = \{q_1,\ldots,q_l\}$
- A (typically binary) *relevance matrix* of size $m \times l$, who can be either the original query issuers (TREC), or (more often) domain experts.
- A test collection is an abstraction of an operational retrieval environment. It allows to test the relative benefits of different strategies in a controlled way

Limits of the 'scientific' approach

- Two limits
  - R*elevance* judgments tend to be *topicality* judgments
  - It does not consider other aspects such as the user interface and its fitness to the user-seeking behavior
- After all, their construction is the real problem

TREC collection

- TREC web site http://trec.nist.gov
- The corpus: 'Ad hoc' track in the first 8 TREC competitions between '92 and '99.
- Several millions documents and 450 information needs
- In TREC, as in many other big collections relevant documents are identified by a *data-pooling* method thus approximating the set of relevant documents from below



Other Benchmark Collections (used in Text Categorization)

Reuters-21578

- The most widely used in text categorization. It consists of newswire articles which are labeled with some number of topical classifications (zero or more). 9603 train + 3299 test documents

Reuters RCV1

- Newstories, larger than the previous (about 810K documents)

20 Newsgroups

- 18941 articles from the 20 Usenet newsgroups