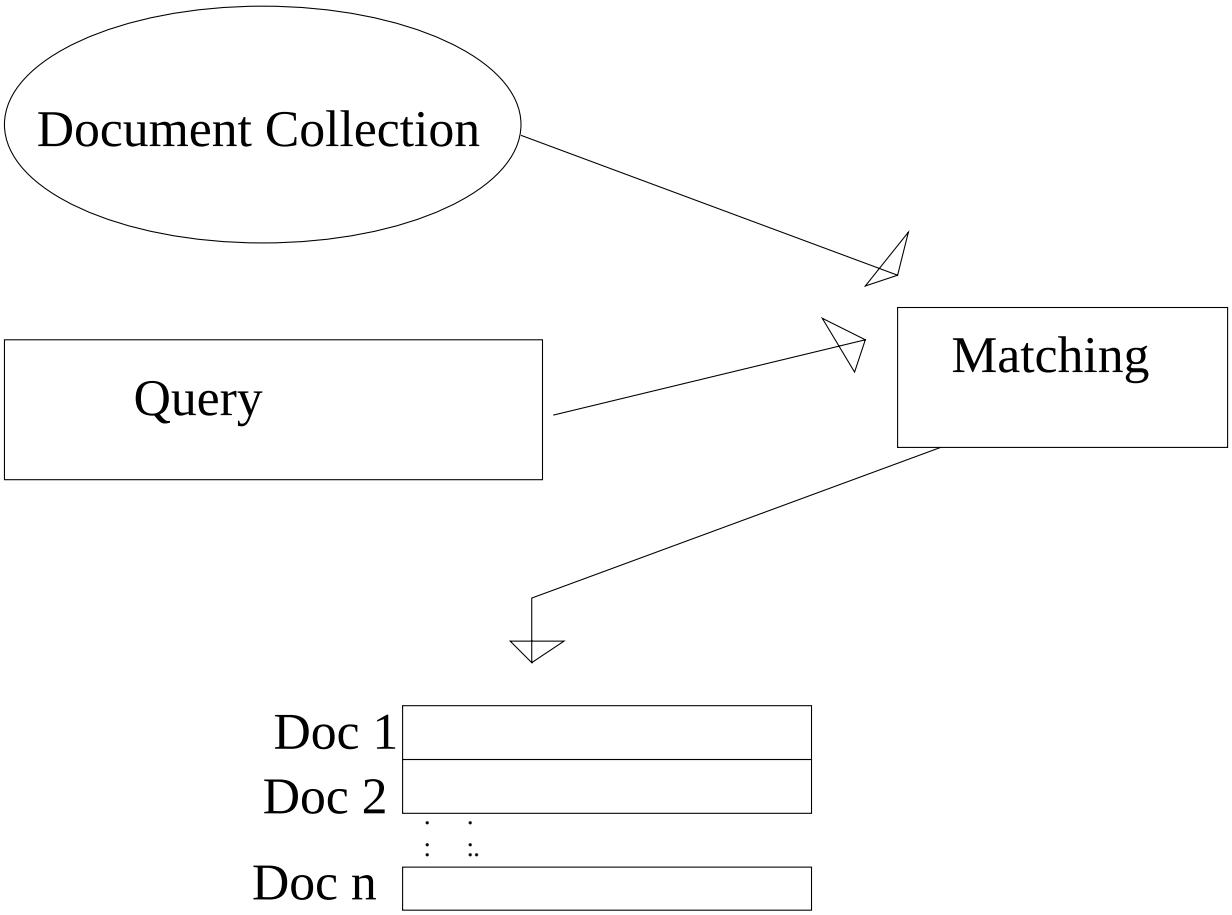
Information Retrieval

What is Information Retrieval (IR)?

An IR system matches a user's query against a document collection, returning a list of documents deemed relevant to the query.

- Query: expression of user's information requirement
  e.g. "have you got books on computational fluid dynamics?"
- Document: a newspaper article, web page, book title, journal paper abstract...
- Document Collection: an database of on-line representations of documents.
- Relevance: "aboutness"

System Archtecture



Retrieved Document List

IR Systems

Examples of Information Retrieval Systems:

- RUG on-line library catalogue
  `http://www.ub.rug.nl/`
- Web search engines: AltaVista, Infoseek, NorthernLight...
- CD-ROM abstract databases:
  INSPEC, MEDLINE

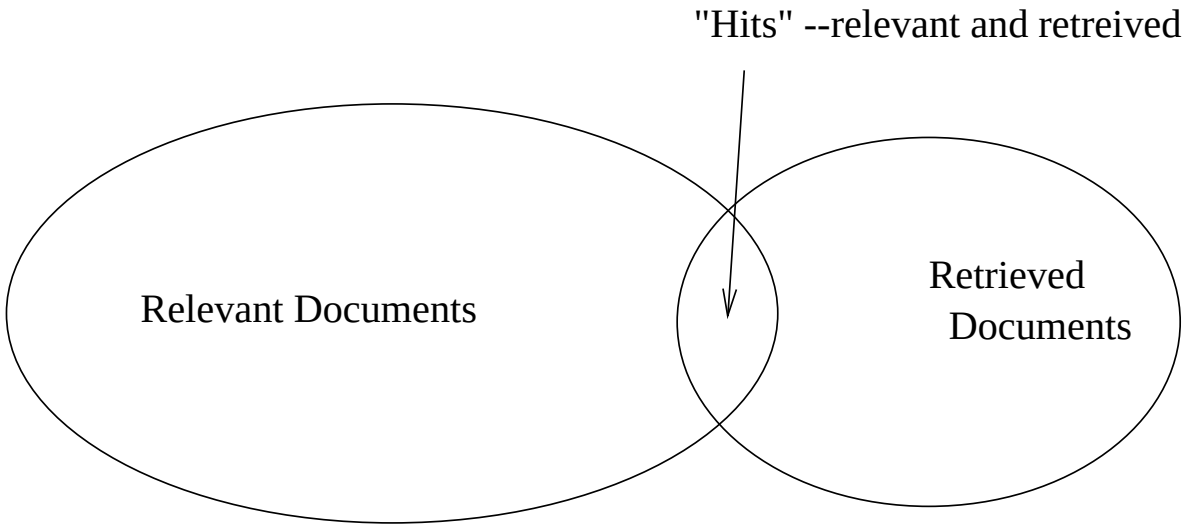
Relevance

- Core concept of IR
- Hard to define objectively

IR research is all about replicating the relevance judgement capabilities of a good librarian. Main issues are:

- Indexing: How to represent the document collection?
- Matching: How to compare query with the doc. collection?
- Performance: How do we know the system is working properly?

Performance



Recall proportion of relevant documents among those retrieved

Precision proportion of retrieved documents that are relevant

Given a Query which retrieves documents Retr

- Rel is the set of relevant documents
- Hits is the intersection of Retr and Rel
- Recall = |Hits|/|Rel|
- Precision = |Hits|/|Retr|

Indexing

Assign a list of keywords or index terms to each document.

Index term - Also known as a concept. A word or phrase indicative of the document content.

e.g. assign "linguistics" and "generative grammar" to Chomsky's *Aspects of the Theory of Syntax*.

Two main approaches:

- Controlled vocabulary systems
- Free text systems

The document is represented in the system as a set of index terms.

Controlled Vocabulary Systems

- Have a predefined vocabulary of allowed index terms e.g. Library of Congress subject headings
- Vocabulary usually built by human experts
- Documents are indexed by choosing most appropriate set of terms from the vocabulary.
- Must analyse the document content and classify it against the vocabulary
- Index terms are usually in a hierarchy

  music $\supset$ contemporary $\supset$ rock $\supset$ punk $\supset$ Offspring

Controlled Vocabulary

- difficult to automate, but attempts have been made
  e.g. NASA's controlled dictionary system

Example 1 Library of Congress subject headings

- It maintains a subject classification hierarchy
- Classifies all new books under relevant part of subject tree

Example2 Yahoo is a manually built internet catalogue - "good" pages manually selected and added by human expert

Example 3 RUG On-line Catalogue

Free Text IR - WWW

- Web search engines use free text methods to manage their indexes
- No restriction on vocabulary: Index terms extracted from documents
- Indexing is done automatically:
  - Turn document into list of words eliminating punctuation, and common function words like "a" and "the" (*stopwords*).
  - Remove all suffixes (*stemming*), or transform all words to canonical form (*lemmatisation*).
  - Select "best" terms from remaining list, assigning an appropriate weight to each

Indexing Example

"Yesterday, after speaking to the senate and the house, President Clinton said he would not be resigning."

- List of words: *yesterday after speaking to the senate and the house president clinton said he would not be resigning*
- Stopword removal: *yesterday speaking senate house president clinton said resigning*
- Stemming: *yesterday speak senat hous presid clinton said resign*

  OR
- Lemmatisation: *yesterday speak senate house president clinton say resign*

Issues in Term Selection and Weighting

Now we must choose the "best" index terms from the derived list and weight them.

- Consider relative importance of term in document.
  More occurences of term $\Rightarrow$ more indicative of content.
  this is the measure *tf* ("term frequency").
- Consider occurence of term in collection as a whole. How well does it discriminate between this document and the rest of the collection?
- It will discriminate well if it occurs more frequently per unit length in this document than in all other documents:
  this is the measure *idf* ("inverse document frequency")

Weighting Terms

- Terms that occur rarely discriminate well but retrieve relatively few documents
- Terms that occur often improve recall but at the expense of precision
- "middle" terms are best - give these the highest weights
- Should we eliminate "bad" terms?
  No - all terms kept but "bad" ones given low weights
- Most common weighting function is *tf.idf*.

Matching

Obtain unordered list of words from user and extract index terms from it

e.g. "machine translation europe EU" yields index terms *europe*, *eu*, *machine translation*

Look for documents which assign a high weight to these terms.

Return a ranked list of matching documents.

Implementation: Vector Space Model, Latent Semantic Indexing, Probabilistic Retrieval, Connectionist Approaches...

The Vector Space Model

- Imagine document collection as an n-dimensional space, one dimension per index term
- Represent document as an n-dimensional vector D.
  $d_i = w_i$ if term $i \in D$ $d_i = 0$ otherwise
- Represent query as an n-dimensional vector Q.
  $q_i = 1$ if term $i \in Q$ $q_i = 0$ otherwise
- Calculate similarity between query vector and all document vectors and return list of top-scoring documents

Toy Example

Example: The Cat Sat on the Mat

Terms = *cat*, *sat*, *mat*, *today*, *yesterday*

Document	Vector
a) the cat sat on the mat	[11100]
b) the cat sat yesterday	[11001]
c) the cat sat	[11000]
d) the cat sat on the mat today	[11110]
Query	**Vector**
cat, *mat*, *today*	[10110]

Matching

Doc	Doc Vector	Query Vector	Score
a	[11100]	[10110]	2
b	[11001]	[10110]	1
c	[11000]	[10110]	1
d	[11110]	[10110]	3

ranking: d, a, b, c

Problems with Free Text IR

- Keyword Ambiguity

  e.g. bank ⇒ financial institution, data repository, aeroplane manoeuvre, land beside river...

  – How do we know which sense of bank a use is interested in?
  – Current research on word sense disambiguation (WSD) relies on contextual clues
  – There are no contextual clues in an IR query
  – WSD for IR is an active area of research
  – Only solution for now is to enter lots of index terms in the query

Problems, cont.

- Polysemy

  e.g. "association football", "soccer", "football', "British football" $\Rightarrow$ FIFA football

  - User entering query on "soccer" will not retrieve references to "football"
  - Query can be automaticaly expanded using a thesaurus (e.g. Wordnet)
  - Thesaurus construction requires human intevention
  - Adding many terms can result in retrieval of many irrelevant documents

Moral of the Story:
Intelligent results require intelligent queries