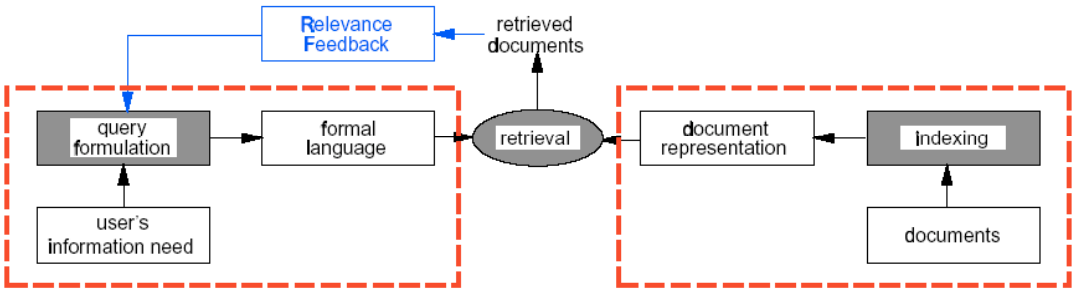
CSA4020

Multimedia Systems:
Adaptive Hypermedia Systems

Lecture 4:
Automatic Indexing
&
Performance Evaluation

Automatic Indexing

- Document Retrieval Model (reminder)



- Indexing Methods

  Before creating an index, you've got to think about why you're creating it

  Garbage in, garbage out

  You cannot retrieve a document using features that you haven't indexed

Objective indexing vs. non-objective

Does your data have a structure?

Do you only ever want to retrieve data based on “database” attributes?

Is your data unstructured?

Do you want to retrieve data based on document contents?

Manual indexing vs. automatic

"she was the apple of my eye"…what has it got to do with fruit?

Automatic indexing means that it is fast and cheap, but may include inaccuracies

Manual indexing, means it is likely to be slow and expensive, but more accurate

Which is more important: speed, cost, or accuracy?

Controlled vocabulary vs uncontrolled

Convert all terms in docs and query into representative term (or concept)

An ontology will give consistency, but you may lose precision (car and truck may both be converted to vehicle – is it what you want?)

Uncontrolled vocabulary is more flexible, but users may not find relevant docs (car and automobile will be treated as different terms!)

Possible to use other techniques resembling controlled vocab, even though uncontrolled vocab is used

Single term indexing vs complex term

"Computer science" can either be indexed as "computer" and "science" or as a complex term

"Computer science" means more than "computer" + "science"

"I like my computer. Science is wonderful."

But it is more expensive

Usually need knowledge base for complex terms, and they are domain specific

Not the same as phrase indexing… a complex term is part of the vocab – a phrase is not

Specificity vs Exhaustivity

- Specificity and controlled vocabularies are at odds with each other: index specifically the terms which are encountered

- Exhaustivity: do we index everything, or just words we are particularly interested in (apart from stopwords), e.g., the first 100 words in a document

Problems with manual indexing

- high labour costs

- inconsistency between trained indexers

  thesauri created by 2 indexers in given subject domain: only 60% of terms in common!

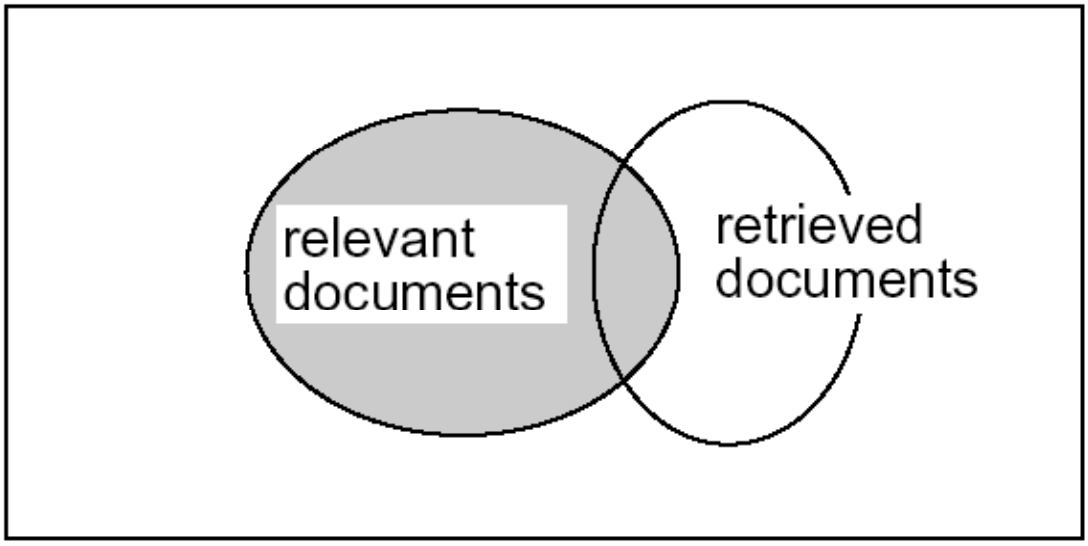
  indices obtained by two indexers from same document with the same thesaurus: 30% of terms in common!

  two users searching same collection with same question: only 40% of documents in common!

  Relevance judgements obtained by two users on same set of documents and same topic: only 60% of documents in common!

Performance Evaluation

- Precision & Recall



$$\text{recall} = \frac{\text{Number of relevant docs retrieved}}{\text{total number of relevant documents}}$$

$$\text{precision} = \frac{\text{Number of relevant docs retrieved}}{\text{total number of docs retrieved}}$$

- But what happens if there are no relevant documents in the collection?
- What happens if no relevant documents are retrieved?
- What happens if all documents are retrieved?
- What happens if all documents in the collection are relevant to the query?

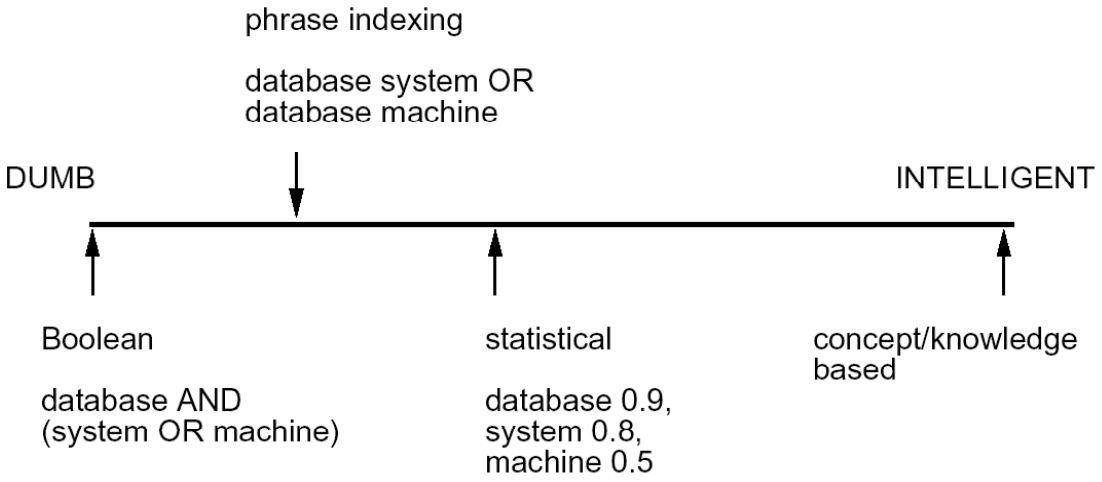
- Fallout

$$\text{fallout} = \frac{\text{no. of nonrelevant docs retrieved}}{\text{total no. of nonrel docs}}$$

Hopefully, there will be some non-relevant documents in the collection!

"Good" system should have high recall and low fallout…

Spectrum of indexing methods



Benchmarking

- It is notoriously difficult to provide evidence that an IR system "works"

- Consider recall and precision… anything vaguely odd about them?

- Who decides which documents in a collection are relevant and non-relevant? (That's the whole point of IR, isn't it?)

  If IR systems could be proven to do it successfully for all domains, the IR problem would be solved

- Problem is, even humans cannot always agree on the relevance of a document to a query!

- Text REtrieval Conference (TREC) provides test collections: collections of documents, queries, and relevance judgements per document/query pair

- Retrieval effectiveness of a system is evaluated on these collections

- Collections available from http://trec.nist.gov