Relevance Feedback and Query Expansion

Dan Smith

University of East Anglia

What is relevance feedback?

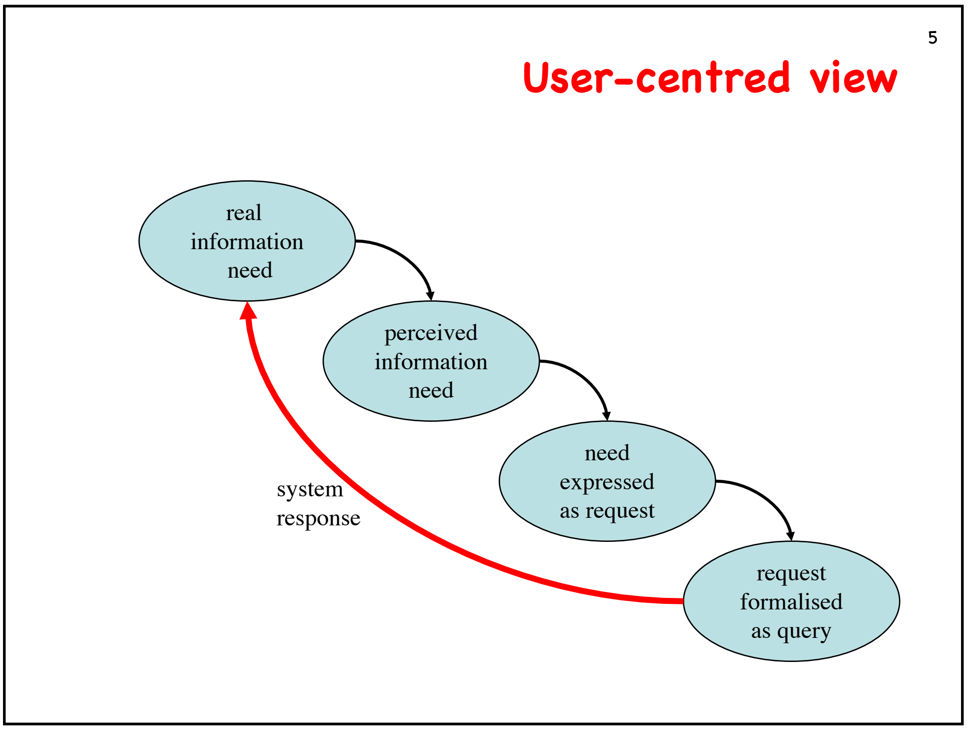
- Relevance feedback is concerned with using information from the results of a search to modify the query or result set to improve its relevance to the user's information need

Query characteristics

- Search engine users follow specific trends with their searches
  - 55% of Google queries > 3 words
  - 7% of Google queries have spelling errors or typos
  - 20% of each day's queries are new or not seen within the last 3 months
- In general
  - Users prefer broad search terms
  - Users do not like to expand their queries either through refining search terms or using Boolean operators
  - Most users only look at the first page of results
- Google handles 3.5bn queries/day

various sources: https://adwords.googleblog.com 14Aug2014, http://www.internetlivestats.com

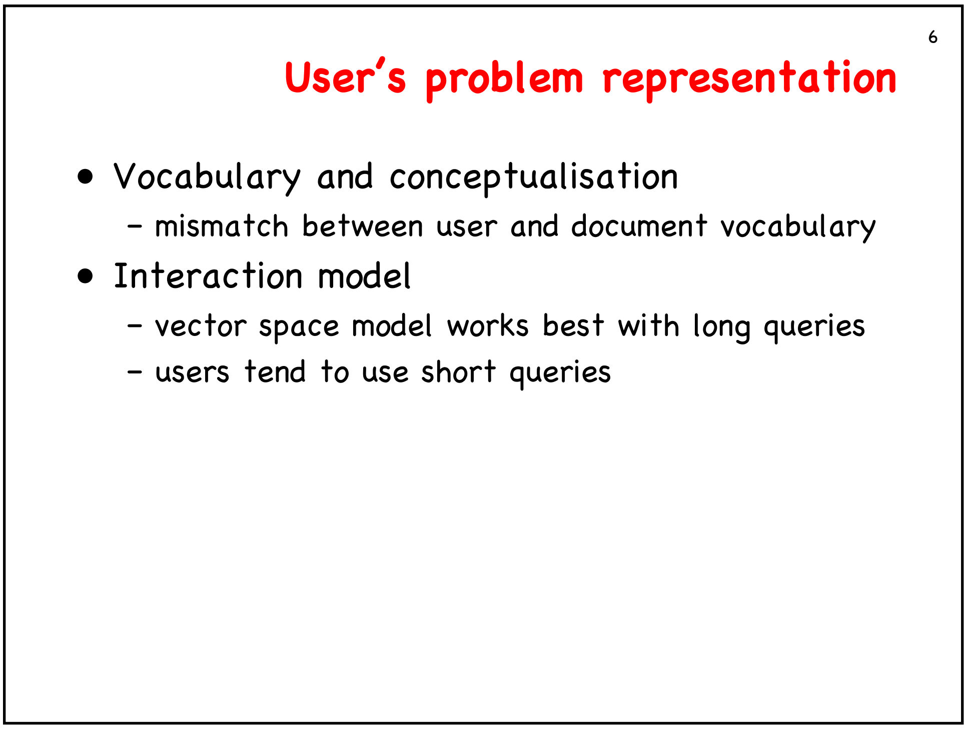
RELEVANCE



5

User-centred view

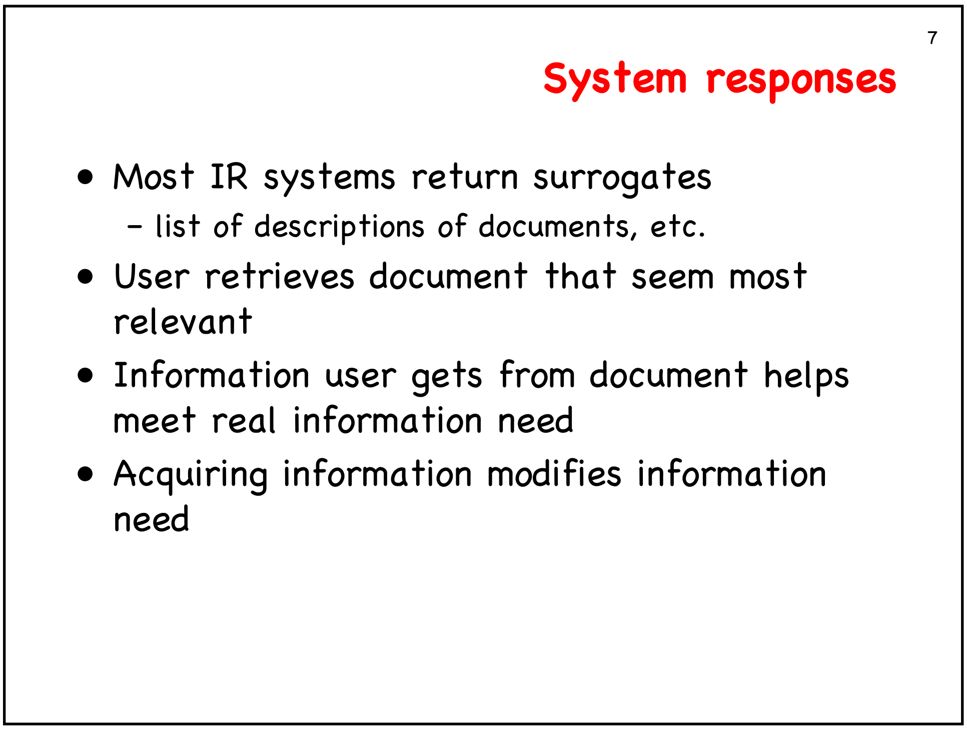
real information need
perceived information need
need expressed as request
request formalised as query
system response



6

User's problem representation

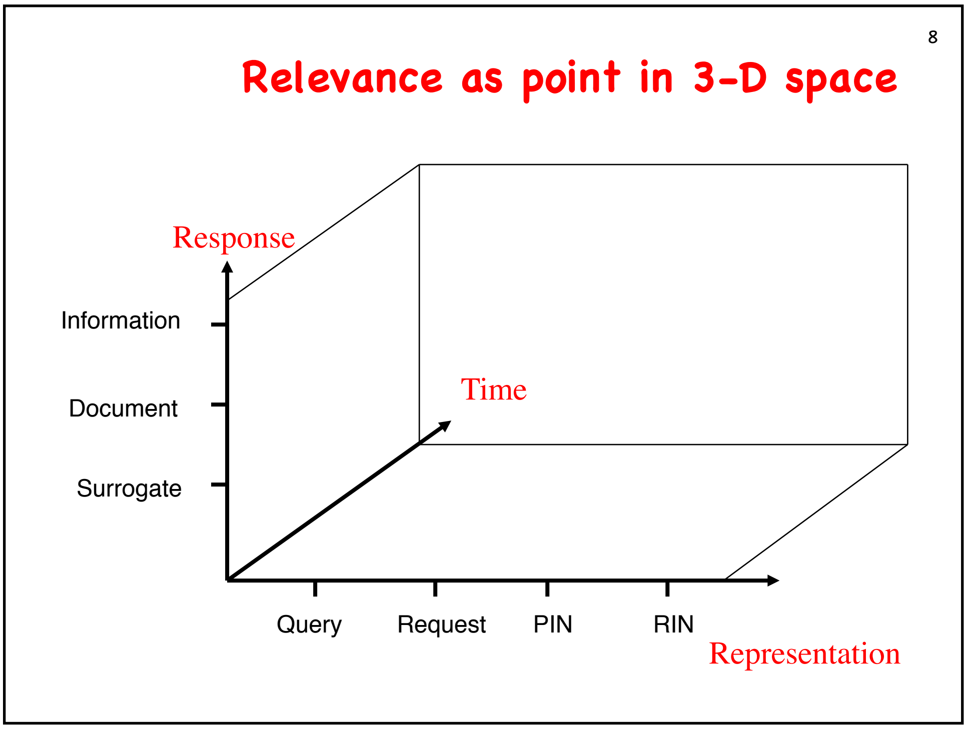
- Vocabulary and conceptualisation
  - mismatch between user and document vocabulary
- Interaction model
  - vector space model works best with long queries
  - users tend to use short queries



7

System responses

- Most IR systems return surrogates
  - list of descriptions of documents, etc.
- User retrieves document that seem most relevant
- Information user gets from document helps meet real information need
- Acquiring information modifies information need



8

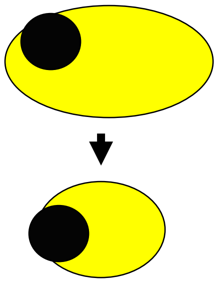
Relevance as point in 3-D space

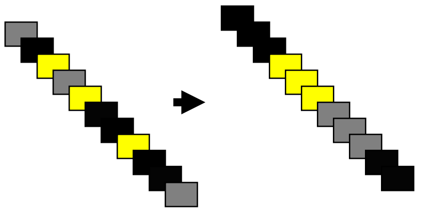
Context: the missing dimension

- Physical
  - Location
  - Co-location
  - Communications channels
- Intellectual
  - Task
  - Prior knowledge and understanding
  - Time, etc. constraints

Relevance feedback

- Assume we return a superset of the relevant documents
- Aim is to either
  - reduce set: discard documents unlike those labeled relevant
  - order set: heavily weight documents like those labeled relevant





11

Getting feedback from a user

- Ask user to rate each document (surrogate) retrieved
  - difficult, as user doesn't know (yet)
- Assume relevance is proportional to time spent viewing
  - can be good for static content-bearing pages (but if the user goes for a coffee...)
- Assume the most relevant documents are downloaded
  - works well if all documents are PDF format or similar
    (e.g. ACM Digital Library)

12

Relevance from group opinions

- Basic idea
  "most people who like *X* also like *Y*"
- Often in recommender systems
  - e.g. Amazon
  - "People who bought this also bought..."
- Can build association lists and show users the most frequent

Linden G., Smith B., and York J. (2003). Amazon.com Recommendations, *IEEE Internet Computing*, Jan.-Feb., 76-80

Konstan J. A., Miller B. N., Maltz D., Herlocker J. L., Gordon L. R., Riedl J. (1997) GroupLens: applying collaborative filtering to Usenet news. *CACM* 40(3) 77-87. DOI= http://doi.acm.org/10.1145/245108.245126

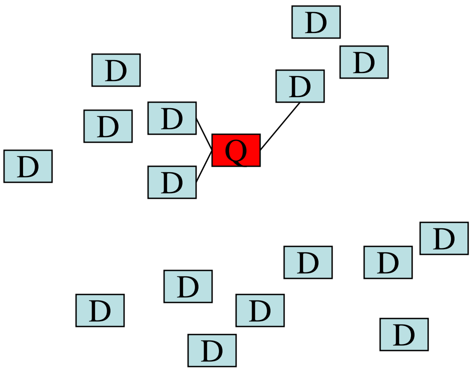
13

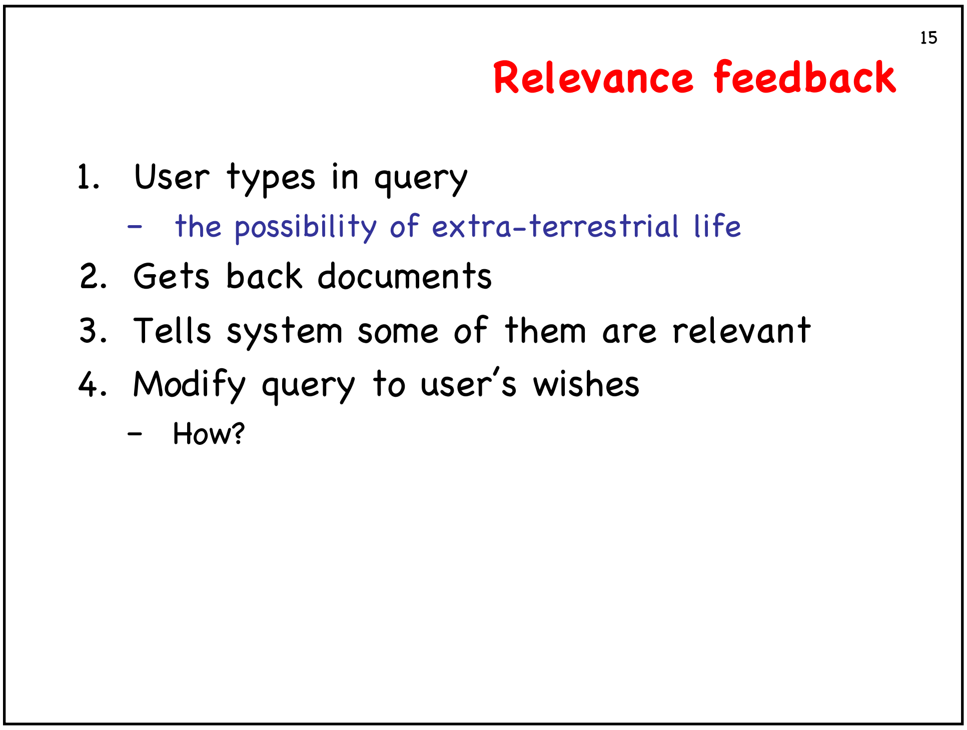
System-centred relevance

- Documents are *about* a topic
- Need to classify (and retrieve) documents according to their topic
- Many approaches to classification
  - sum of weighted terms
  - relative frequency of terms occurring in known relevant documents
  - latent semantic indexing
  - NLP-based
  - …

14

System-centred view

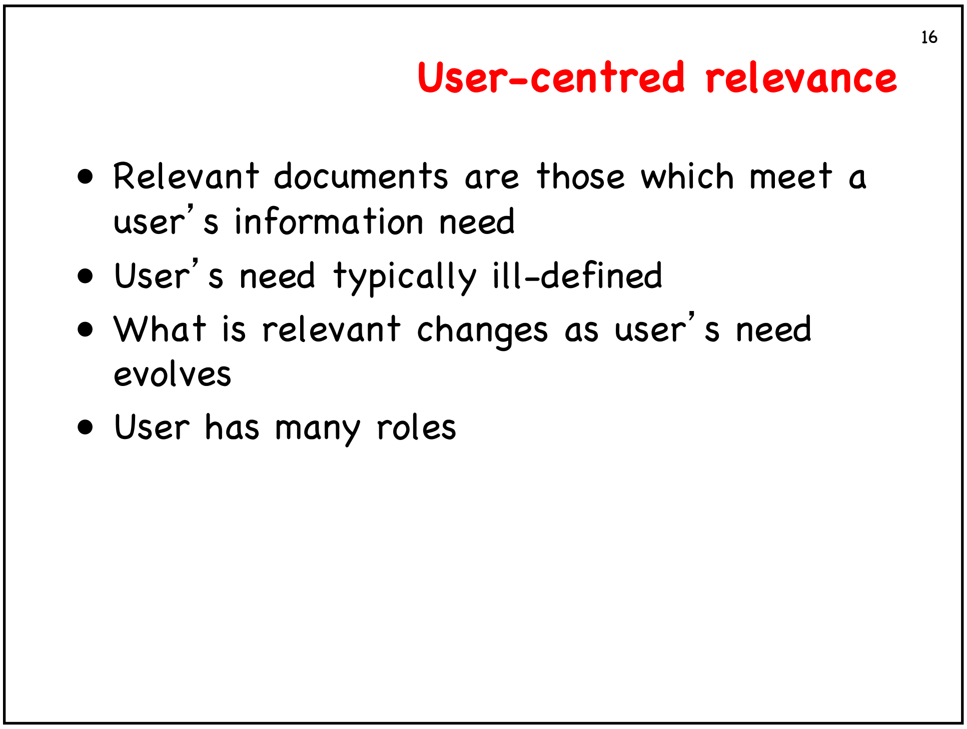




15

Relevance feedback

1. User types in query
   - the possibility of extra-terrestrial life
2. Gets back documents
3. Tells system some of them are relevant
4. Modify query to user's wishes
   - How?



16

User-centred relevance

- Relevant documents are those which meet a user' s information need
- User' s need typically ill-defined
- What is relevant changes as user' s need evolves
- User has many roles

17

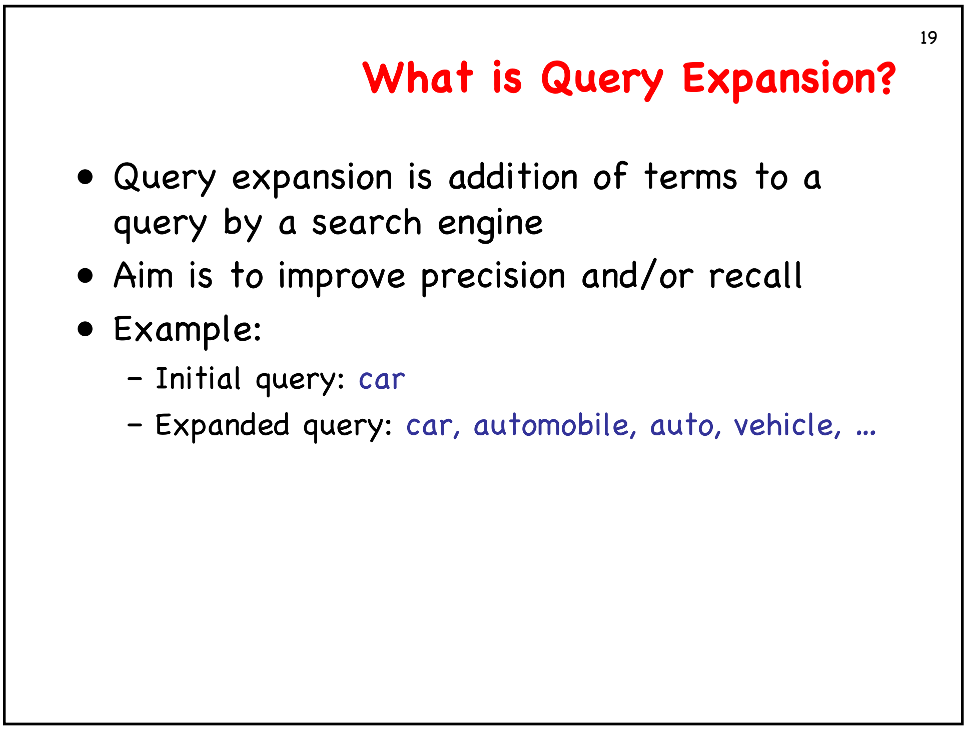
Relevance feedback: problems

- Users don't like giving feedback during search tasks
- Users' information needs (and hence perceptions of relevance) change

- So …
- Can we get the system to approximate user judgments?

18

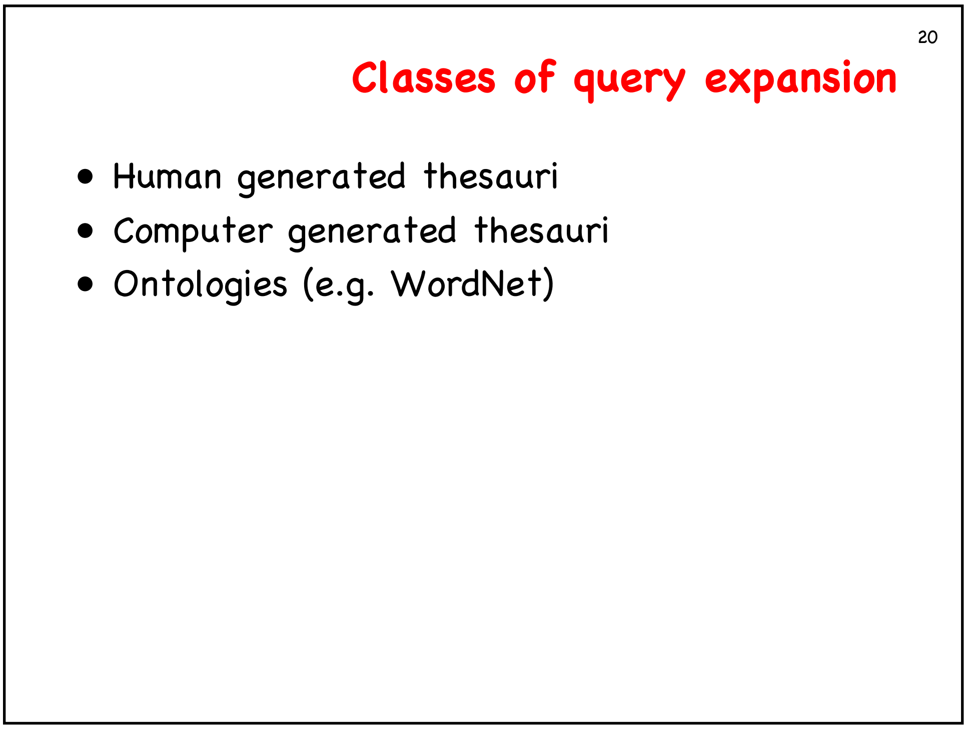
QUERY EXPANSION



19

What is Query Expansion?

- Query expansion is addition of terms to a query by a search engine
- Aim is to improve precision and/or recall
- Example:
  - Initial query: car
  - Expanded query: car, automobile, auto, vehicle, …



20

Classes of query expansion

- Human generated thesauri
- Computer generated thesauri
- Ontologies (e.g. WordNet)

Query expansion issues

- Two major issues
  - Which terms to include?
  - Which terms to weight more?
- Concept-based query expansion
  - Add terms which describe the query concepts
- Term-based query expansion
  - Add terms which are synonymous or similar to those in the query

Thesauri

- What is a thesaurus in the IR world?
  - "Any data structure that defines semantic relatedness between words."
    - *Schutze and Pedersen (1997)*
  - Often more complex than normal thesaurus
    - these are generally too broad to be useful

The need for thesauri

- Assumption: adding words from a thesaurus improves
  - **Recall**: the number of relevant documents retrieved
  - **Precision**: the proportion of relevant documents in the result set
    (or the number of relevant documents in the top *n* results)
- The car example: car or car, auto, automobile, vehicle, sedan, …
  - Which would retrieve most documents?
  - Less is more?

Thesaurus generation

- Early work in the 1960s
  - *Thesaurofacet* – detailed list of engineering terms

    Aitchison, J. (1970). The thesaurofacet: A multipurpose retrieval language tool. *Journal of Documentation*, 26(1), 187–203.
- Combines the idea of faceted search with lists of similar terms
- Largely used in such industries as medicine, aerospace, and other technological fields

25

Drawbacks of handcrafted thesauri

- Cost
  - Development
  - Maintenance
  - Cost often outweighs benefit
- Time
  - Takes a long time to develop a thesaurus
  - Hard to keep up with the pace of scientific and technological development
- Quality
  - Depends on quality of experts

26

Automatically generated thesauri

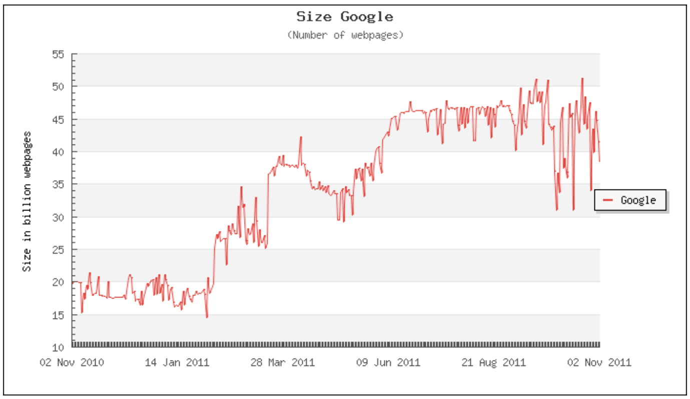
- Quicker to develop
- No longer have the cost of experts to generate thesauri
- Quality depends on the quality of construction techniques

Automatically generated thesauri

- Three steps.
  1. Extract word co-occurrences
  2. Define word similarities using
     - word co-occurrence
     - lexical relationship
  3. Cluster words based upon their similarities
- Not proven very successful

Relevance of query expansion

- The web keeps on growing
  - Google had 135 million pages in 1999
  - Indexable web in early 2005 was 11.5 billion pages
  - Current estimates suggest Google indexes about 40bn pages



http://www.worldwidewebsize.com/

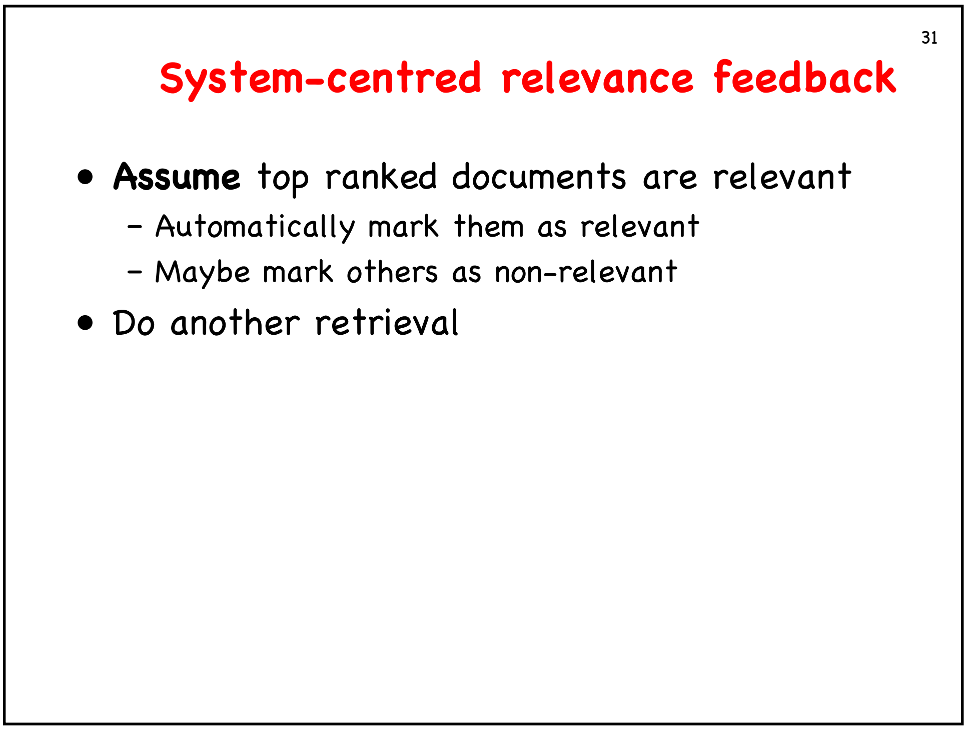
29

Query expansion: *idf* differences

- Compute *idf* in non-relevant docs
  - Approximate to main collection
- Compute *idf* in relevant docs
  - Rank terms on their difference
  - Add top *n* terms to query
  - Do another retrieval
  - Seems to work well
    - Harman, D. (1992): Relevance feedback revisited, *ACM SIGIR,* Copenhagen, 1-10

30

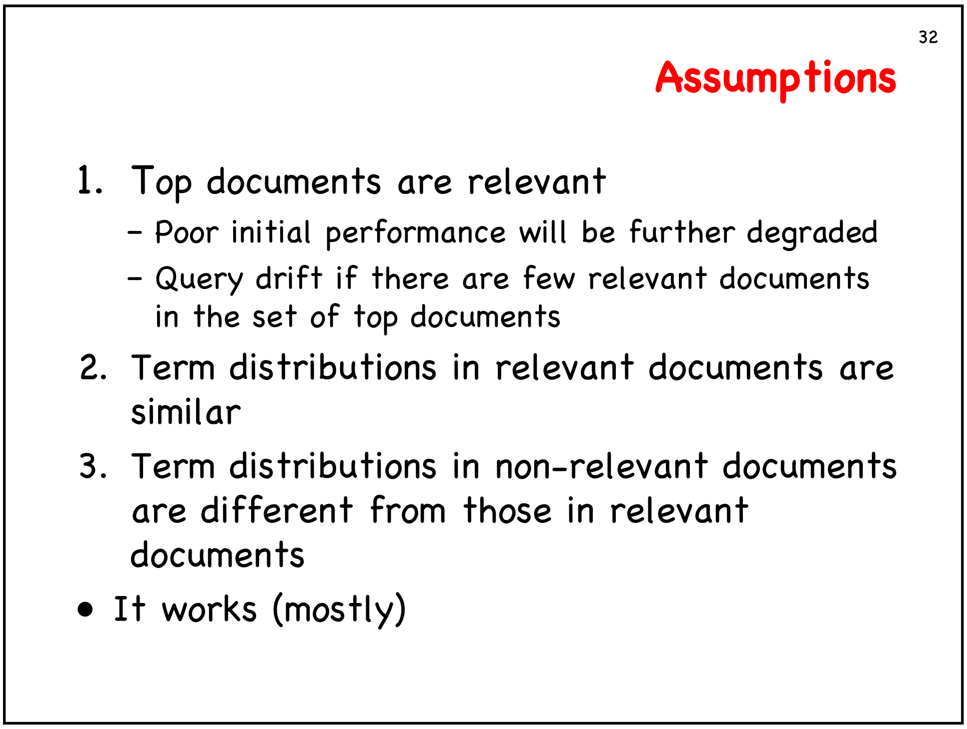
RELEVANCE FEEDBACK



31

System-centred relevance feedback

- **Assume** top ranked documents are relevant
  - Automatically mark them as relevant
  - Maybe mark others as non-relevant
- Do another retrieval



32

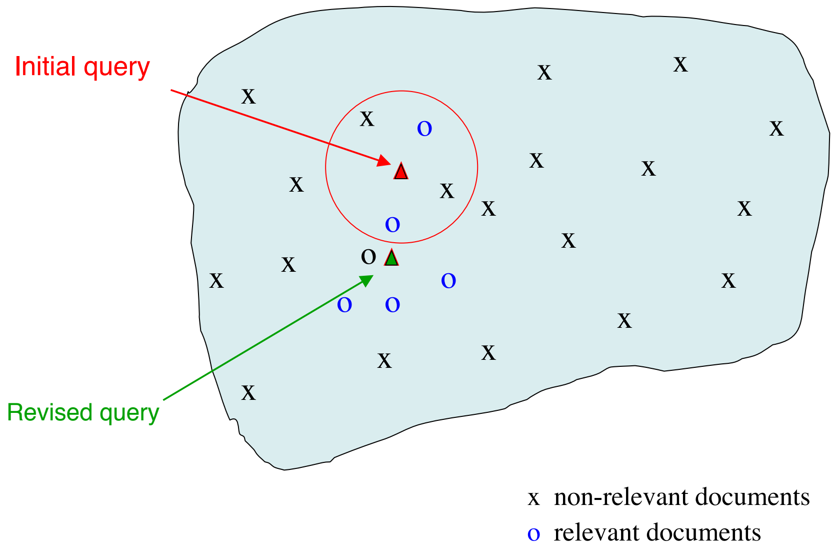
Assumptions

1. Top documents are relevant
   - Poor initial performance will be further degraded
   - Query drift if there are few relevant documents in the set of top documents
2. Term distributions in relevant documents are similar
3. Term distributions in non-relevant documents are different from those in relevant documents

- It works (mostly)

33

Relevance feedback: pictorial view



34

Relevance feedback: Rocchio algorithm

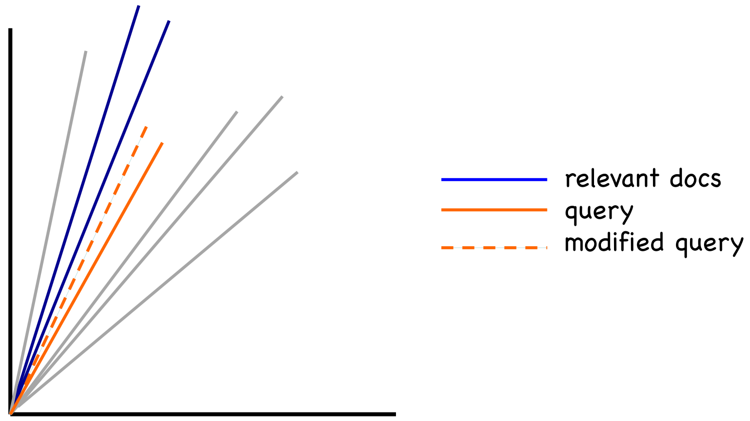
$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

q_m = modified query vector;
q_0 = original query vector;
α, β, γ: weights (hand-chosen or set empirically);
D_r = set of known relevant doc vectors;
D_{nr} = set of known irrelevant doc vectors

- The new query
  - moves towards relevant documents
  - away from non-relevant documents

35

Rocchio algorithm: geometric view



- The query vector is moved closer to the relevant documents

36

Rocchio example

$$\text{query vector} = \alpha \cdot \text{original query vector}$$
$$+ \beta \cdot \text{positive feedback vector}$$
$$- \gamma \cdot \text{negative feedback vector}$$

Typically, $\gamma < \beta$

Original query	0	4	0	8	0	0	$\alpha = 1.0$	0	4	0	8	0	0	
Positive Feedback	2	4	8	0	0	2	$\beta = 0.5$	1	2	4	0	0	1	(+)
Negative feedback	8	0	4	4	0	16	$\gamma = 0.25$	2	0	1	1	0	4	(-)
							New query	-1	6	3	7	0	-3	

37

Rocchio notes

- Many systems don't use only relevant documents:

$$\gamma = 0$$

- If non-relevant documents are used, negative weights are set to 0

38

Words are not enough?

- Query: cosmonaut
- Query: astronaut, moon

- Should they match?
  - Conventional system, no
  - Maybe they should?
    - How?

39

Lexical co-occurrence

- Instead of looking at the frequency of terms in a document, look at their proximity
- Context of words becomes important
- Some performance improvement shown in small document collections

40

Term co-occurrence

- Relationship between words based upon their co-occurrence in documents
- Clustering
  - Documents that share a significant number of terms are grouped together
  - A thesaurus is then generated from the terms in these classes
- Issues
  - Categories sometimes too narrow or broad
  - Does not account for synonyms

Resources

- Manning C. D., Raghavan R. and Schutze H. (2008) *Introduction to Information Retrieval*, CUP, Ch. 9
- Belew R. (2000) *Finding Out About*, Ch. 4
- Buckley C., Salton G., Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *ACM SIGIR* Dublin, 292-300
- Sarwar, B., Karypis G., Konstan J., Reidl J. (2001) Item-based collaborative filtering recommendation algorithms. *Proc. WWW '01* Hong Kong,
  doi= http://doi.acm.org/10.1145/371920.372071
- Cilibrasi R., Vitanyi P. (2006) Similarity of Objects and meaning of Words, TAMC, 21-45
  doi= http://dx.doi.org/10.1007/11750321_2