Information Needs

- An *information need* is the underlying cause of the query that a person submits to a search engine
  - Sometimes called *information problem* to emphasise that information need is generally related to a **task**
- Categorised using a variety of dimensions
  - e.g., number of relevant documents being sought
  - Type of information that is needed
  - Type of task that led to the requirement for information

3

Queries and Information Needs

- A query can represent very different information needs
  - May require different search techniques and ranking algorithms to produce the best rankings
- A query can be a poor representation of the information need
  - Users may find it difficult to express the information need
  - Users are encouraged to enter short queries both by the search engine interface, and by the fact that long queries often don't work
  - Ambiguity: the same query string may represent different information needs

4

Query Ambiguity

- 16+% of user queries are ambiguous (Song et al., IP&M 2009)
  - e.g., what is a user issuing the query *'ash'* after?

Query Formulation Problem

- Difficult to generate well formulated queries without
  - Knowledge of collection
    - How terms are distributed, type of docs, etc.
  - Retrieval environment
    - Term weighting, query language, retrieval strategy, etc
- First query is a trial run
  - Practically used to retrieve few useful items from a collection
  - Learn from those relevant ones
    - Query term modification

IR is an iterative process

Why is it difficult to formulate a query?

6

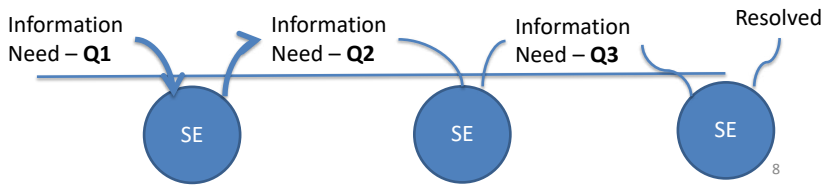
User Interaction

- Interaction with the system occurs
  - During query formulation and reformulation
  - While browsing the search results
- Key aspect of effective retrieval
  - Users can't change the ranking algorithm but can change the results through **interaction**
  - Helps refine the description of information need
    - e.g., same initial query, different information needs
    - How do users describe what they don't know?

7

Interaction & Query Reformulation

- The user needs to find some information. Why?
  - To do a task
  - 'Anomalous' state of knowledge [Belkin 1982]
  - Maybe the user is uncertain about something
  - The user formulates a query (information need is transformed into a query)
    - System helps user to refine the query and … accomplish the task



8

ASK Hypothesis

- Belkin et al (1982) proposed a model called Anomalous State of Knowledge
- ASK hypothesis:
  - Difficult for people to define exactly what their information need is, because that information is a **gap** in their knowledge
  - Search engine should look for information that fills those gaps
- Interesting idea, little practical impact (yet)

9

(Explicit) Interactions

"System helps user to refine the query"

Examples of explicit interaction:

•Relevance Feedback: the search engine permits the user to say what documents already retrieved are relevant or non-relevant

•Query Expansion/Query Term Suggestion: the search engine suggests additional query terms for the user

•Query Suggestions: the search engine suggests related queries for the user

10

Relevance Feedback

- After initial retrieval results are presented, allow the user to provide feedback on the relevance of one or more of the retrieved documents
- Use this feedback information to reformulate the query
- Produce new results based on reformulated query
- Allows more interactive, multi-pass process

11

Relevance Feedback Example

Permit user to select relevant documents

1. Badmans **Tropical Fish**
   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...
2. **Tropical Fish**
   Notes on a few species and a gallery of photos of African cichlids.
3. The **Tropical** Tank Homepage - **Tropical Fish** and Aquariums
   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...
4. **Tropical Fish** Centre
   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.
5. **Tropical fish** - Wikipedia, the free encyclopedia
   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...
6. **Tropical Fish** Find
   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...
7. Breeding **tropical fish**
   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...
8. FishLore
   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.
9. Cathy's **Tropical Fish** Keeping
   Information on setting up and maintaining a successful freshwater aquarium.
10. **Tropical Fish** Place
    Tropical Fish information for your freshwater **fish** tank ... great amount of information about a great hobby, a freshwater **tropical fish** tank. ...

Top 10 documents for "tropical fish"

12

Relevance Feedback Example

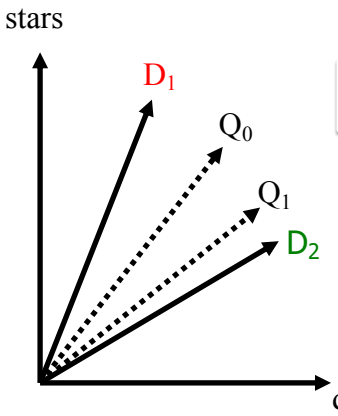
- Document 7 ("Breeding tropical fish") has ***explicitly*** been indicated to be relevant
- In this document, the most frequent terms are:

  breeding (4), fish (4), tropical (4), marine (2), pond (2), coldwater (2), keeping (1), interested (1)
- We can add these terms back into the query
  - This makes the query more similar to the relevant documents
  - E.g. {tropical fish} -> {tropical^5 fish^5 breeding^4 marine^2 …}
- Specific weights and scoring methods used for relevance feedback depend on retrieval model

13

Relevance Feedback Using Vectors

- Documents relevant to a particular query resemble each other in the sense that they are represented by reasonably similar vectors

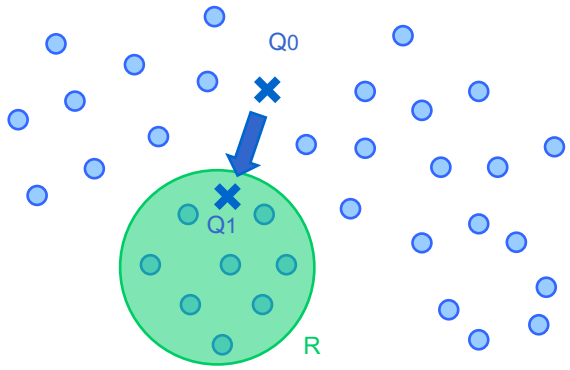


Moving a given query towards the relevant items and away from the non-relevant ones

Relevance Feedback within a Vector Space Model is due to J. J. Rocchio, 1971

14

Conceptual View of Relevance Feedback



15

Relevance Feedback

- Usually **both** of the following occur using the feedback:
  - Expand query with new terms
  - Re-weight terms in query
- There are many variations
  - Usually positive weights for terms from relevant docs
    - Found to be much valuable
  - Sometimes negative weights for terms from non-relevant docs
  - Remove terms that ONLY appear in non-relevant documents

16

Query Reformulation for VSM

- Change query vector using vector algebra
- **Add** the vectors for the **relevant** documents to the query vector
- **Subtract** the vectors for the **non-relevant** docs from the query vector

- This adds both positive and negatively weighted terms to the query as well as re-weighting the initial terms

Optimal Query

- Assume that the relevant set of documents C_r are known

- Then the best query that ranks all and only the relevant documents at the top is:

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j$$

where N is the total number of documents.

Standard Rochio Method

- Since all relevant documents are unknown, just use the **known** relevant (D_r) and non-relevant (D_n) sets of documents and include the initial query q_0

$$\vec{q}_{new} = \alpha\vec{q}_0 + \frac{\beta}{|D_r|}\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for non-relevant documents.
Usually α=1, β = 0.75 and γ = 0.25

Rocchio method vs optimal query?
what is the difference....
why is it needed?

19

Example: Rocchio Calculation

$D_{r1} = (.030, 0.00, 0.00, .025, .025, .050, 0.00, 0.00, .120)$ Relevant docs
$D_{r2} = (.020, .009, .020, .002, .050, .025, .100, .100, .120)$

$D_{n1} = (.030, .010, .020, 0.00, .005, .025, 0.00, .020, 0.00)$ Non-rel doc
$q_0 = (0.00, 0.00, 0.00, 0.00, .500, 0.00, .450, 0.00, .950)$ Original Query

$\alpha = 1$
$\beta = 0.75$ Constants
$\gamma = 0.25$

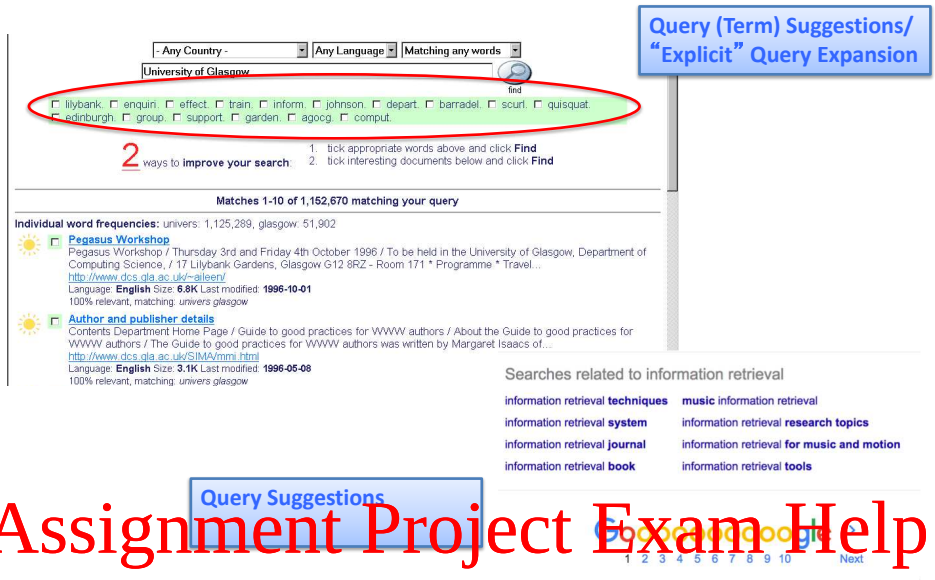
$$q_{new} = \alpha q_0 + \left(\frac{\beta}{2} \times (D_{r1} + D_{r2})\right) - \left(\frac{\gamma}{1} \times D_{n1}\right)$$ Rocchio Calculation

Resulting Feedback Query

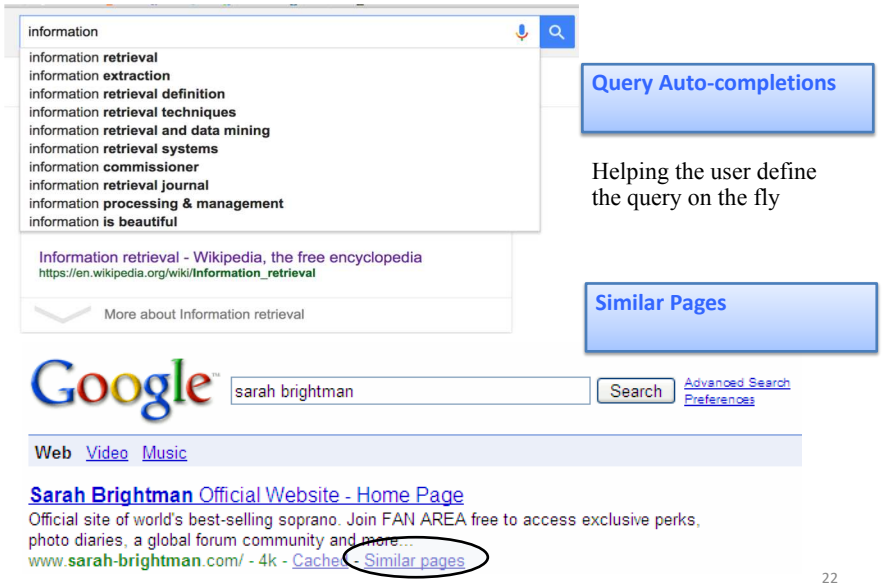
$q_{new} = (0.011, 0.000875, 0.002, 0.01, 0.527, 0.022, 0.488, 0.033, 1.04)$

20

Other forms of Explicit Interaction (1)



Other forms of Explicit Interaction (2)



Relevance Feedback (RF) Performance

- RF generally improves retrieval performance (recall and precision)
- RF is most useful for increasing *recall* in situations where recall is important
  - Users can be expected to review results and to take time to iterate
- Positive feedback is more valuable than negative feedback
  - E.g. set $\gamma < \beta$; e.g. $\gamma = 0.25$, $\beta = 0.75$
  - Many systems only allow positive feedback ($\gamma = 0$)

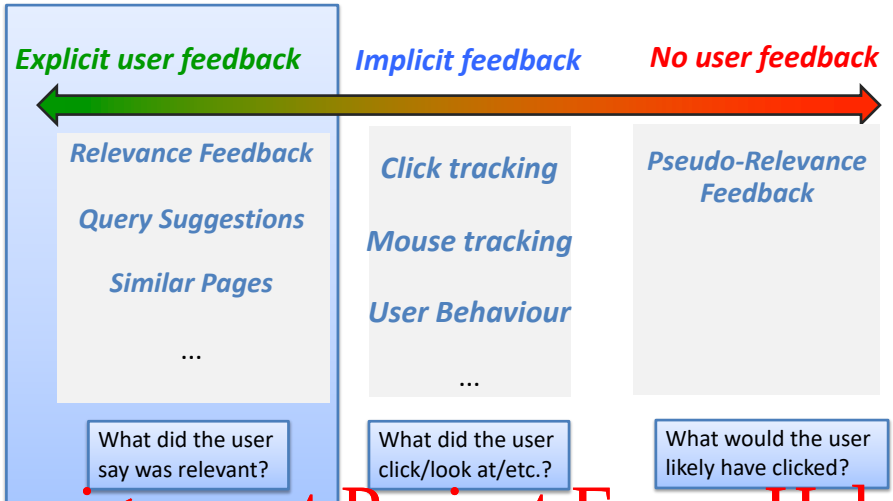
23

Why is Relevance Feedback Not Widely Used?

- Users are in general reluctant to provide explicit feedback (see Excite & Altavista experiments)
- Results in long queries that require more computation to retrieve
  - Search engines process lots of queries and allow little time for each one
- Makes it harder to understand why a particular document was retrieved

24

Types of User Feedback



25

Pseudo-Relevance Feedback

- Use relevance feedback methods without explicit user input
- Just assume the top *m* retrieved documents are relevant, and use them to reformulate the query:
  - Look at the statistics of the terms in the top retrieved documents
  - Add the terms with highest weight to the query
  - Do relevance feedback (e.g. Rocchio) and carry out the retrieval with new query
- Aka (*automatic*) **Query Expansion (QE)**
  - A number of such QE models are implemented in Terrier

26

Pseudo Relevance Feedback: Illustrative Example

- TREC Query: *Scottish highland games*
- Retrieve top 3 documents using the original query, and use the index to determine what other terms occur in those documents
- Using **Terrier's Bo1 QE** mechanism and Weak Stemming
  - Bo1 generalises Rocchio's approach by using a refined parameter-free statistical term weighting model (no **model** parameters are needed)
- The expanded query:
  - *Scottish highland games* Ligonier kilt caber clan toss Scot tartan grandfather artist heavy tradition dance Celtic dancer athlete heather competitor
- In the expanded query (using the relevance assessment)
  - These terms are helpful: Ligonier kilt caber clan toss Scot tartan
  - These terms bring noise: grandfather artist heavy
  - The rest of the added query terms are neutral, e.g. dancer, tradition

27

Query Re-weighting

- Assume that in the expanded query, the terms are not re-weighted
- Assume that *qtw* is the frequency of the query term
- For the expanded query:

  "*Scottish highland games* Ligonier kilt caber clan toss scot tartan tradition dance Celtic dancer athlete heather competitor grandfather artist heavy"
  - *qtw* = 1 for each query term
  - We can't differentiate the informative terms from the others
- Therefore, there is a need for re-weighting the query terms, including the expanded ones

28

Term	Weight	Term	Weight
Scottish	*1.5000*	*highland*	*1.4087*
games	*1.3105*	*Ligonier*	*0.3609*
kilt	*0.2897*	*caber*	*0.1347*
clan	*0.1291*	*tradition*	*0.1189*
dance	*0.1115*	*Celtic*	*0.1067*
toss	*0.1062*	*dancer*	*0.1013*
grandfather	*0.1009*	*Scot*	*0.0895*
athlete	*0.0745*	*heather*	*0.0673*
artist	*0.0643*	*heavy*	*0.0606*
tartan	*0.0587*	*competitor*	*0.0427*

29

Pseudo-Relevance Feedback Works

- Pseudo-relevance feedback automates the "manual" part of true (explicit) relevance feedback
- Works well on average in many experimental settings
- Example:
  - TREC 2005 Terabyte Track adhoc task
  - Using Terrier's TF-IDF, MAP=0.3024
  - Using Terrier's TF-IDF+**Bo1**, MAP=0.3428
  - A significant improvement (p-value=0.008169)

30

Problems with Pseudo-Relevance Feedback

- If the initial query is hard
  - That means the initial results are poor
    - Then pseudo-RF creates problems often drifting away from an optimal query
- May not be rewarding if the number of relevant documents is small
  - Adding more query terms cannot bring many relevant documents
  - e.g. query '*Homepage of Glasgow Information Retrieval Group*' has only a unique relevant document. It is not helpful to expand the query

31

Other Sources of Evidence for Query Expansion

- **Query Expansion** can examine different sources of evidence to automatically reformulate a query:
  - Top-ranked documents: based on term co-occurrence with query terms (pseudo-relevance feedback)
  - Also the entire corpus (incl. an external corpus), using a similar term co-occurrence analysis
  - Previous query reformulations of users (query logs)
  - Thesauri & dictionaries (e.g. WordNet) & **word embeddings**
    - However, they are not necessarily effective, as they do not take context into account

32

Thesaurus

- A thesaurus provides information on synonyms and semantically related words and phrases.
- Example:

```
physician
  syn: ||croaker, doc, doctor, MD,
medical, mediciner, medico, ||sawbones
  rel: medic, general practitioner,
surgeon, …
```


33

Thesaurus-based Query Expansion

- For each term, *t*, in a query, expand the query with synonyms and related words of *t* from the thesaurus
- You might weight the added terms less than the original query terms
- Generally increases recall
- It could significantly decrease precision, particularly with ambiguous terms
  - "interest rate" → "interest rate fascinate evaluate"

34

WordNet

- A more detailed database of semantic relationships between English words
- Developed by well-known cognitive psychologist George Miller and a team at Princeton University.
- About 155,000 English words.
- Nouns, adjectives, verbs, and adverbs grouped into about 175k synonym sets called ***synsets***

https://wordnet.princeton.edu/

33

WordNet Synset Relationships

- Antonym: front → back
- Attribute: benevolence → good (noun to adjective)
- Pertainym: alphabetical → alphabet (adjective to noun)
- Similar: unquestioning → absolute
- Cause: kill → die
- Entailment: breathe → inhale
- Holonym: chapter → text (part to whole)
- Meronym: computer → cpu (whole to part)
- Hyponym: plant → tree (specialization)
- Hypernym: apple → fruit (generalization)

36

Query Expansion with WordNet

- The expanded query terms are usually synonyms of the original query terms (i.e. words in the same synset)
- Example: for "*Scottish highland games*", we have the following expanded terms:
  - Scotch, Gaelic, upland, hilly, mountainous, bet, gage, stake, punt etc.
- It is also possible to add hyponyms (specialised terms) or to add hypernyms to generalise a query

37

Statistical Thesaurus

- Existing human-developed thesauri are not easily available in all languages
  - Human thesuari are **limited** in the type and range of synonymy and semantic relations they represent
- Semantically related terms can be discovered from the statistical analysis of corpora such as term co-occurrence:
  - If term t1 occurs in X documents, and a term t2 occurs in Y documents and they co-occur in C documents:
    - Cosine (t1, t2) =C/sqrt(X*Y)
    - Dice (t1, t2) =2C/(X+Y)
- Other approaches involve distributional semantics
  - **Word embeddings** - word2vec, glove, etc
- Does not usually work well compared to pseudo-relevance feedback

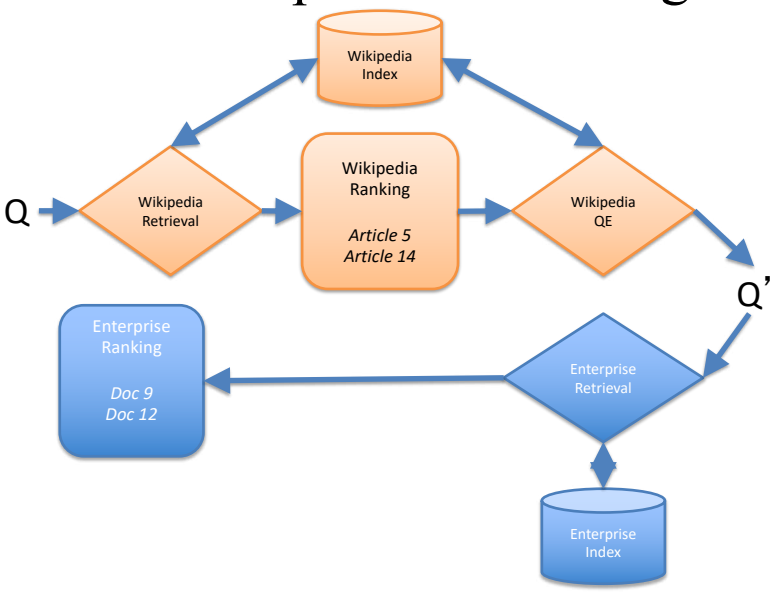
38

Query Expansion Using External Resource

- A **high-quality external** resource can bring useful information that improves the quality of the pseudo-relevance set
- [Kwok 2003]
  - The pseudo-relevance set contains top-ranked documents returned by Google!
  - Aka Collection Enrichment or External Query Expansion
- Works well on large collections, including Web & Microblog corpora
  - The larger the external resource is, the better the terms follow a random distribution
  - Advanced term weighting models are based on randomness statistical laws
  - A large external resource provides better collection statistics for the term weighting

39

Example of External QE For Enterprise Search Engine



40

Selective Query Expansion

- **Idea**: Disable query expansion if the pseudo-relevance set is predicted to be poor
  - Various **query performance predictors** have been proposed [He & Ounis, 2004]
  - For a given query, estimate how well a given query will do using a **statistical analysis** of query/corpus and/or the retrieved set of documents
    - E.g. are the query terms not informative enough, or is the query ambiguous (e.g. very topically dissimilar terms)
- Works well, and particularly appropriate for retrieval tasks that require early precision (e.g. Web search)

41

Wisdom of the Crowds

- Consider the query: Scotch whisky shop Bellevue
- There are no such shops in Bellevue, so the user may reformulate to Scotch whisky shop Seattle
- If many users use similar query reformulations, from the **query logs** we can understand that Bellevue → Seattle is a possible reformulation/ expansion:

Source	Target	Frequency/Score
bellevue	seattle	6
bellevue	bellevue ne	5
bellevue	bellevue washington	4

42

Summary

- Relevance feedback is an effective means for user-directed query modification
- Modification can be done with either **direct** or **indirect** user input
- Modification can be done based on an individual's or a group's past input

43