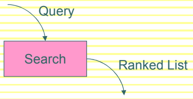
LBSC 796/INFM 718R: Week 8

Relevance Feedback

Jimmy Lin
College of Information Studies
University of Maryland

Monday, March 27, 2006

The IR Black Box

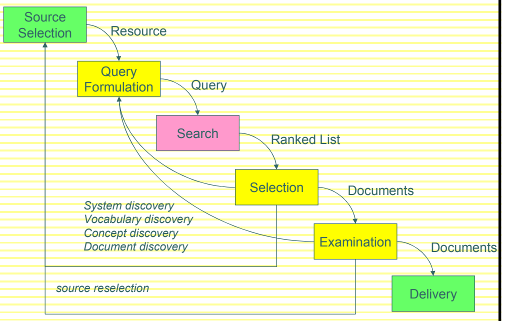

Query
Search
Ranked List

Anomalous State of Knowledge

- Basic paradox:
  - Information needs arise because the user doesn't know something: "an anomaly in his state of knowledge with respect to the problem faced"
  - Search systems are designed to satisfy these needs, but the user needs to know what he is looking for
  - However, if the user knows what he's looking for, there may not be a need to search in the first place
- Implication: computing "similarity" between queries and documents is fundamentally wrong
- How do we resolve this paradox?

Nicholas J. Belkin. (1980) Anomalous States of Knowledge as a Basis for Information Retrieval. *Canadian Journal of Information Science*, 5, 133-143.

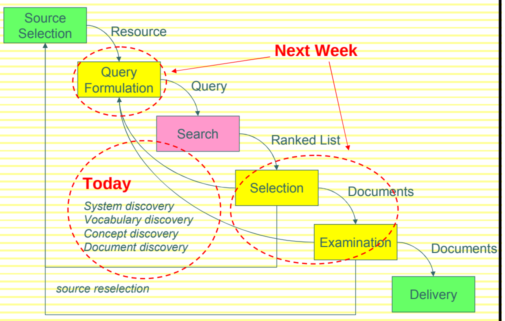
The Information Retrieval Cycle


Source Selection
Resource
Query Formulation
Query
Search
Ranked List
Selection
Documents
Examination
Documents
Delivery

System discovery
Vocabulary discovery
Concept discovery
Document discovery

source reselection

Upcoming Topics


Source Selection
Resource
Next Week
Query Formulation
Query
Search
Ranked List
Today
System discovery
Vocabulary discovery
Concept discovery
Document discovery
Selection
Documents
Examination
Documents
Delivery
source reselection

Different Types of Interactions

- System discovery – learning capabilities of the system
  - Playing with different types of query operators
  - "Reverse engineering" a search system
- Vocabulary discovery – learning collection-specific terms that relate to your information need
  - The literature on aerodynamics refers to *aircrafts*, but you query on *planes*
  - How do you know what terms the collection uses?

Different Types of Interactions

- Concept discovery – learning the concepts that relate to your information need
  - What's the name of the disease that Reagan had?
  - How is this different from vocabulary discovery?
- Document discovery – learning about the types of documents that fulfill your information need
  - Were you looking for a news article, a column, or an editorial?

Relevance Feedback

- Take advantage of user relevance judgments in the retrieval process:
  - User issues a (short, simple) query and gets back an initial hit list
  - User marks hits as relevant or non-relevant
  - The system computes a better representation of the information need based on this feedback
  - Single or multiple iterations (although little is typically gained after one iteration)
- Idea: you may not know what you're looking for, but you'll know when you see it

Outline

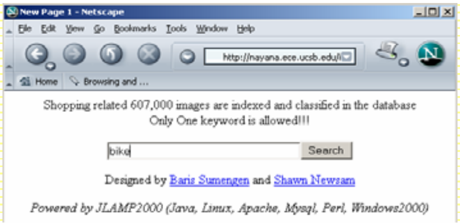
- Explicit feedback: users explicitly mark relevant and irrelevant documents
- Implicit feedback: system attempts to infer user intentions based on observable behavior
- Blind feedback: feedback in absence of any evidence, explicit or otherwise

Why relevance feedback?

- You may not know what you're looking for, but you'll know when you see it
- Query formulation may be difficult; simplify the problem through iteration
- Facilitate vocabulary and concept discovery
- Boost recall: "find me more documents like this…"

Relevance Feedback Example

Image Search Engine
http://nayana.ece.ucsb.edu/imsearch/imsearch.html



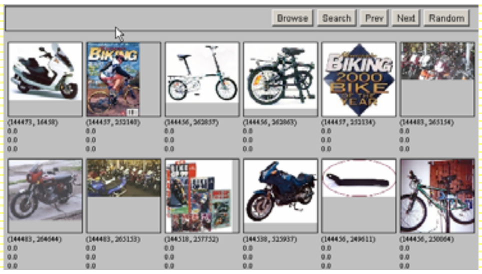
Shopping related 607,000 images are indexed and classified in the database
Only One keyword is allowed!!!

bike

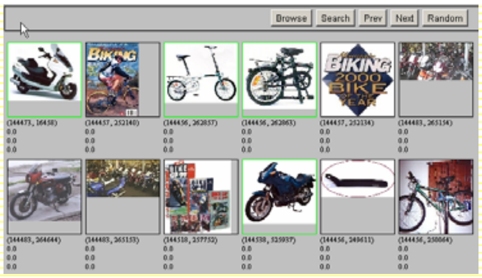
Designed by Baris Sumengen and Shawn Newsam

Powered by JLAMP2000 (Java, Linux, Apache, Mysql, Perl, Windows2000)

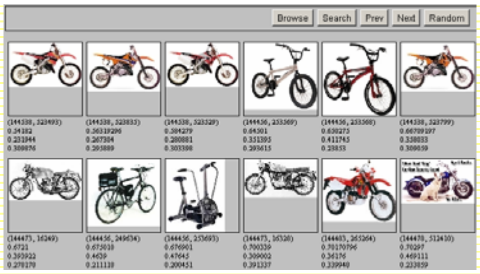
Initial Results



Relevance Feedback



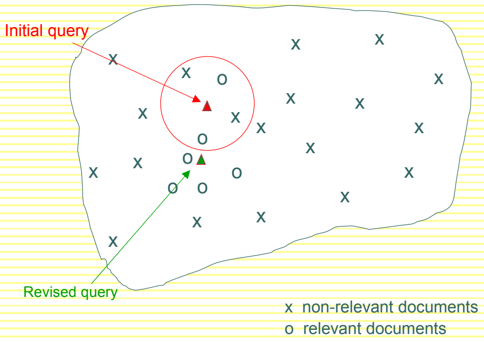
Revised Results



Updating Queries

- Let's assume that there is an optimal query
  - The goal of relevance feedback is to bring the user query closer to the optimal query
- How does relevance feedback actually work?
  - Use relevance information to update query
  - Use query to retrieve new set of documents
- What exactly do we "feed back"?
  - Boost weights of terms from relevant documents
  - Add terms from relevant documents to the query
  - Note that this is hidden from the user

Picture of Relevance Feedback



Rocchio Algorithm

- Used in practice:

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

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

- New query
  - Moves toward relevant documents
  - Away from irrelevant documents

Rocchio in Pictures

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

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	

Relevance Feedback: Assumptions

- A1: User has sufficient knowledge for a reasonable initial query
- A2: Relevance prototypes are “well-behaved”

Violation of A1

- User does not have sufficient initial knowledge
- Not enough relevant documents are retrieved in the initial query
- Examples:
  - Misspellings (Brittany Speers)
  - Cross-language information retrieval
  - Vocabulary mismatch (e.g., cosmonaut/astronaut)

Relevance Prototypes

- Relevance feedback assumes that relevance prototypes are “well-behaved”
  - All relevant documents are clustered together
  - Different clusters of relevant documents, but they have significant vocabulary overlap
- In other words,
  - Term distribution in relevant documents will be similar
  - Term distribution in non-relevant documents will be different from those in relevant documents

Violation of A2

- There are several clusters of relevant documents
- Examples:
  - Burma/Myanmar
  - Contradictory government policies
  - Opinions

Evaluation

- Compute standard measures with q_0
- Compute standard measures with q_m
  - Use all documents in the collection
    - Spectacular improvements, but… it’s cheating!
    - The user already selected relevant documents
  - Use documents in residual collection (set of documents minus those assessed relevant)
    - More realistic evaluation
    - Relative performance can be validly compared
- Empirically, one iteration of relevance feedback produces significant improvements
  - More iterations don’t help

Relevance Feedback: Cost

- Speed and efficiency issues
  - System needs to spend time analyzing documents
  - Longer queries are usually slower
- Users often reluctant to provide explicit feedback
- It’s often harder to understand why a particular document was retrieved

Koenemann and Belkin's Work

- Well-known study on relevance feedback in information retrieval
- Questions asked:
  - Does relevance feedback improve results?
  - Is user control over relevance feedback helpful?
  - How do different levels of user control effect results?

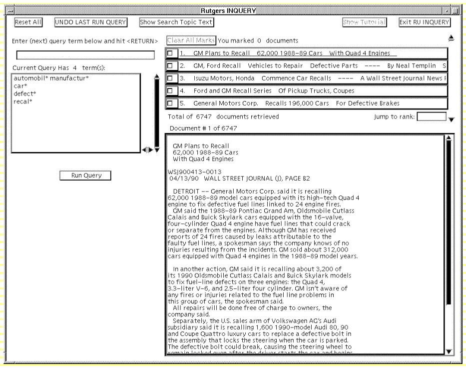
Jürgen Koenemann and Nicholas J. Belkin. (1996) A Case For Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. *Proceedings of SIGCHI 1996 Conference on Human Factors in Computing Systems (CHI 1996).*

What's the best interface?

- Opaque (black box)
  - User doesn't get to see the relevance feedback process
- Transparent
  - User shown relevance feedback terms, but isn't allowed to modify query
- Penetrable
  - User shown relevance feedback terms and is allowed to modify the query

Which do you think worked best?

Query Interface



You marked 0 documents

☐ 1. GM Plans to Recall 62,000 1988-89 Cars With Quad 4 Engines
☐ 2. GM, Ford Recall Vehicles to Repair Defective Parts ---- By Neal Templin S
☐ 3. Isuzu Motors, Honda Commence Car Recalls ---- A Wall Street Journal News I
☐ 4. Ford and GM Recall Series Of Pickup Trucks, Coupes
☐ 5. General Motors Corp. Recalls 196,000 Cars For Defective Brakes

Total of 6747 documents retrieved

Document # 1 of 6747

GM Plans to Recall
62,000 1988-89 Cars
With Quad 4 Engines

WSJ900413-0013
04/13/90 WALL STREET JOURNAL (J), PAGE B2

DETROIT -- General Motors Corp. said it is recalling 62,000 1988-89 model cars equipped with its high-tech Quad 4 engine to fix defective fuel lines linked to 24 engine fires.

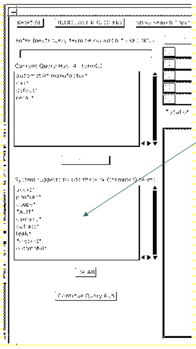
GM said the 1988-89 Pontiac Grand Am, Oldsmobile Cutlass Calais and Buick Skylark cars equipped with the 16-valve, four-cylinder Quad 4 engine have fuel lines that could crack or separate from the engines. Although GM has received reports of 24 fires caused by leaks attributable to the faulty fuel lines, a spokesman says the company knows of no injuries resulting from the incidents. GM sold about 312,000 cars equipped with Quad 4 engines in the 1988-89 model years.

In another action, GM said it is recalling about 3,200 of its 1990 Oldsmobile Cutlass Calais and Buick Skylark models to fix fuel-line defects on three engines: the Quad 4, 3.3-liter V-6, and 2.5-liter four cylinder. GM isn't aware of any fires or injuries related to the fuel line problems in this group of cars, the spokesman said.

All repairs will be done free of charge to owners, the company said.

Separately, the U.S. sales arm of Volkswagen AG's Audi subsidiary said it is recalling 1,600 1990-model Audi 80, 90 and Coupe Quattro luxury cars to replace a defective bolt in the assembly that locks the steering when the car is parked. The defective bolt could break, causing the steering wheel to

Penetrable Interface

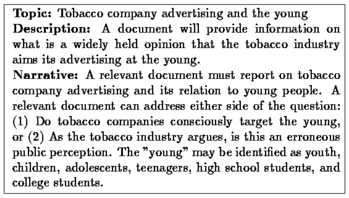


Users get to select which terms they want to add

Study Details

- Subjects started with a tutorial
  - 64 novice searchers (43 female, 21 male)
- Goal is to keep modifying the query until they've developed one that gets high precision
- INQUERY system used
- TREC collection (Wall Street Journal subset)
- Two search topics:
  - Automobile Recalls
  - Tobacco Advertising and the Young
- Relevance judgments from TREC and experimenter

Sample Topic

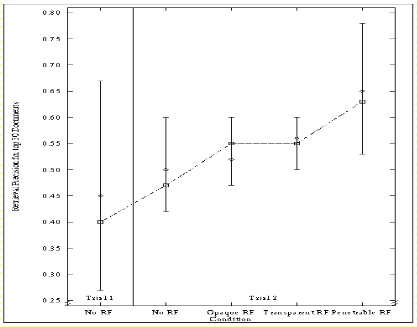


Topic: Tobacco company advertising and the young
Description: A document will provide information on what is a widely held opinion that the tobacco industry aims its advertising at the young.
Narrative: A relevant document must report on tobacco company advertising and its relation to young people. A relevant document can address either side of the question: (1) Do tobacco companies consciously target the young, or (2) As the tobacco industry argues, is this an erroneous public perception. The "young" may be identified as youth, children, adolescents, teenagers, high school students, and college students.

Procedure

- Baseline (Trial 1)
  - Subjects get tutorial on relevance feedback
- Experimental condition (Trial 2)
  - Shown one of four modes: no relevance feedback, opaque, transparent, penetrable
- Evaluation metric used: precision at 30 documents

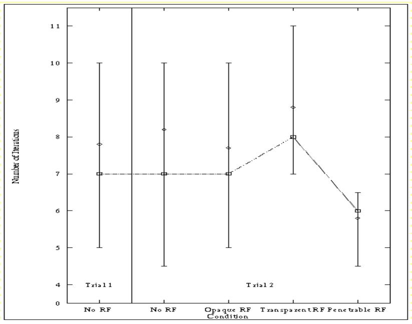
Precision Results



Relevance feedback works!

- Subjects using the relevance feedback interfaces performed 17-34% better
- Subjects in the penetrable condition performed 15% better than those in opaque and transparent conditions

Number of Iterations



Behavior Results

- Search times approximately equal
- Precision increased in first few iterations
- Penetrable interface required fewer iterations to arrive at final query
- Queries with relevance feedback are much longer
  - But fewer terms with the penetrable interface — users were more selective about which terms to add

Implicit Feedback

- Users are often reluctant to provide relevance judgments
  - Some searches are precision-oriented
  - They're lazy!
- Can we gather feedback without requiring the user to do anything?
- Idea: gather feedback from observed user behavior

Observable Behavior

		Minimum Scope		
		Segment	**Object**	**Class**
Behavior Category	**Examine**	View Listen	Select	
	Retain	Print	Bookmark Save Purchase Delete	Subscribe
	Reference	Copy / paste Quote	Forward Reply Link Cite	
	Annotate	Mark up	Rate Publish	Organize

Discussion Point

- How might user behaviors provide clues for relevance feedback?

So far…

- Explicit feedback: take advantage of user-supplied relevance judgments
- Implicit feedback: observe user behavior and draw inferences
- Can we perform feedback without having a user in the loop?

Blind Relevance Feedback

- Also called “pseudo relevance feedback”
- Motivation: it’s difficult to elicit relevance judgments from users
  - Can we automate this process?
- Idea: take top *n* documents, and simply *assume* that they are relevant
- Perform relevance feedback as before
- If the initial hit list is reasonable, system should pick up good query terms
- Does it work?

BRF Experiment

- Retrieval engine: Indri
- Test collection: TREC, topics 301-450
- Procedure:
  - Used topic description as query to generate initial hit list
  - Selected top 20 terms from top 20 hits using *tf.idf*
  - Added these terms to the original query

BRF Example

Number: 303
Title: Hubble Telescope Achievements

Description:
Identify positive accomplishments of the Hubble telescope since it was launched in 1991.

Narrative:
Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. Documents limited to the shortcomings of the telescope would be irrelevant. Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.

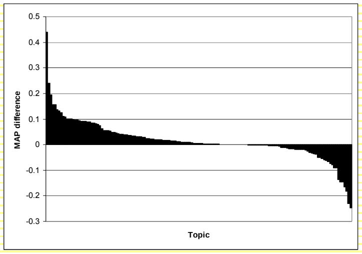
Terms added →

telescope	1041.33984032195
hubble	573.896477205696
space	354.090789112131
nasa	346.475671454331
ultraviolet	242.588034029191
shuttle	230.448255669841
mirror	184.794966339329
telescopes	155.290920607708
earth	148.865466409231
discovery	146.718067628756
orbit	142.597040178043
flaw	141.832019493907
scientists	132.384677410089
launch	116.322861618261
stars	116.205713485691
universe	114.705686405825
mirrors	113.677943638299
light	113.59717006967
optical	106.198288687586
species	103.555123536418

Results

	MAP	R-Precision
No feedback	0.1591	0.2022
With feedback	0.1806 (+13.5%)	0.2222 (+9.9%)

Blind relevance feedback doesn't always help!



The Complete Landscape

- Explicit, implicit, blind feedback: it's all about manipulating terms
- Dimensions of query expansion
  - "Local" vs. "global"
  - User involvement vs. no user involvement

Local vs. Global

- "Local" methods
  - Only considers documents that have be retrieved by an initial query
  - Query specific
  - Computations must be performed on the fly
- "Global" methods
  - Takes entire document collection into account
  - Does not depend on the query
  - Thesauri can be computed off-line (for faster access)

User Involvement

- Query expansion can be done automatically
  - New terms added without user intervention
- Or it can place a user in the loop
  - System presents suggested terms
  - Must consider interface issues

Query Expansion Techniques

- Where do techniques we've discussed fit?

	Local	**Global**
Manual		
Automatic		

Global Methods

- Controlled vocabulary
  - For example, MeSH terms
- Manual thesaurus
  - For example, WordNet
- Automatically derived thesaurus
  - For example, based on co-occurrence statistics

Using Controlled Vocabulary

PubMed Query:

```
("neoplasms"[MeSH Terms] OR cancer[Text Word])
```

Thesauri

- A thesaurus may contain information about lexical semantic relations:
  - Synonyms: similar words
    e.g., violin → fiddle
  - Hypernyms: more general words
    e.g., violin → instrument
  - Hyponyms: more specific words
    e.g., violin → Stradivari
  - Meronyms: parts
    e.g., violin → strings

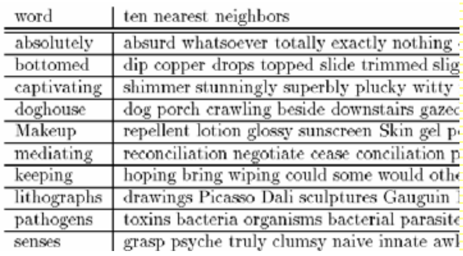
Using Manual Thesauri

- For each query term *t*, added synonyms and related words from thesaurus
  - feline → feline cat
- Generally improves recall
- Often hurts precision
  - "interest rate" → "interest rate fascinate evaluate"
- Manual thesauri are expensive to produce and maintain

Automatic Thesauri Generation

- Attempt to generate a thesaurus automatically by analyzing the document collection
- Two possible approaches
  - Co-occurrence statistics (co-occurring words are more likely to be similar)
  - Shallow analysis of grammatical relations
    - Entities that are grown, cooked, eaten, and digested are more likely to be food items.

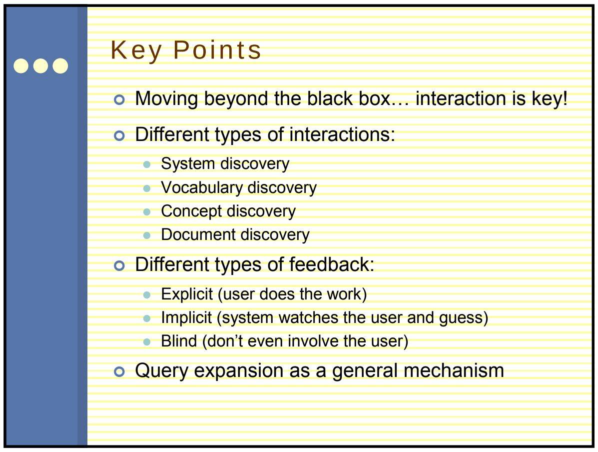
Automatic Thesauri: Example



word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gaze
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasit
senses	grasp psyche truly clumsy naive innate awl

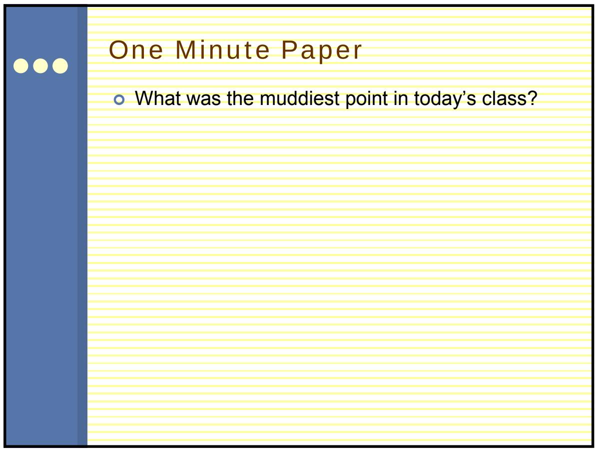
Automatic Thesauri: Discussion

- Quality of associations is usually a problem
- Term ambiguity may introduce irrelevant statistically correlated terms.
  - "Apple computer" → "Apple red fruit computer"
- Problems:
  - False positives: Words deemed similar that are not
  - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents



Key Points

- Moving beyond the black box… interaction is key!
- Different types of interactions:
  - System discovery
  - Vocabulary discovery
  - Concept discovery
  - Document discovery
- Different types of feedback:
  - Explicit (user does the work)
  - Implicit (system watches the user and guess)
  - Blind (don't even involve the user)
- Query expansion as a general mechanism



One Minute Paper

- What was the muddiest point in today's class?