Introduction to Information Retrieval

Relevance Feedback, Query Expansion

This Lecture

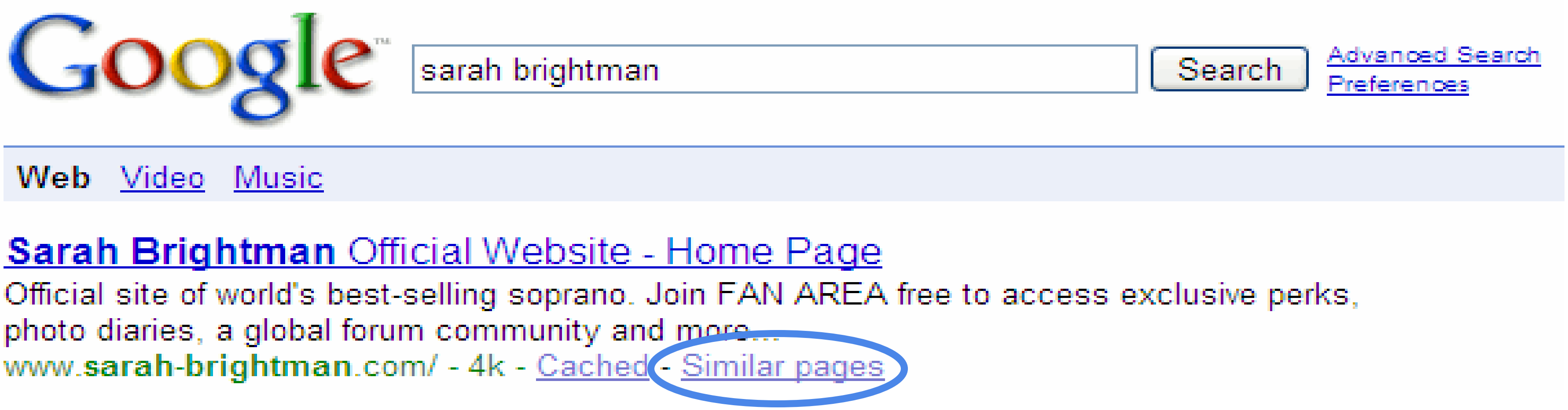
- Improving results
    - For high recall. E.g., searching for aircraft doesn't match with plane; nor thermodynamic with heat
- Options for improving results...
    - Local methods
        - Relevance feedback
        - Pseudo relevance feedback
    - Global methods
        - Query expansion
            - Thesauri
            - Automatic thesaurus generation

Relevance Feedback

Relevance Feedback

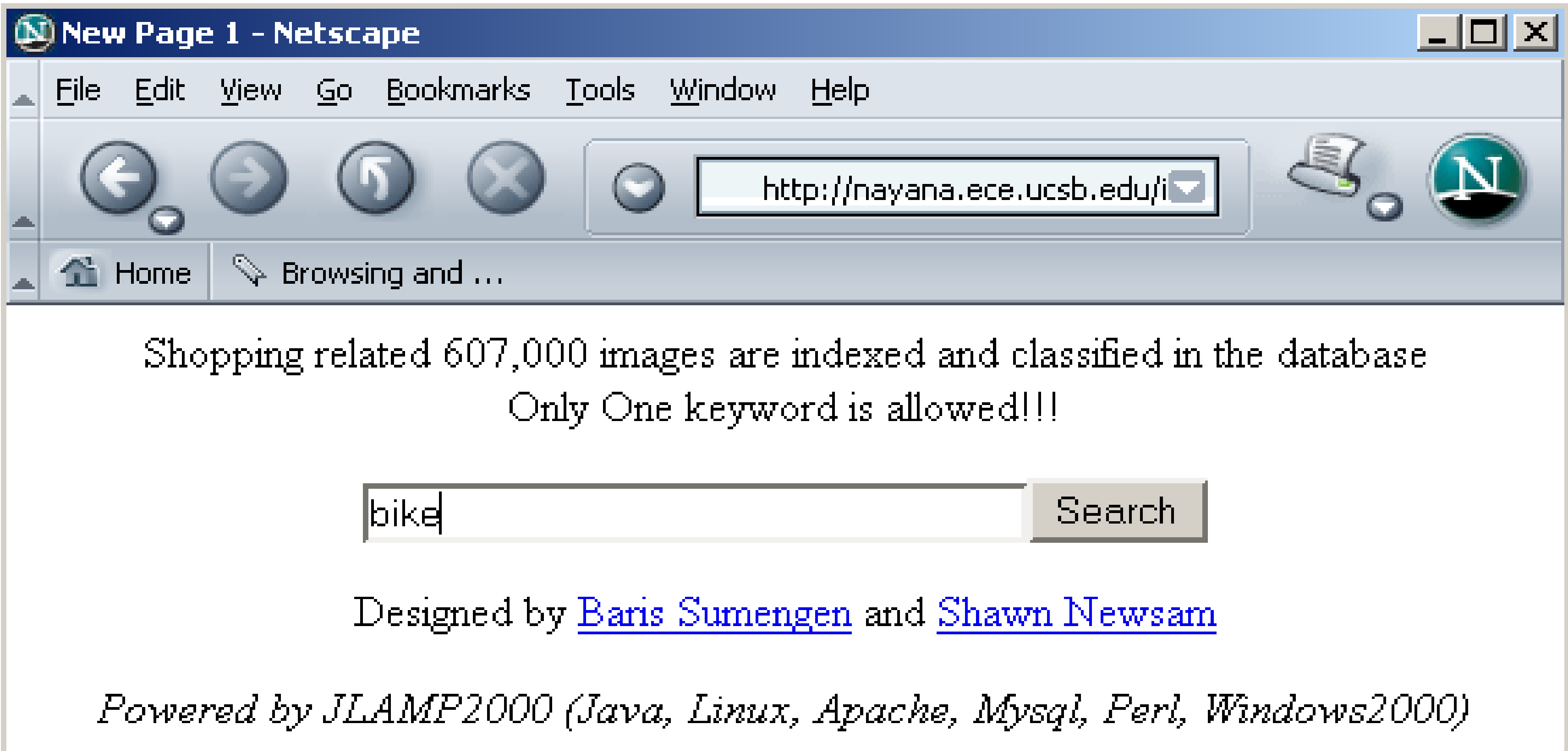
- Relevance feedback: user feedback on relevance of docs in initial set of results
  - User issues a (short, simple) query
  - The user marks some results as relevant or non-relevant.
  - The system computes a better representation of the information need based on feedback.
  - Relevance feedback can go through one or more iterations.
- Idea: it may be difficult to formulate a good query when you don’t know the collection well, so iterate
- Examples

Similar pages

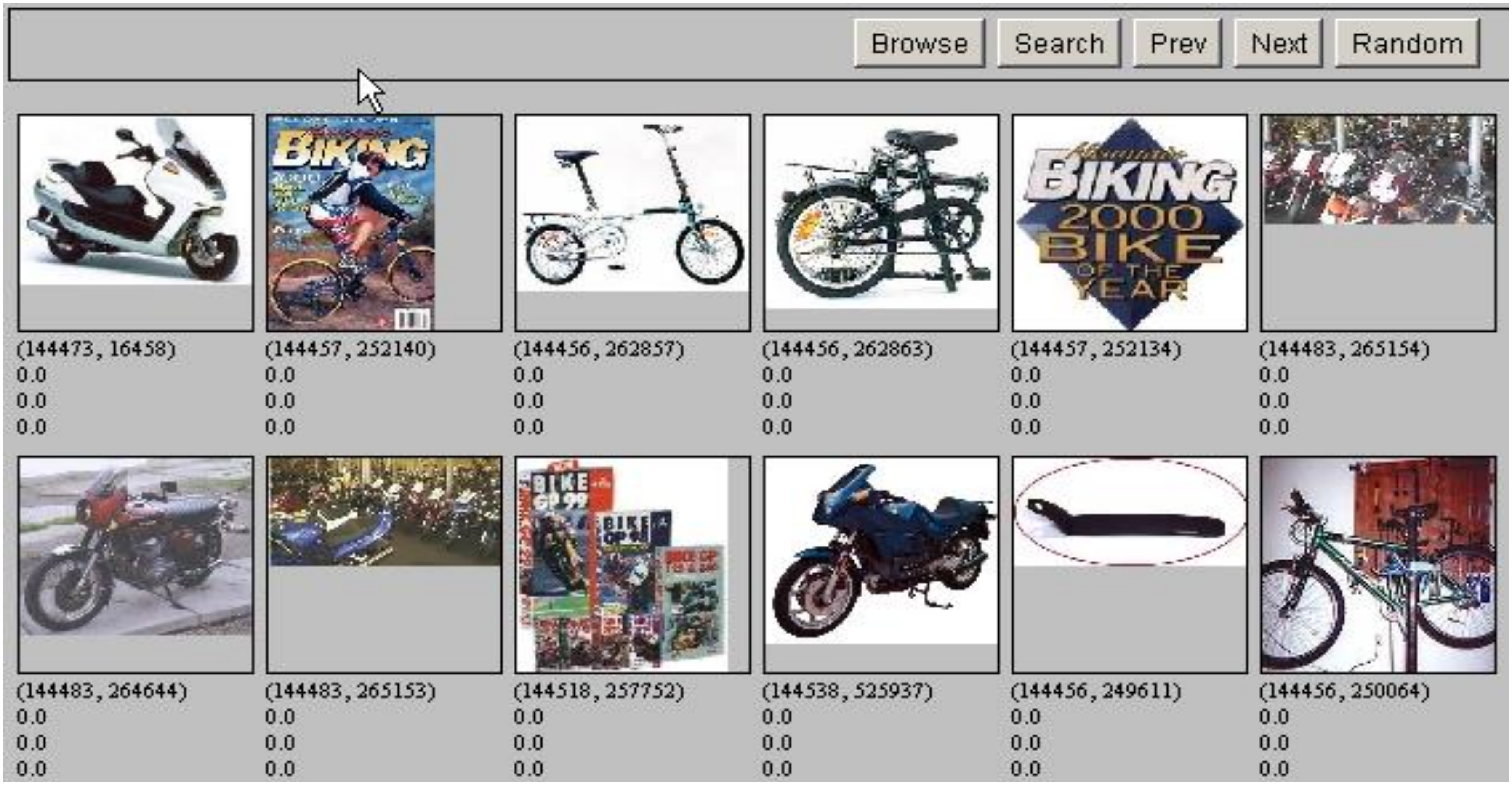


Relevance Feedback: Example

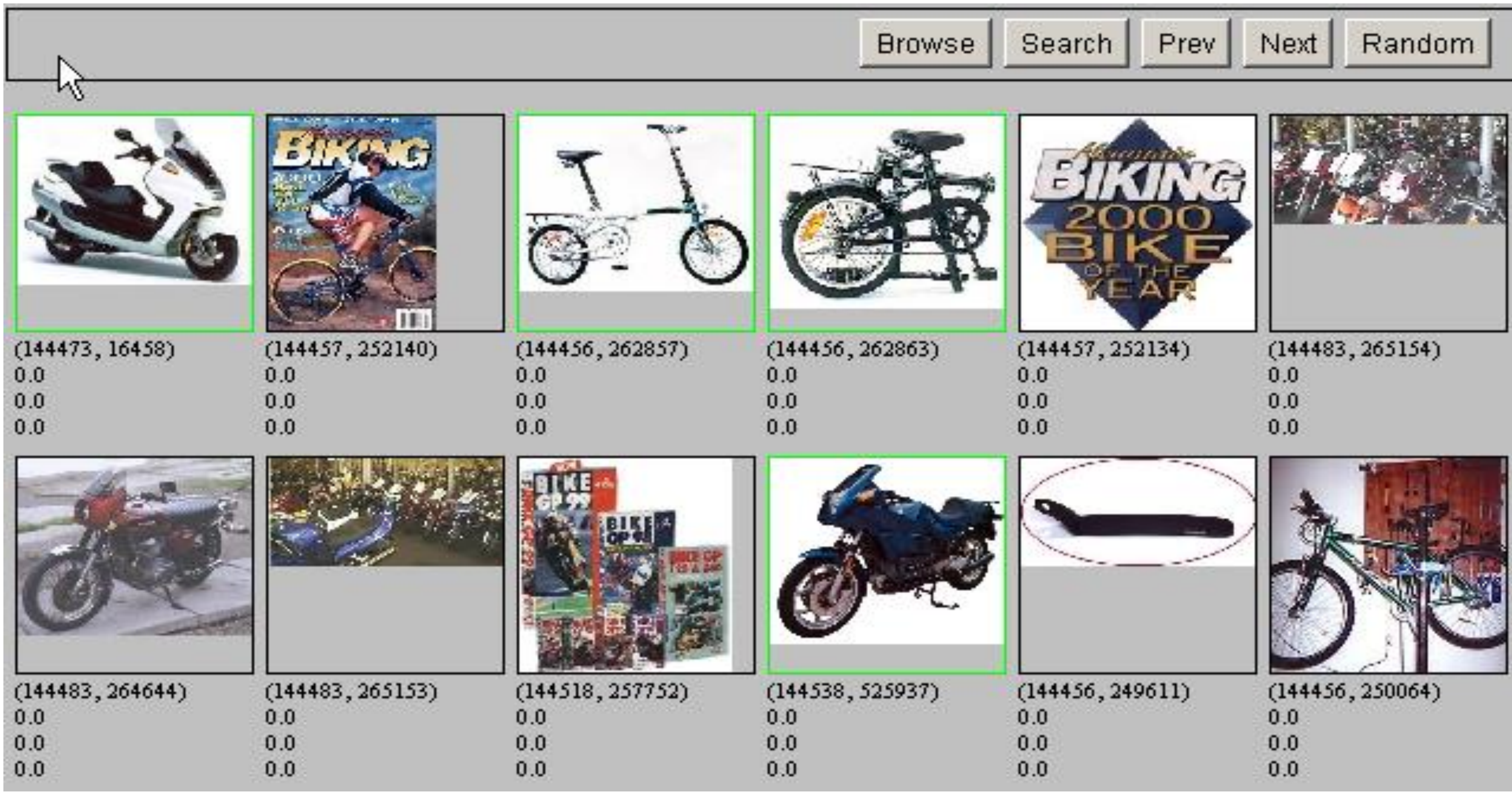
- Image search engine http://nayana.ece.ucsb.edu/imsearch/imsearch.html



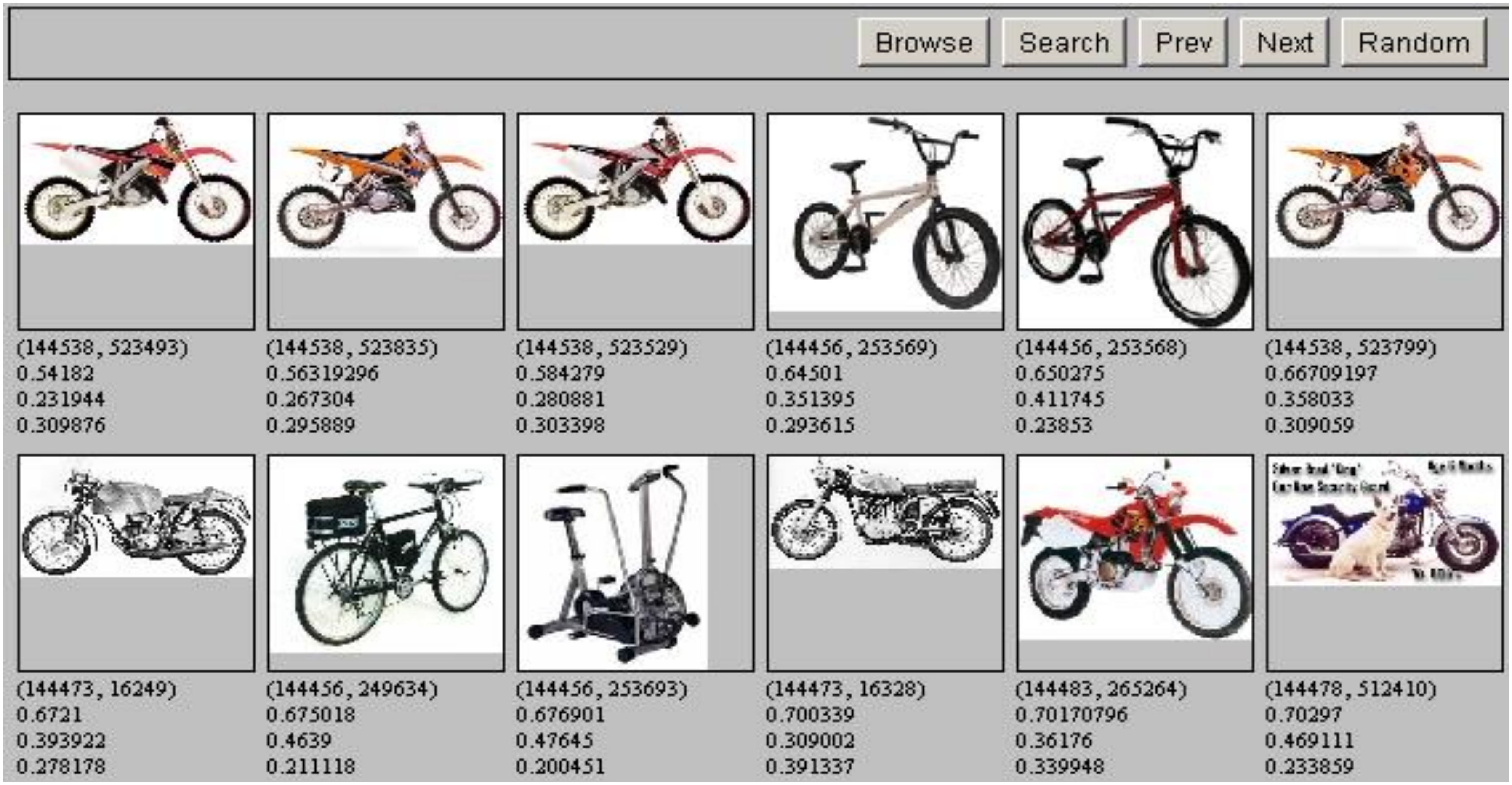
Results for Initial Query



Relevance Feedback



Results after Relevance Feedback



Initial query/results

- Initial query: New space satellite applications

+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes

4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget

5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research

6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate

7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada

+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

- User then marks relevant documents with "+".

Expanded query after relevance feedback

- 2.074 new
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil

15.106 space

5.660 application

5.196 eos

3.972 aster

3.446 arianespace

2.806 ss

2.053 scientist

1.172 earth

0.646 measure

Results for expanded query

2 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

1 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

8 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Key concept: Centroid

- The centroid is the center of mass of a set of points
- Recall that we represent documents as points in a high-dimensional space
- Definition: Centroid

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

where D_c is a set of documents in class c and $\vec{v}(d)$ is the vector representation of document d.

Rocchio Algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance feedback query
- Rocchio seeks the query q_{opt} that maximizes

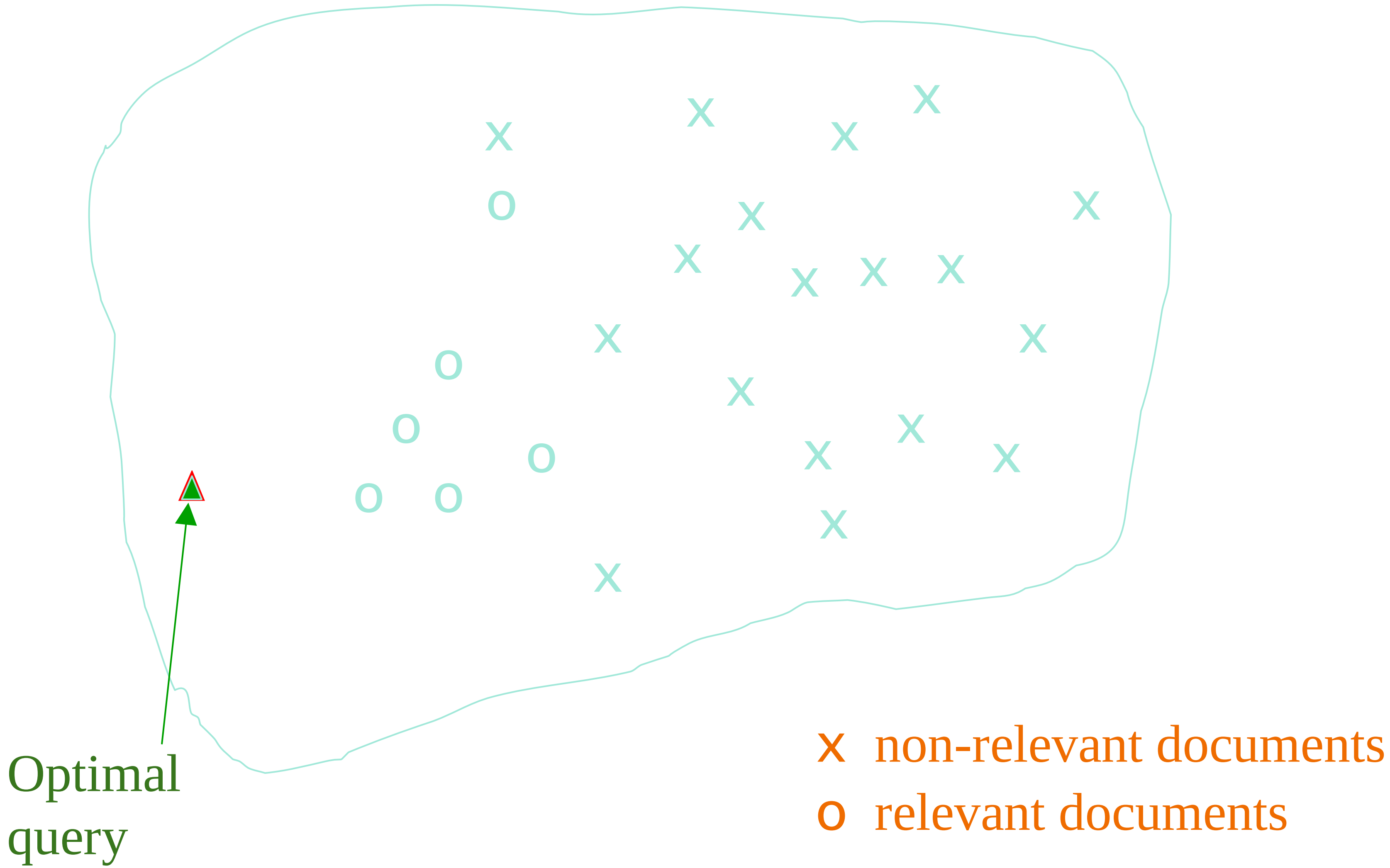
$$\vec{q}_{opt} = \arg\max_{\vec{q}} [\text{sim}(\vec{q}, C_r) - \text{sim}(\vec{q}, C_{nr})],$$

- Tries to separate docs marked relevant and non-relevant

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \in C_{nr}} \vec{d}_j$$

- Problem: we don't know the truly relevant docs

The Theoretically Best Query



Rocchio 1971 Algorithm (SMART)

- Used in practice:

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

- D_r = set of known relevant doc vectors
- D_{nr} = set of known irrelevant doc vectors
  - Different from C_r and C_{nr} !
- q_m = modified query vector; q_0 = original query vector; α,β,γ: weights (hand-chosen or set empirically)
- New query moves toward relevant documents and away from irrelevant documents

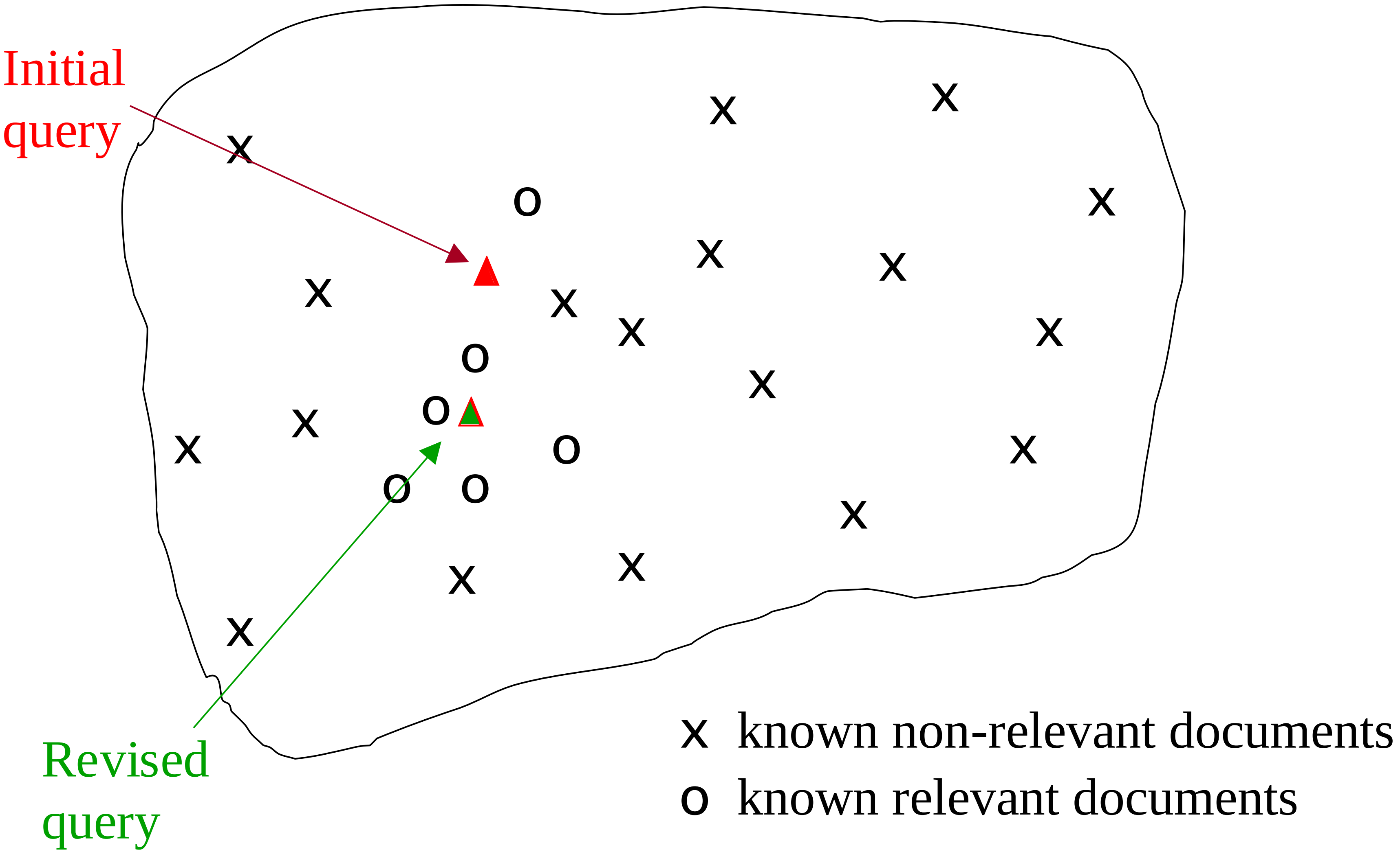
Subtleties to note

- Tradeoff α vs. β/γ: If we have a lot of judged documents, we want a higher β/γ.
- Some weights in query vector can go negative
  - Negative term weights are ignored (set to 0)

Positive vs Negative Feedback

- Positive feedback is more valuable than negative feedback (so, set $\gamma < \beta$; e.g. $\gamma = 0.25$, $\beta = 0.75$).
- Many systems only allow positive feedback (γ=0).

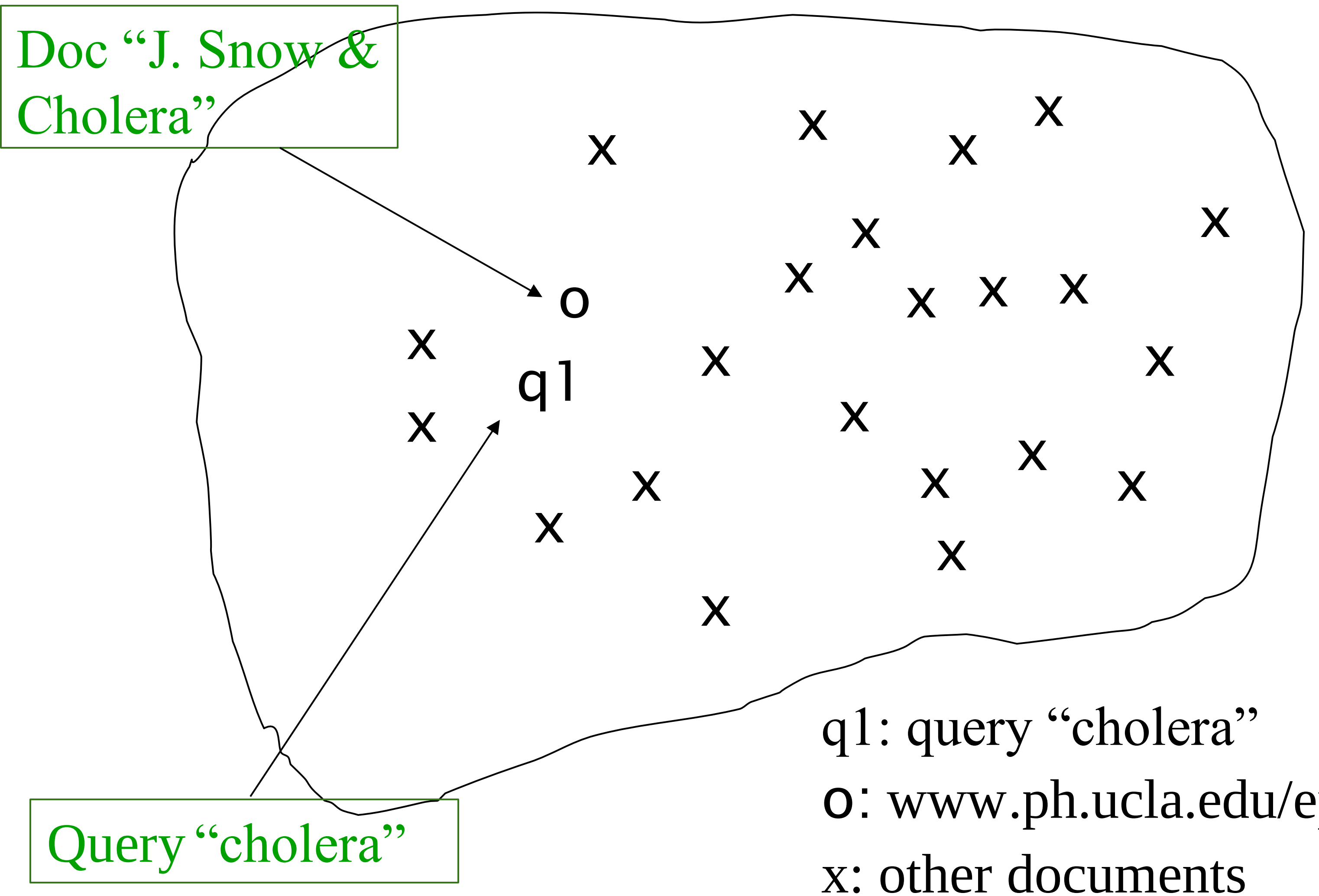
Relevance feedback on initial query



Relevance Feedback in vector spaces

- We can modify the query based on relevance feedback and apply standard vector space model.
- Use only the docs that were marked for relevance feedback.
- Relevance feedback can potentially improve recall and precision
- However, relevance feedback 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

Aside: Vector Space can be Counterintuitive.



q1: query “cholera”

o: www.ph.ucla.edu/epi/snow.html

x: other documents

High-dimensional Vector Spaces

- The queries “cholera” and “john snow” are far from each other in vector space.
- How can the document “John Snow and Cholera” be close to both of them?
- Our intuitions for 2- and 3-dimensional space don't work in >10,000 dimensions.
- high dimensions: If a document is close to many queries, then some of these queries must be close to each other.

Relevance feedback: Assumptions

- When can relevance feedback enhance recall?
- Assumption A1: The user knows the terms in the collection well enough for an initial query.
- Assumption A2: Relevant documents contain similar terms (so I can “hop” from one relevant document to a different one when giving relevance feedback).

22

Violation of A1

- Assumption A1: The user knows the terms in the collection well enough for an initial query.
- Violation: Mismatch of searcher's vocabulary and collection vocabulary
- Example: cosmonaut / astronaut

23

Violation of A2

- Assumption A2: Relevant documents are similar.
- Example for violation: [contradictory government policies]
- Several unrelated “prototypes”
  - Subsidies for tobacco farmers vs. high tariffs on imports from developing countries
- Relevance feedback on tobacco docs will not help with finding docs on developing countries.

24

Relevance Feedback: Problems

- Long queries are inefficient for typical IR engine.
  - Long response times for user.
  - High cost for retrieval system.
  - Partial solution:
    - Only reweight certain prominent terms
      - Perhaps top 20 by term frequency
- Users are often reluctant to provide explicit feedback
- It's often harder to understand why a particular document was retrieved after applying relevance feedback (so only recall is considered)

Why?

Evaluation of relevance feedback strategies

- Use q_0 and compute precision and recall graph
- Use q_m and compute precision recall graph
  - Assess on all documents in the collection
    - Spectacular improvements, but … it’s cheating!
    - Partly due to known relevant documents ranked higher
    - Must evaluate with respect to documents not seen by user

Initial query: New space satellite applications

1. **0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer**
2. **0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan**
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
8. **0.509, 12/02/87, Telecommunications Tale of Two Companies**

Revised Query Results

1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Evaluation of relevance feedback

- Second method – assess only the docs not rated by the user in the first round
  - Measures usually then lower than for original query
  - Could make relevance feedback look worse than it really is
  - Can still assess relative performance of algorithms
- Most satisfactory – use two collections each with their own relevance assessments
  - q_0 and user feedback from first collection
  - q_m run on second collection and measured

Empirically, one round of relevance feedback is often very useful. Two rounds is sometimes marginally useful.

Initial query: New space satellite applications

1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer

2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan

3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes

4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget

5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research

6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate

7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada

8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

Revised Query Results

~~1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan [~~

~~2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer~~

3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own

4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

~~5. 0.492, 12/02/87, Telecommunications Tale of Two Companies~~

6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use

7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers

8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Evaluation of relevance feedback

- Method I– assess only the docs not rated by the user in the first round
  - Measures usually then lower than for original query
  - Could make relevance feedback look worse than it really is
  - Can still assess relative performance of algorithms
- Method- II – use two collections each with their own relevance assessments
  - q_0 and user feedback from first collection D_1 = <d_{11},d_{12}.........d_{1n}>
  - q_m run on second collection and measured D_2 = <d_{21},d_{22}.........d_{2m}>
  - D_1 \inter D_2 = \phi

Empirically, one round of relevance feedback is often very useful. Two rounds is sometimes marginally useful.

Evaluation: Caveat

- True evaluation of usefulness must compare to other methods taking the same amount of time.
- Alternative to relevance feedback: User revises and resubmits query.
- Users may prefer revision/resubmission to having to judge relevance of documents.
- There is no clear evidence that relevance feedback is the “best use” of the user’s time.

Relevance Feedback on the Web

- Some search engines offer a similar/related pages feature (this is a trivial form of relevance feedback)
  - Google (link-based)
  - Altavista
  - Stanford WebBase

$\alpha/\beta/\gamma$??

- But some don’t because it’s hard to explain to average user:
  - Alltheweb
  - bing
  - Yahoo
- Excite initially had true relevance feedback, but abandoned it due to lack of use.

Excite Relevance Feedback

Spink et al. 2000

- Only about 4% of query sessions from a user used relevance feedback option
  - Expressed as “More like this” link next to each result
- But about 70% of users only looked at first page of results and didn’t pursue things further
  - So 4% is about 1/8 of people extending search
- Relevance feedback improved results about 2/3 of the time

Explicit Feedback is largely not in use now

Pseudo relevance feedback (Blind Feedback)

- Pseudo-relevance feedback automates the “manual” part of true relevance feedback.
- Pseudo-relevance algorithm:
  - Retrieve a ranked list of hits for the user’s query
  - Assume that the top k documents are relevant.
  - Do relevance feedback (e.g., Rocchio)
- Works very well on average
- But can go horribly wrong for some queries.
- Several iterations can cause query drift. (Why?)

Example Bank of London River Bank in London – River Sea boat

Implicit Feedback

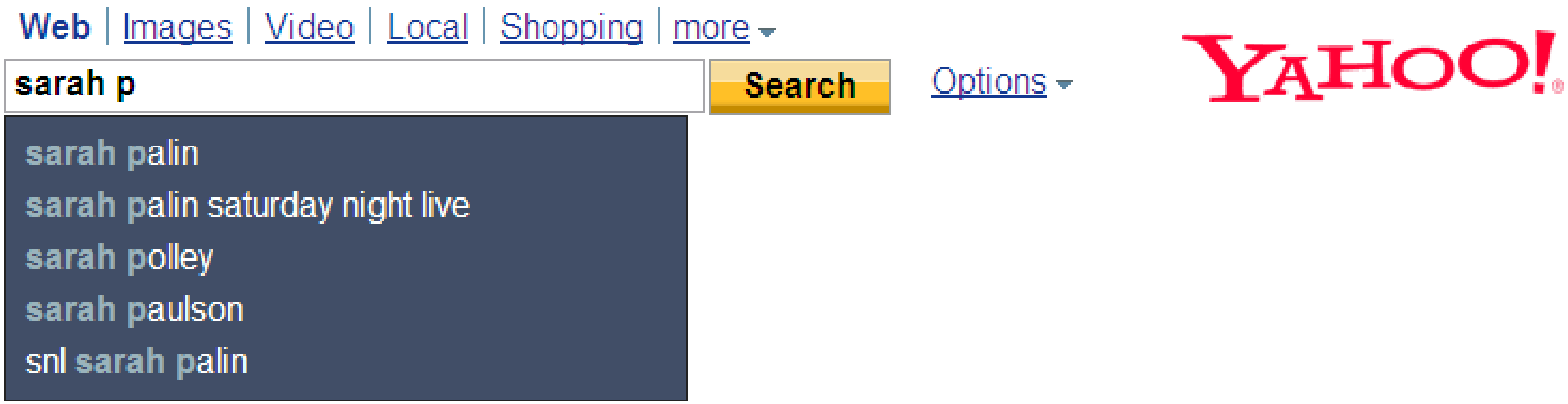
- User behavior
  - documents they do and do not select for viewing,
  - the duration of time spent viewing a document, (dwell time)
  - page browsing or scrolling actions
  - Clickstream mining
- User is not necessarily informed that their behavior will be used as relevance feedback

Query Expansion

Query Expansion

- In relevance feedback, users give additional input (relevant/non-relevant) on **documents**, which is used to reweight terms in the documents
- In query expansion, users give additional input (good/bad search term) on **words or phrases**
- **Query Log Analysis**

Query assist



Query Expansion

- Users give additional terms
- Search Engine suggest related queries in response to a query

How do we augment the user query?

- Manual thesaurus
  - UMLS (United Medical Language System) – canonical term for each concept. Cancer neoplasm.
  - May not have a canonical term like normal thesaurus

39

Thesaurus-based query expansion

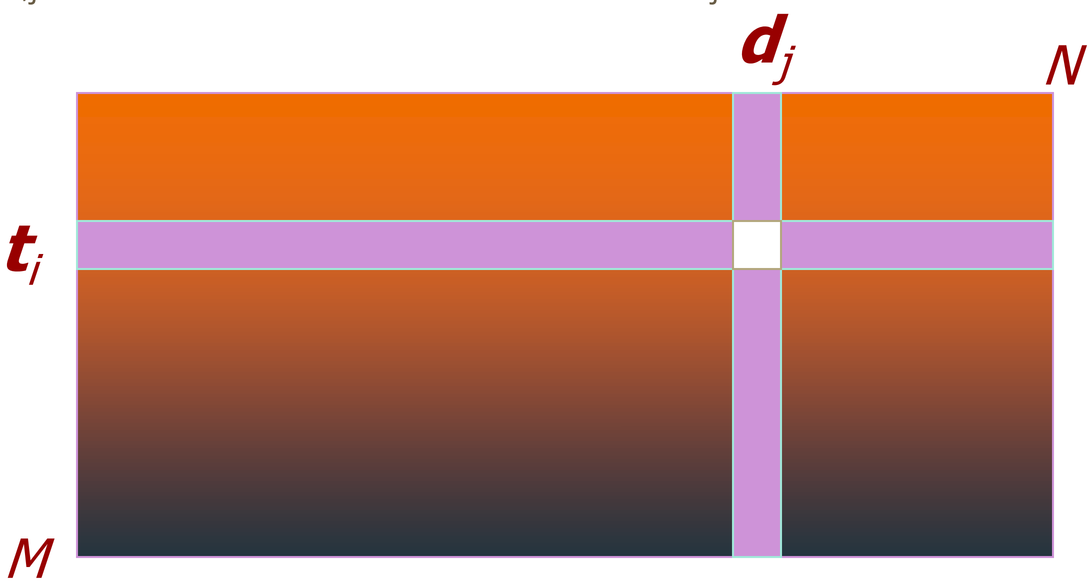
- For each term, t, in a query, expand the query with synonyms and related words of t from the thesaurus
  - feline → feline cat
- May weight added terms less than original query terms.
- Generally increases recall
- Widely used **in many science/engineering fields – useful in obtaining domain specific keywords**
  - There is a high cost of manually producing and updating a scientific thesaurus
- May significantly decrease precision, particularly with ambiguous terms.
  - “interest rate” → “interest rate fascinate evaluate”

Automatic Thesaurus Generation

- Attempt to generate a thesaurus automatically by analyzing the collection of documents
- Fundamental notion: similarity between two words
  - Definition 1: Two words are similar if they co-occur with similar words.
  - Definition 2: Two words are similar if they occur in a given grammatical relation with the same words.
- You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.

Co-occurrence Thesaurus

- Simplest way to compute one is based on term-term similarities in $C = AA^T$ where A is term-document matrix.
- $w_{i,j}$ = (normalized) weight for (t_i, d_j)



What does *C* contain if *A* is a term-doc incidence (0/1) matrix?

- For each t_i, pick terms with high values in C

$$
\mathrm{A} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & D_3 & D_4 \end{array} \\ \begin{pmatrix} 1 & 5 & 2 & 3 \\ 4 & 0 & 1 & 1 \\ 2 & 2 & 2 & 0 \end{pmatrix} \end{array}
$$

$$
A^T = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} 1 & 4 & 2 \\ 5 & 0 & 2 \\ 2 & 1 & 2 \\ 3 & 1 & 0 \end{pmatrix} \end{array}
$$

$$
AA^T = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} \mathbf{39} & 9 & 16 \\ 9 & 18 & 10 \\ 16 & 10 & 12 \end{pmatrix} \end{array}
$$

Automatic Thesaurus Generation 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 gazec
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 parasite
senses	grasp psyche truly clumsy naive innate awk

Automatic Thesaurus Generation: Discussion

- Quality of associations is usually a problem.
- Term ambiguity may introduce irrelevant statistically correlated terms.
    - “Apple computer” → “Apple red fruit computer”
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

Query reformulation

- Query reformulations based on query log mining.
- The manual query reformulations of other users to make suggestions to a new user.
- This requires a huge query volume, and is thus particularly appropriate to web search

Most effective – Query assist and reformulation

Thank You

Questions?