CIS 833 – Information Retrieval and Text Mining

Lecture 19

Link Analysis

November 5, 2015

Credits for slides: Allan, Arms, Manning, Lund, Noble, Page.

Next

- Web Search
  - Textbook Chapter 21 – Web analysis

Connectivity-Based Ranking

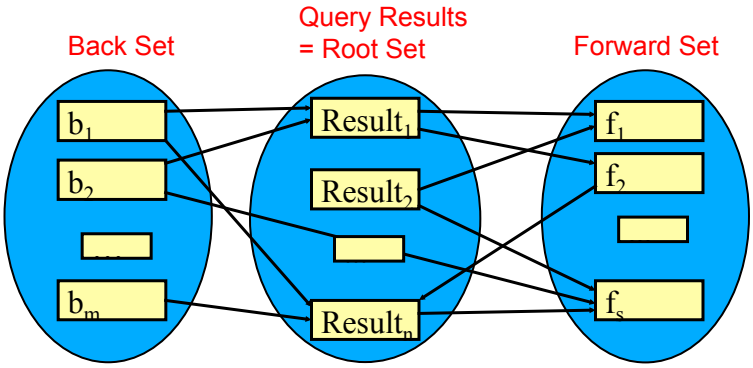
Ranking based on hyperlink analysis

- Query-independent ranking
  - PageRank: authorities
- Query-dependent ranking
  - HITS: authorities and hubs

- *Authorities* are pages that are recognized as providing significant, trustworthy, and useful information on a topic.
- *Hubs* are index pages that provide lots of useful links to relevant content pages (topic authorities).

Neighborhood Graph (Base Set)

- Subgraph associated to each query



HITS Iterative Algorithm

Initialize for all $p \in S$: $a_p = h_p = 1$

For i = 1 to k:

For all $p \in S$: $a_p = \sum_{q:q \to p} h_q$ (*update auth. scores*)

For all $p \in S$: $h_p = \sum_{q:p \to q} a_q$ (*update hub scores*)

For all $p \in S$: $a_p = a_p / c$ c: $\sum_{p \in S} \left(a_p / c\right)^2 = 1$ (*normalize* **a**)

For all $p \in S$: $h_p = h_p / c$ c: $\sum_{p \in S} \left(h_p / c\right)^2 = 1$ (*normalize* **h**)

Convergence

- Algorithm converges to a *fix-point* if iterated indefinitely.
- Define *A* to be the adjacency matrix for the subgraph defined by *S.*
  - $A_{ij} = 1$ for $i \in S$, $j \in S$ iff $i \to j$
- Authority vector, ***a***, converges to the principal eigenvector of $A^T A$
- Hub vector, ***h***, converges to the principal eigenvector of AA^T
- In practice, 20 iterations produce fairly stable results.

Japan Elementary Schools

Hubs

- schools
- LINK Page-13
- “ú–{‚ÌŠw Z
- a‰„ ¬Šw Zƒz [ƒ ƒy [ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ‚l‚f‚j ¬Šw Z‚U”N‚P ‘g•¨Œê
- ÒŠ—’¬—§ ÒŠ—“Œ ¬Šw Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y“ì ¬Šw Z‚Ìƒz [ƒ ƒy [ƒW
- UNIVERSITY
- ‰J—³ ¬Šw Z DRAGON97-TOP
- Â‰ª ¬Šw Z‚T”N‚P ‘gƒz [ƒ ƒy [ƒW
- ¶µ°é¼ÂÁ© ¥á¥Ë¥â¡¼ ¥á¥Ë¥â¡¼

Authorities

- The American School in Japan
- The Link Page
- ‰ª è s—§ˆä“c ¬Šw Zƒz [ƒ ƒy [ƒW
- Kids' Space
- ˆÀ é s—§ˆÀ é ¼•” ¬Šw Z
- ‹{ é‹³ç ‘âŠw• ‘® ¬Šw Z
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- _“Þ ìŒ§ E‰¡•ì s—§’† ì ¼ ¬Šw Z‚Ìƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

Things to Note

- Pulls together good pages regardless of language of page content.
- Use *only* link analysis after base set assembled
  - iterative scoring is query-independent.
- Iterative computation only after retrieving the root set based on query - significant overhead.

Results

- Authorities for query: “Java”
  - java.sun.com
  - comp.lang.java FAQ
- Authorities for query “search engine”
  - Google.com
  - Yahoo.com
  - Bing.com
- Authorities for query “Gates”
  - Microsoft.com
  - roadahead.com

Result Comments

- In most cases, the final authorities were not in the initial root set.
- Authorities were brought in from linked and reverse-linked pages and then HITS computed their high authority score.

Finding Similar Pages Using Link Structure

- Given a page, *P*, let *R* (the root set) be *t* (e.g., 200) pages that point to *P*.
- Grow a base set *S* from *R*.
- Run HITS on *S*.
- Return the best authorities in *S* as the best similar-pages for *P*.
- Finds authorities in the "link neighborhood" of *P*.

Similar Page Results

- Given "honda.com"
  - toyota.com
  - ford.com
  - bmwusa.com
  - saturncars.com
  - nissanmotors.com
  - audi.com
  - volvocars.com

Issues

- Topic Drift
  - Off-topic pages can cause off-topic "authorities" to be returned
    - E.g., the neighborhood graph can be about a "super topic"
- Mutually Reinforcing Affiliates
  - Affiliated pages/sites can boost each others' scores
    - Linkage between affiliated pages is not a useful signal

HITS for Clustering

- An ambiguous query can result in the principal eigenvector only covering one of the possible meanings.
- Non-principal eigenvectors may contain hubs & authorities for other meanings.
- Example: "jaguar":
  - Atari video game (principal eigenvector)
  - NFL Football team (2nd non-princ. eigenvector)
  - Automobile (3rd non-princ. eigenvector)

PageRank vs. HITS

- Computation:
  - Once for all documents and queries (offline)
- Query-independent – requires combination with query-dependent criteria
- Hard(er) to spam

- Computation:
  - Requires computation for each query
- Query-dependent
- Relatively easy to spam
- Quality depends on quality of start set
- Gives hubs as well as authorities

Relevance Feedback

Where we are heading next…

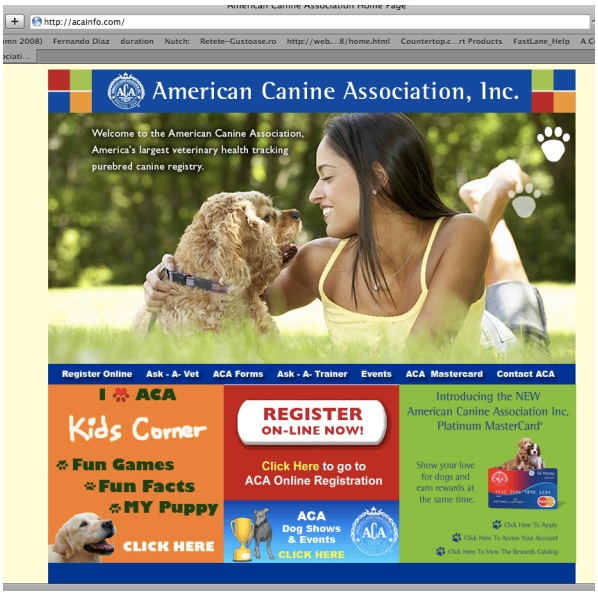
- Improving IR results - recall
  - E.g., searching for *canine* doesn't match with *dog.*
- Options for improving results…
  - The complete landscape
    - Local methods
      - Relevance feedback
      - Pseudo relevance feedback
    - Global methods
      - Query expansion
        - Thesauri
        - Automatic thesaurus generation
    - Focus on relevance feedback first (Ch. 9 textbook)

First, some intuition

Suppose you search for "canine".

Score = 10

Term	Freq.
Canine	10
Dog	0
Cat	0



Score = 6

Term	Freq.
Canine	6
Dog	0
Cat	0

Dog

From Wikipedia, the free encyclopedia

This article is about the domestic dog. For other uses, see Dog (disambiguation).

The **dog** (*Canis lupus familiaris*)[2] is a domesticated subspecies of the gray wolf, a mammal of the Canidae family of the order Carnivora. The term encompasses both feral and pet varieties and is also sometimes used to describe wild canids of other subspecies or species. The domestic dog has been one of the most widely kept working and companion animals in human history, as well as being a food source in some cultures. There are estimated to be 400 million dogs in the world.[3]

The dog has developed into hundreds of varied breeds. Height measured to the withers ranges from a few inches in the Chihuahua to a few feet in the Irish Wolfhound; color varies from white through grays (usually called *blue*) to black, and browns from light (tan) to dark ("red" or "chocolate") in a wide variation of patterns; and, coats can be very short to many centimeters long, from coarse hair to something akin to wool, straight or curly, or smooth.[4] It is common for most breeds to shed this coat, however some non-shedding breeds are also popular.

Contents [hide]

1 Etymology and related terminology
2 Taxonomy
3 Origin and evolution
3.1 Origins
3.2 Ancestry and history of domestication
3.3 Development of dog breeds
3.3.1 Breed popularity
4 Differences from wolves
4.1 Physical characteristics
4.2 Behaviour
4.3 Trainibility
5 Physical characteristics
5.1 Sight
5.2 Hearing
5.3 Smell
5.4 Coat color
5.5 Tail
6 Behavior and intelligence
6.1 Intelligence

Domestic dog
Fossil range: Late Pleistocene – Recent

A Labrador Retriever

Conservation status: Domesticated

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Carnivora
Family:	Canidae
Genus:	*Canis*
Species:	*C. lupus*
Subspecies:	***C. l. familiaris***

Score = 2

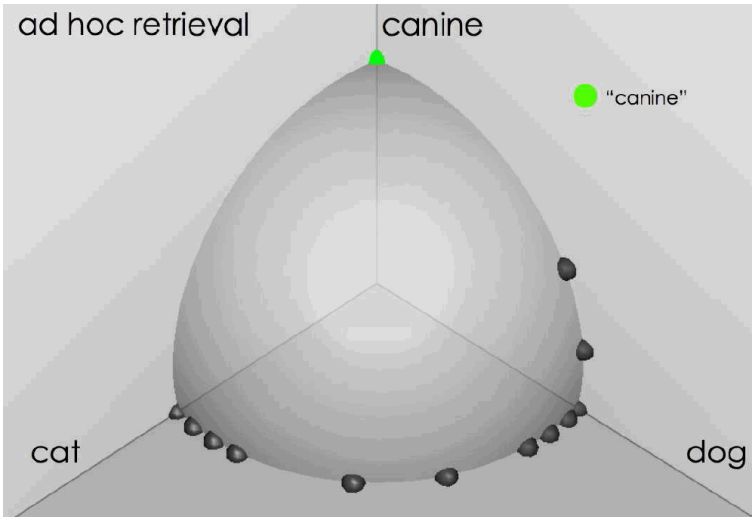
Term	Freq.
Canine	2
Dog	10
Cat	2

Score = 0

Term	Freq.
Canine	0
Dog	10
Cat	2

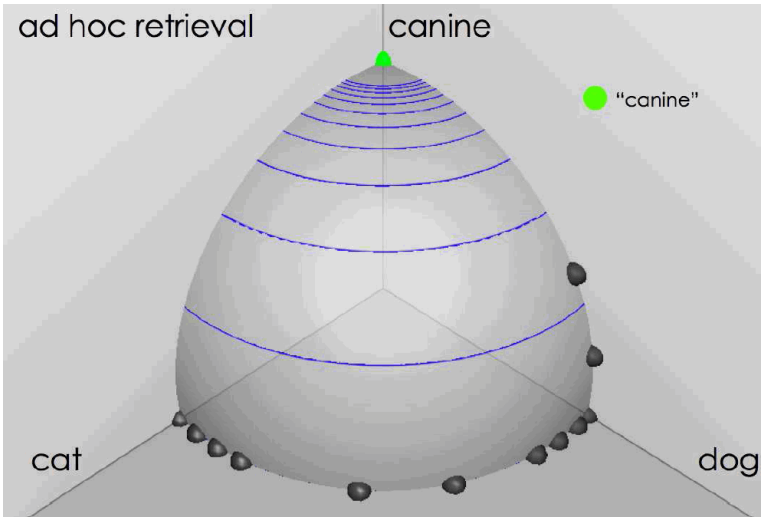
Ad hoc results for query *canine*

source: Fernando Diaz



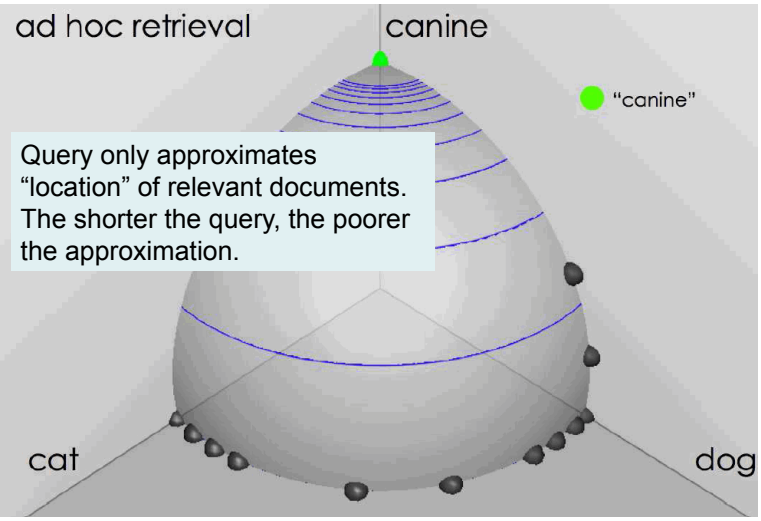
Ad hoc results for query *canine*

source: Fernando Diaz



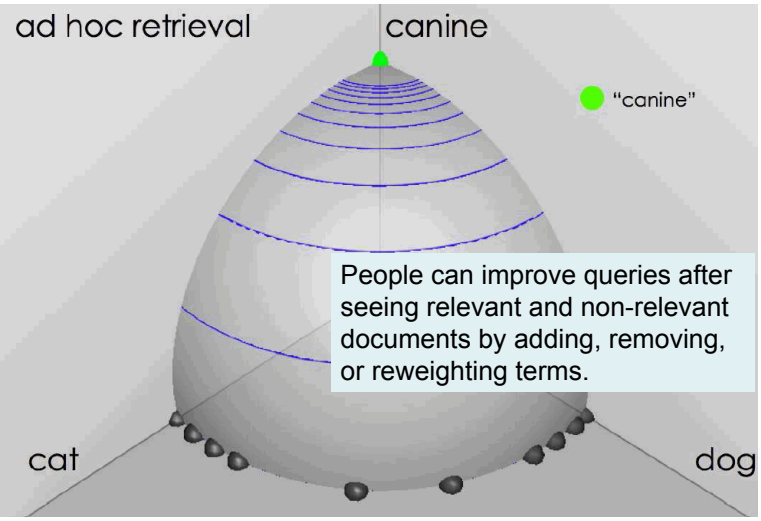
Ad hoc results for query *canine*

source: Fernando Diaz



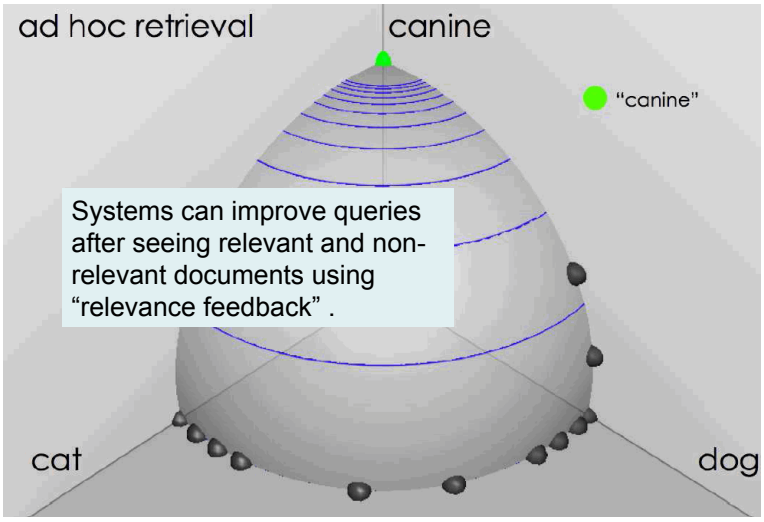
Ad hoc results for query *canine*

source: Fernando Diaz



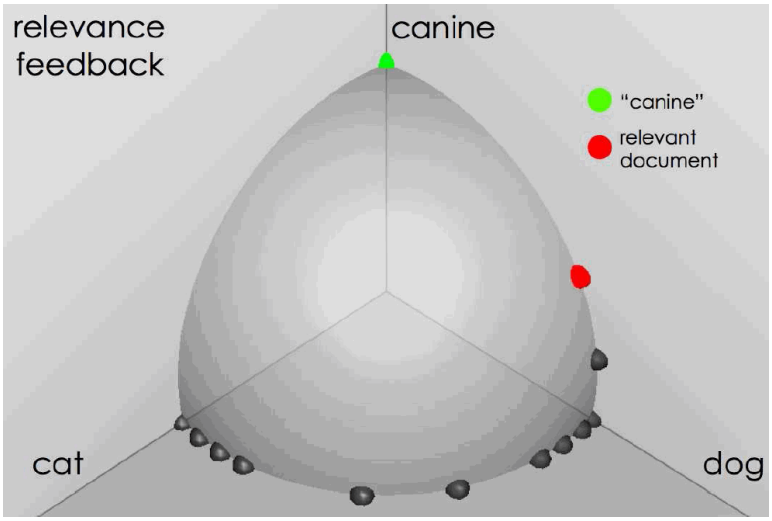
Ad hoc results for query *canine*

source: Fernando Diaz



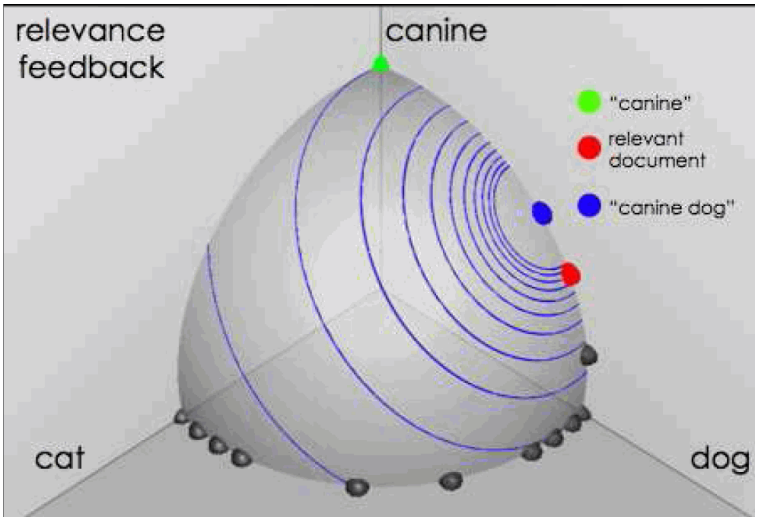
User feedback: Select what is relevant

source: Fernando Diaz



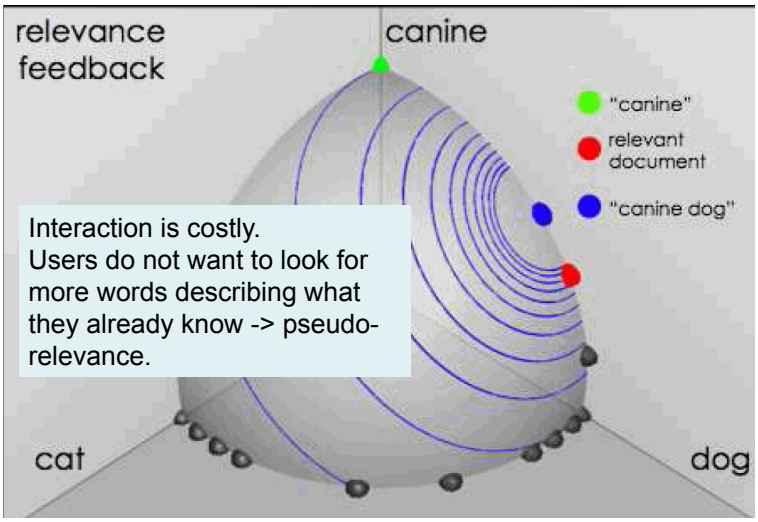
Results after relevance feedback

source: Fernando Diaz



Results after relevance feedback

source: Fernando Diaz



Interaction is costly.
Users do not want to look for more words describing what they already know -> pseudo-relevance.

Big Picture

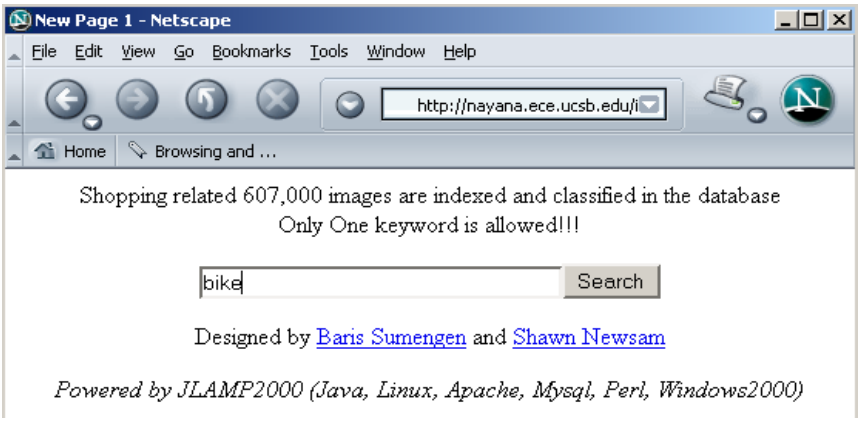
- Relevance feedback
  - Adjust query with direct interaction
    - User looks at returned list of documents and provides feedback
    - System returns a revised ranked list
  - Adjust query with indirect interaction
    - By observing what the user looks at.
    - System returns a revised ranked list
  - Can a better query be created automatically by analyzing relevant and non-relevant documents?
- Pseudo-relevance feedback
  - Adjust query without interaction
    - Generate ranked list but do not present it
    - Use information to create a new ranked list that *is* presented
  - Can a better query be created automatically by assuming that some documents are relevant?

Relevance Feedback

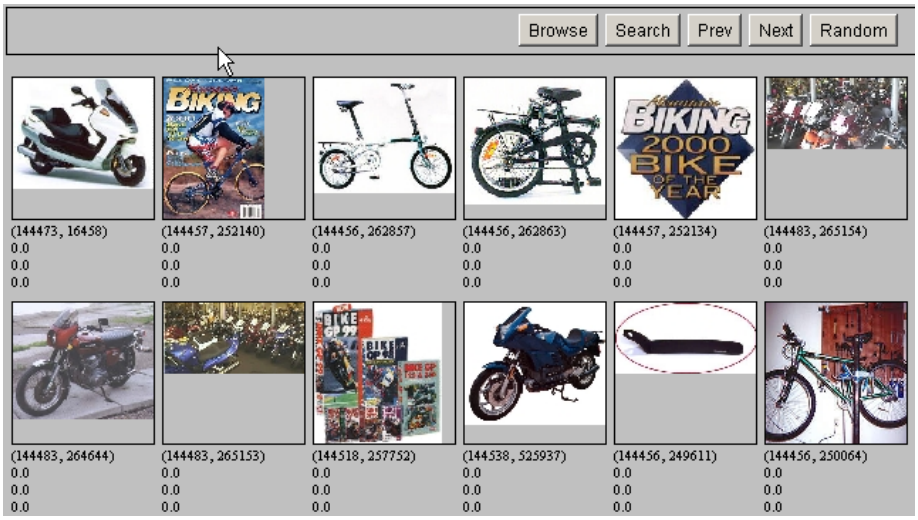
- User feedback on relevance of documents in initial set of results
- Occurs in different models
  - Vector space is used most often (we'll focus on it)
  - Probabilistic models
- How it works:
  - User issues a (short, simple) query, a set of results is retrieved
  - User marks some results as relevant 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.

Relevance Feedback: Example

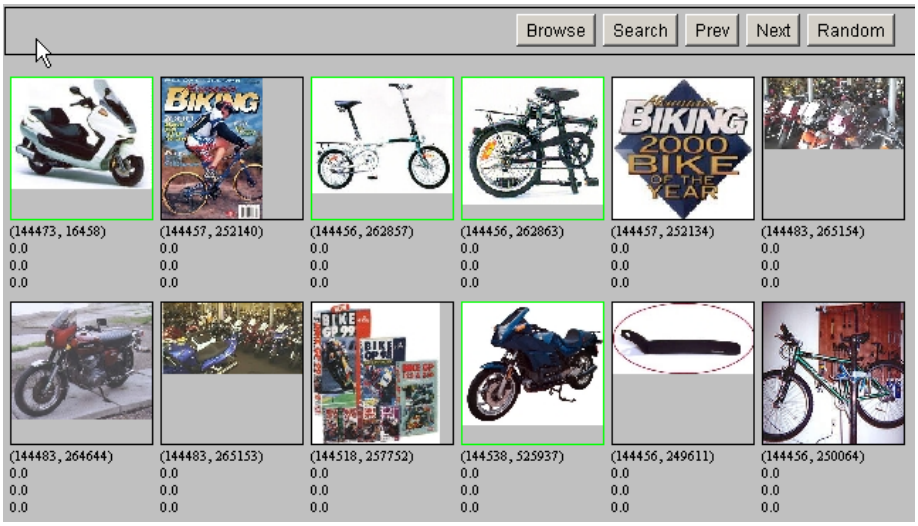
- Image search engine



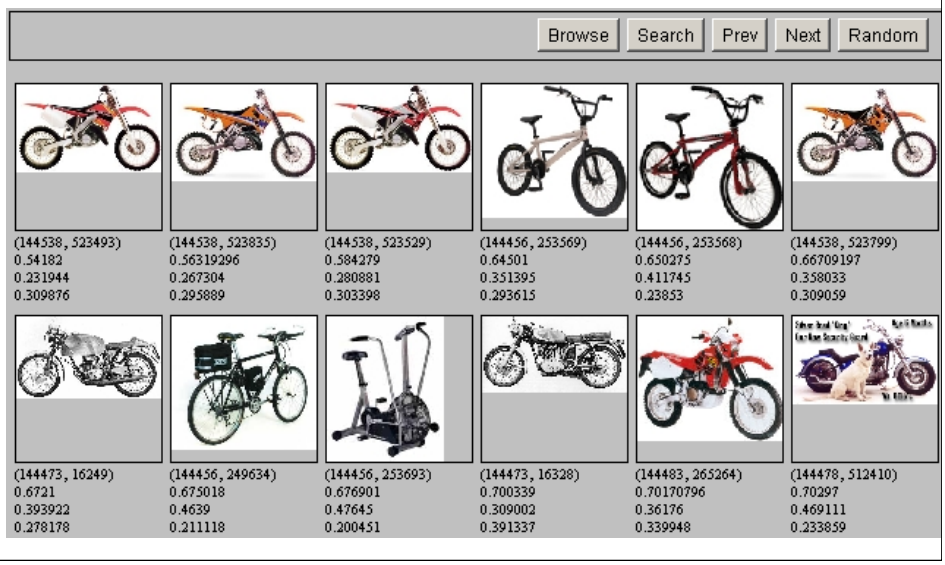
Results for Initial Query



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

Relevance Feedback 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

$$\vec{\mu}(C) = \frac{1}{|C|}\sum_{d \in C}\vec{d}$$

where C is a set of documents.

Rocchio Algorithm

- The Rocchio algorithm uses the vector space model to pick a relevance fed-back query
- Rocchio seeks the query $\vec{q}_{opt}$ that maximizes

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

Rocchio Algorithm

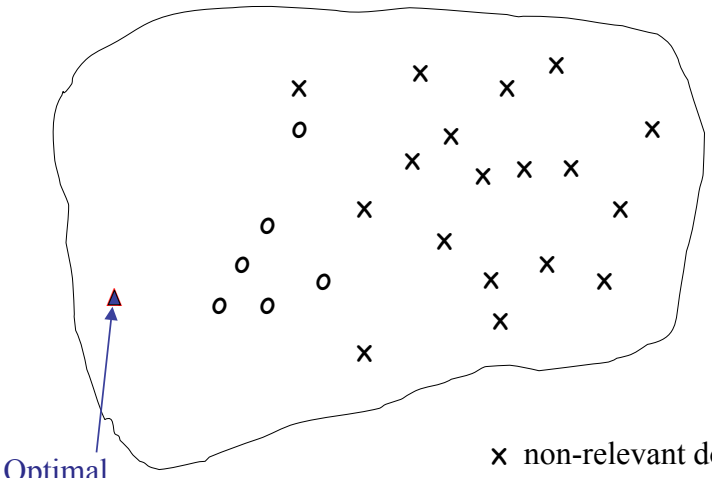
- The Rocchio algorithm uses the vector space model to pick a relevance fed-back query
- Rocchio seeks the query $\vec{q}_{opt}$ that maximizes

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

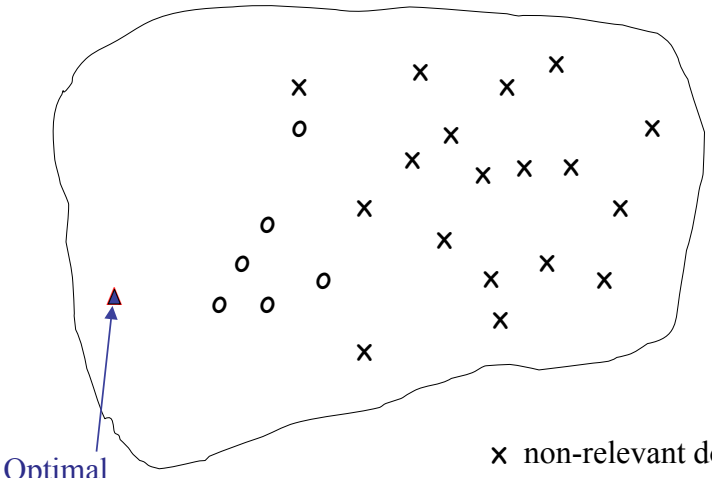
- Tries to separate docs marked relevant and non-relevant

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

The Theoretically Best Query

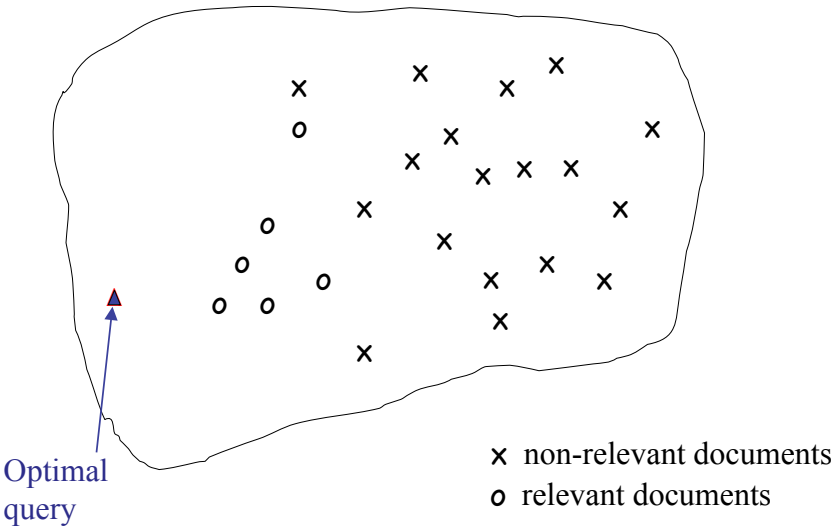


The Theoretically Best Query



Problem:?

The Theoretically Best Query



Problem: we don't know the truly relevant docs

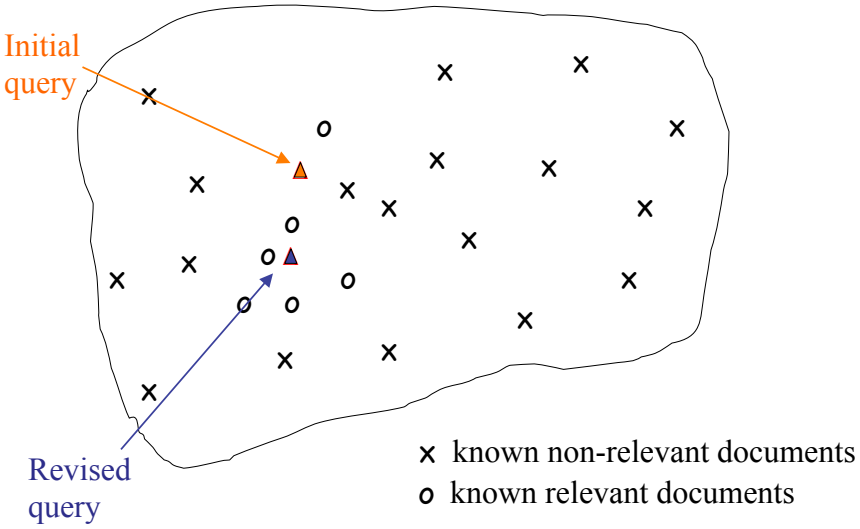
Rocchio 1971 Algorithm (SMART)

- 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)
- **Add** the vectors for **relevant** docs to the query doc
- **Subtract** the vectors for **irrelevant** docs from the query doc
- New query moves toward relevant documents and away from irrelevant documents

Relevance Feedback on Initial Query



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 become negative
  - Negative term weights are ignored (set to 0)

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

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.
- Relevance feedback can improve recall and precision
- 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

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 ($\gamma=0$).

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

Rocchio Variant I: Ide Regular Method

- Since more feedback should perhaps increase the degree of reformulation, do not normalize for amount of feedback:

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

α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for irrelevant documents.

Rocchio Variant II: Ide "Dec Hi" Method

- Bias towards rejecting **just** the highest ranked of the irrelevant documents:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \max_{non-relevant}(\vec{d}_j)$$

α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for irrelevant document.

Comparison of Methods

- Overall, experimental results indicate no clear preference for any one of the specific methods.
- All methods generally improve retrieval performance (recall & precision) with feedback.
- More insights: when does relevance feedback work?

Relevance Feedback: Assumptions

- A1: User has sufficient knowledge for initial query.
- A2: Relevance prototypes are "well-behaved".
  - Term distribution in relevant documents will be similar
  - Term distribution in non-relevant documents will be different from those in relevant documents
    - Either: All relevant documents are tightly clustered around a single prototype.
    - Or: There are different prototypes, but they have significant vocabulary overlap.
    - Similarities between relevant and irrelevant documents are small

Violation of A1

- User does not have sufficient initial knowledge.
- Examples:
  - Misspellings (Brittany Speers).
  - Cross-language information retrieval (hígado).
  - Mismatch of searcher's vocabulary vs. collection vocabulary
    - Cosmonaut/astronaut

Violation of A2

- There are several relevance prototypes.
- Examples:
  - Burma/Myanmar
  - Contradictory government policies
  - Pop stars that worked at Burger King
- Often: instances of a general concept
- Good editorial content can address problem
  - Report on contradictory government policies