CS3245

Information Retrieval

10

Lecture 10: Relevance Feedback and Query Expansion

Last Time

- Evaluating a search engine
  - Benchmarks: 3 components
    Queries, documents and relevance judgments
  - Precision-recall curves
  - Composite, single number summaries
  - A/B Testing

- XML Retrieval – the space between free text retrieval and structured (DB) retrieval
- Matching – Lexicalized Subtrees
  - Structure (Context Similarity)
  - Content (Standard VSM)

Today

Chapter 9

1. Relevance Feedback

   Document Level

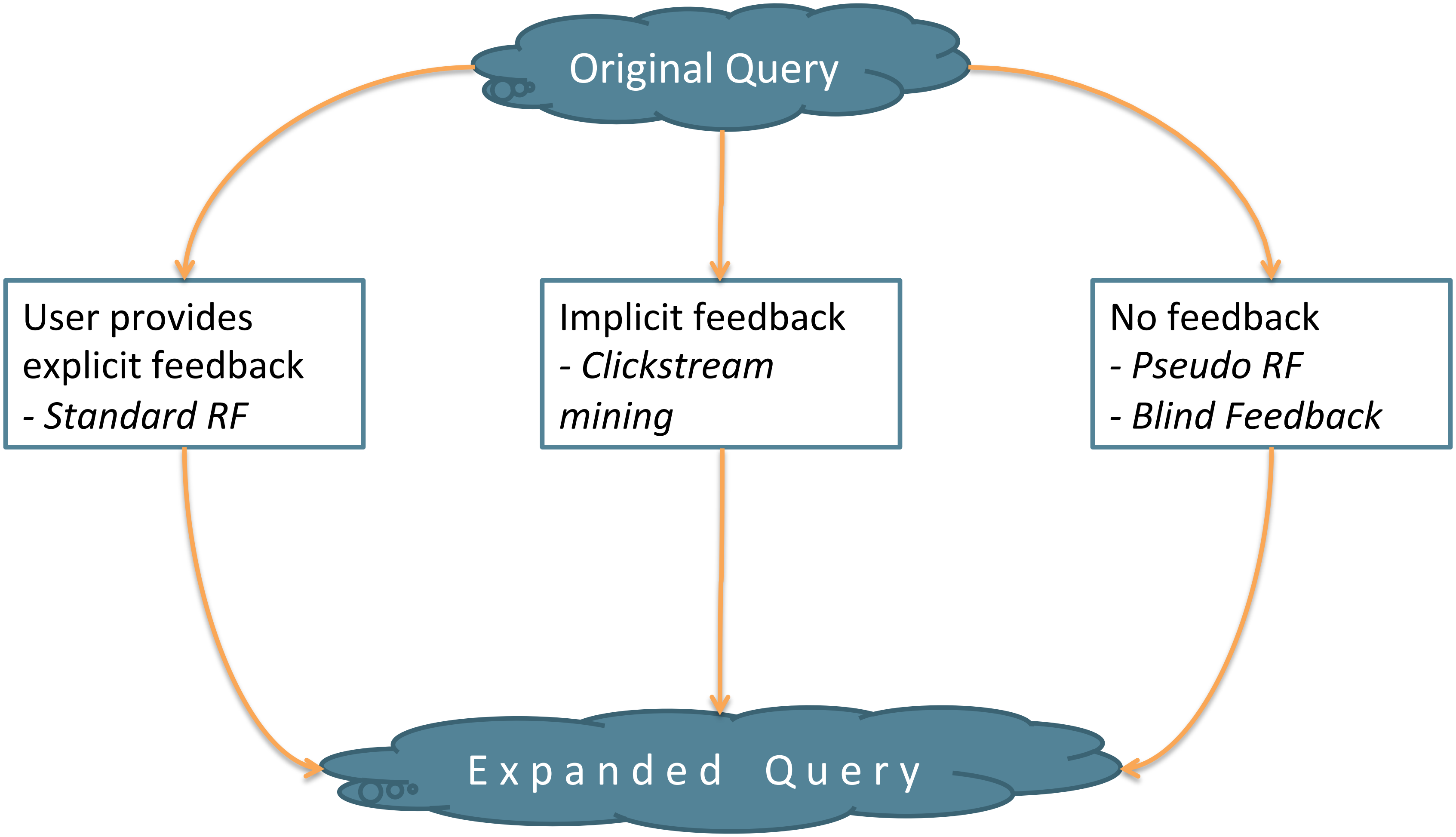
   - Explicit RF – Rocchio (1971)
   - When does it work?
   - Variants – Implicit and Blind

2. Query Expansion

   Term Level

   - Controlled Vocabularies
   - WordNet
   - Automatic Thesaurus Generation

Relevance Feedback



Original Query

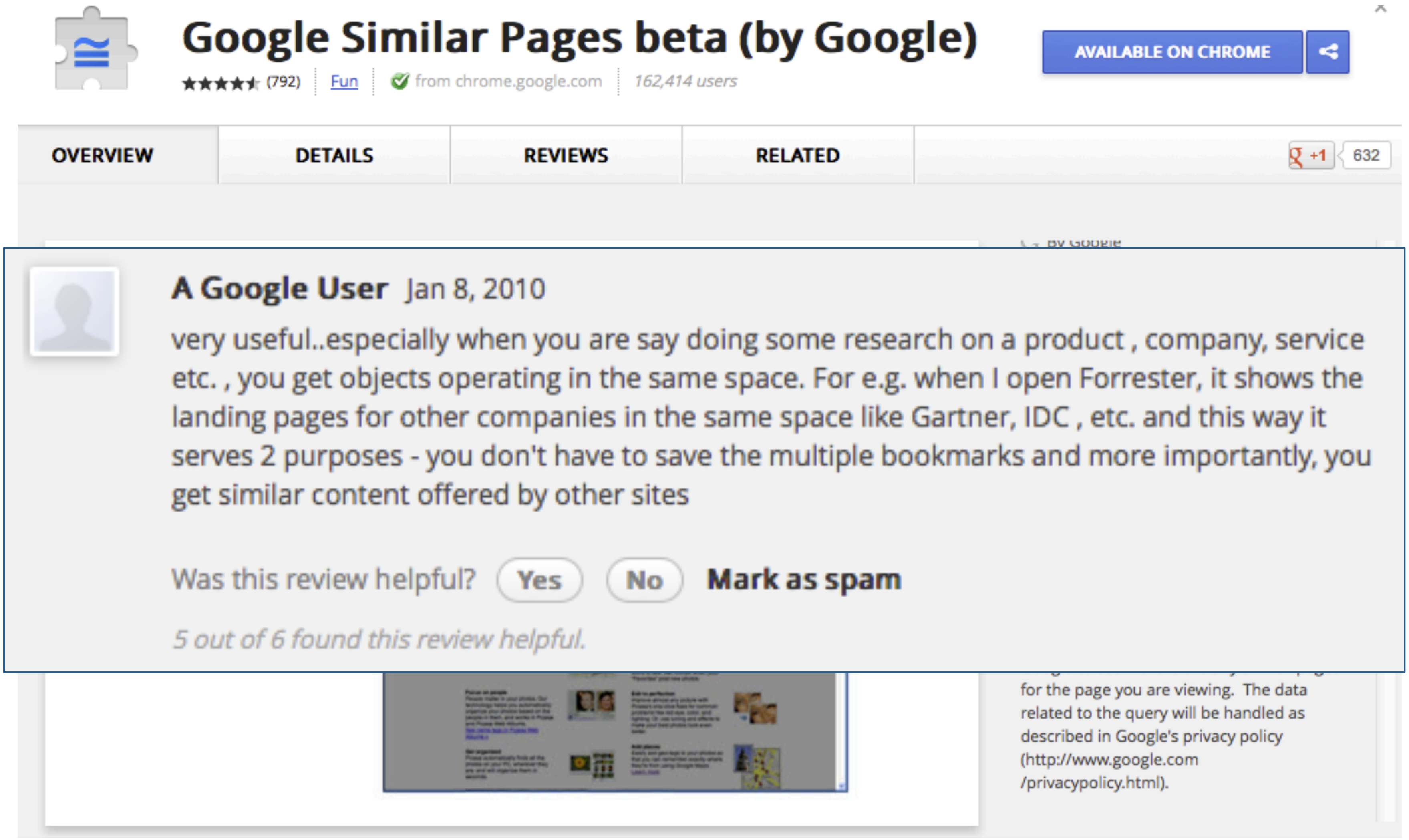
User provides explicit feedback
- *Standard RF*

Implicit feedback
- *Clickstream mining*

No feedback
- *Pseudo RF*
- *Blind Feedback*

Expanded Query

Similar pages



A Google User Jan 8, 2010

very useful..especially when you are say doing some research on a product , company, service etc. , you get objects operating in the same space. For e.g. when I open Forrester, it shows the landing pages for other companies in the same space like Gartner, IDC , etc. and this way it serves 2 purposes - you don't have to save the multiple bookmarks and more importantly, you get similar content offered by other sites

Was this review helpful? Yes No **Mark as spam**

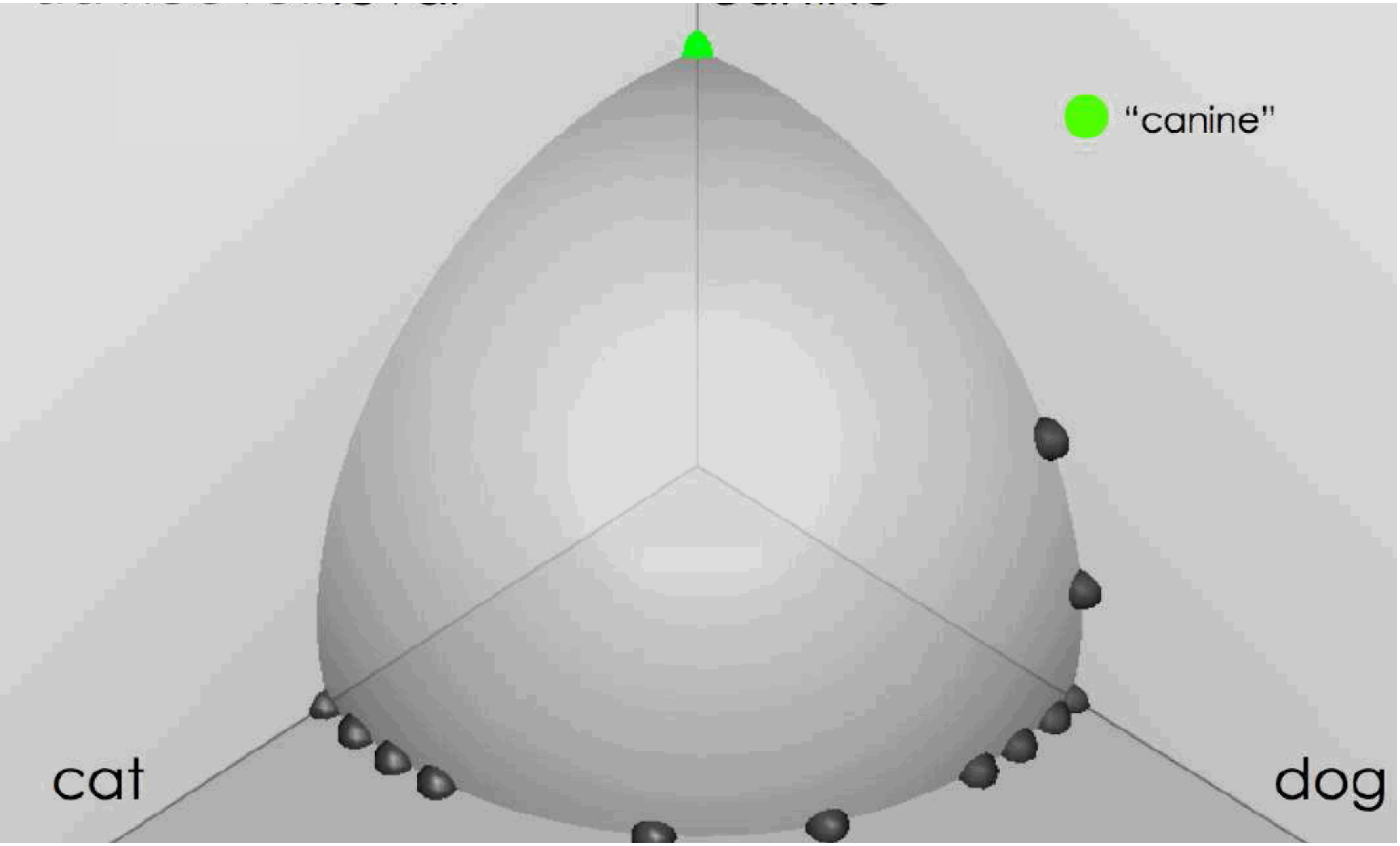
5 out of 6 found this review helpful.

for the page you are viewing. The data related to the query will be handled as described in Google's privacy policy (http://www.google.com /privacypolicy.html).

So how does it work?

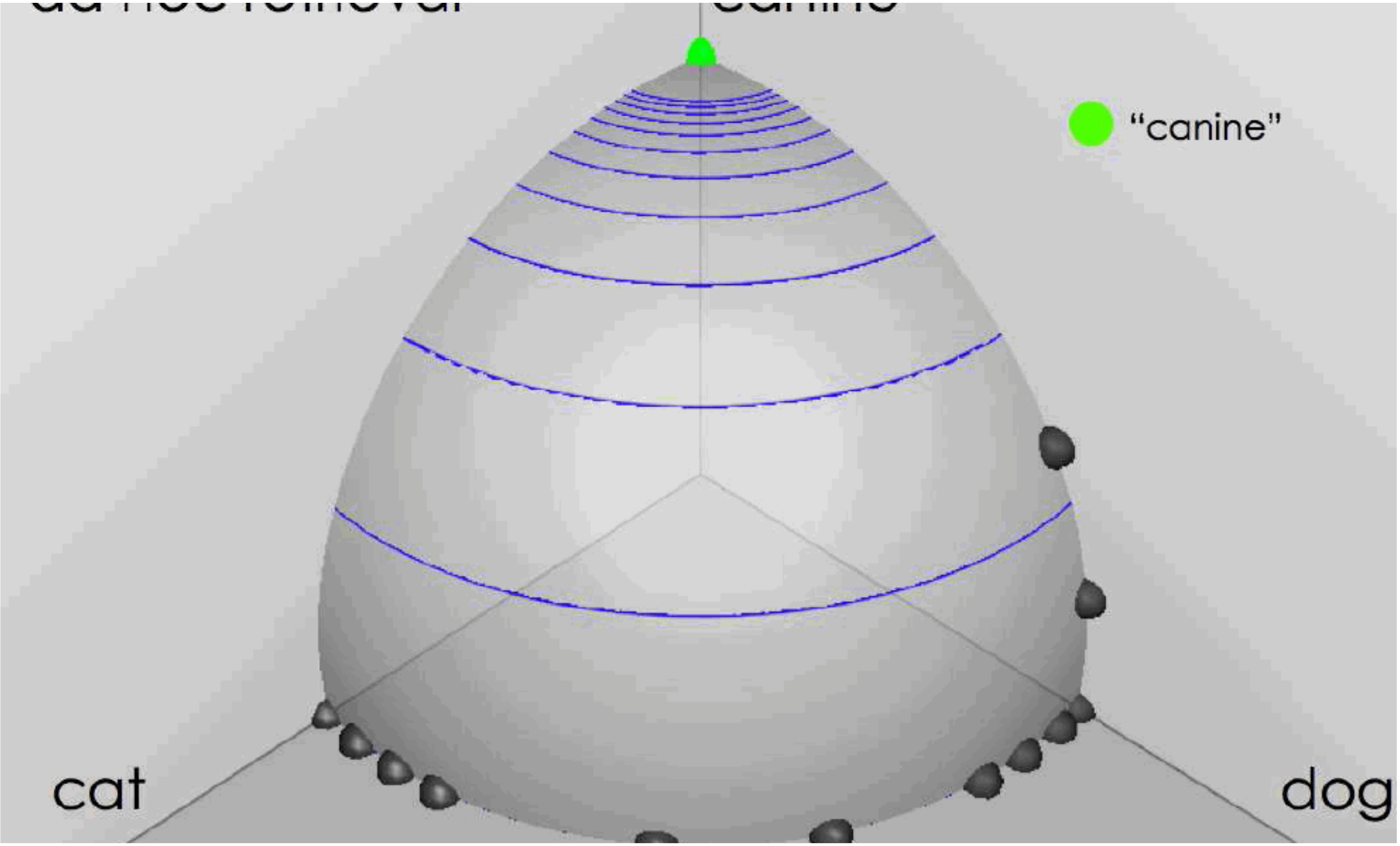
Initial results for query *canine*

source: Fernando Diaz



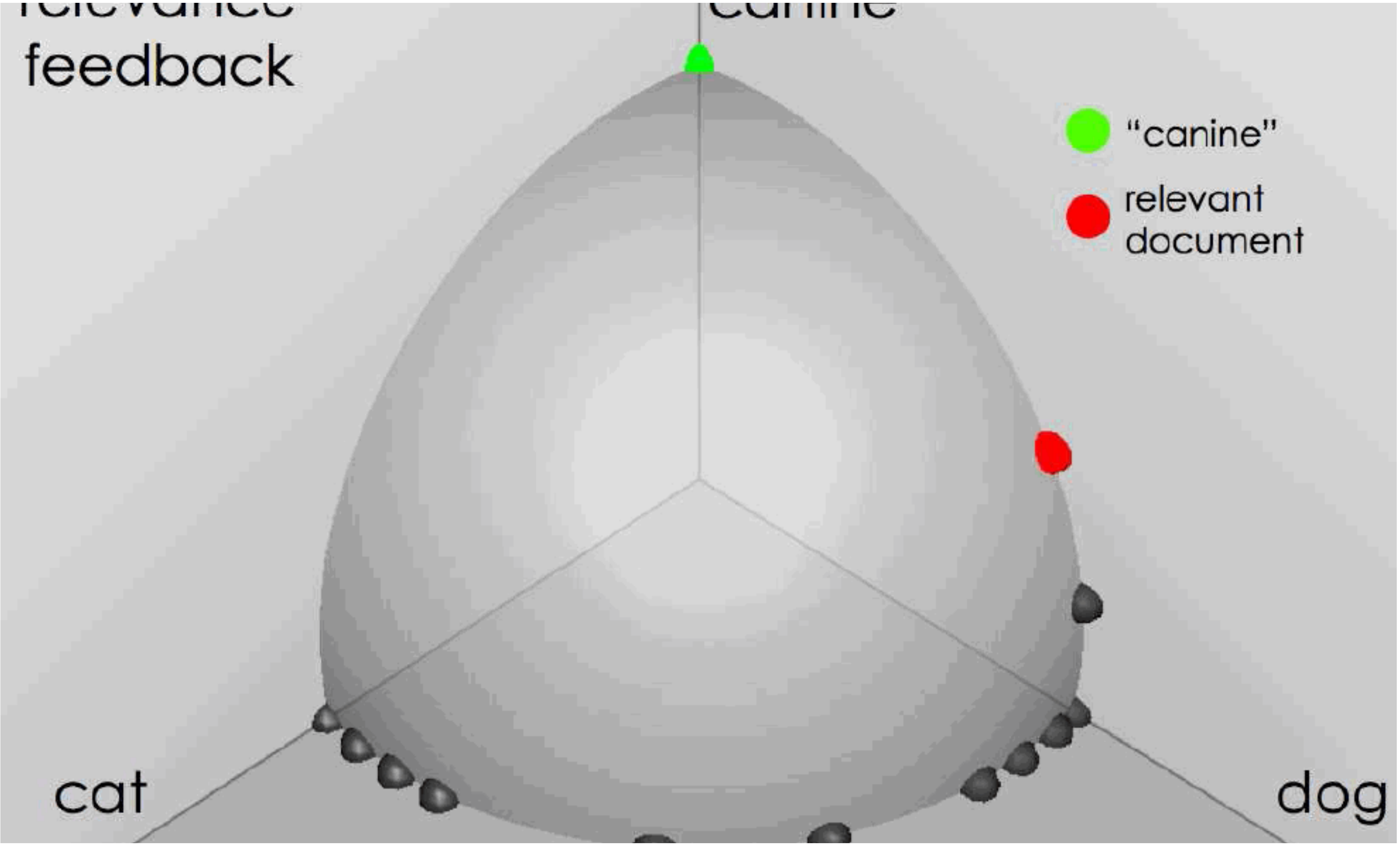
Ad hoc results for query *canine*

source: Fernando Diaz



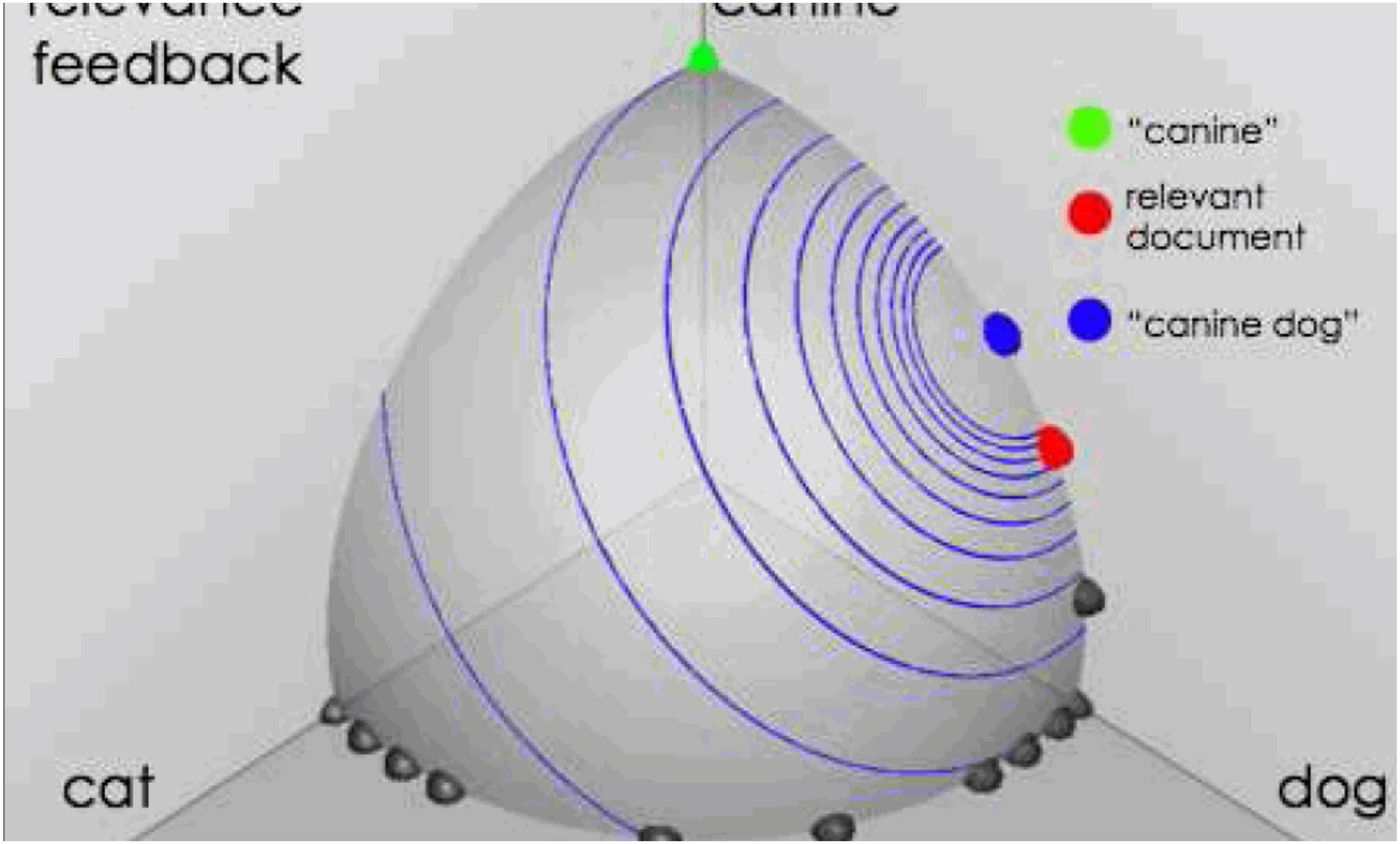
User feedback: Select what is relevant

source: Fernando Diaz



Results after relevance feedback

source: Fernando Diaz



Initial query/results

Initial query: *New space satellite applications*

User marks relevant items

+ 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
− 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

Assume others as nonrelevant

Expanded query after relevance feedback

2.074	new	15.10	space
30.81	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

Results for the 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

Originally Marked Relevant Documents

Key concept: Centroid

- The **centroid** is the center of mass of a set of points.

Definition: Centroid

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

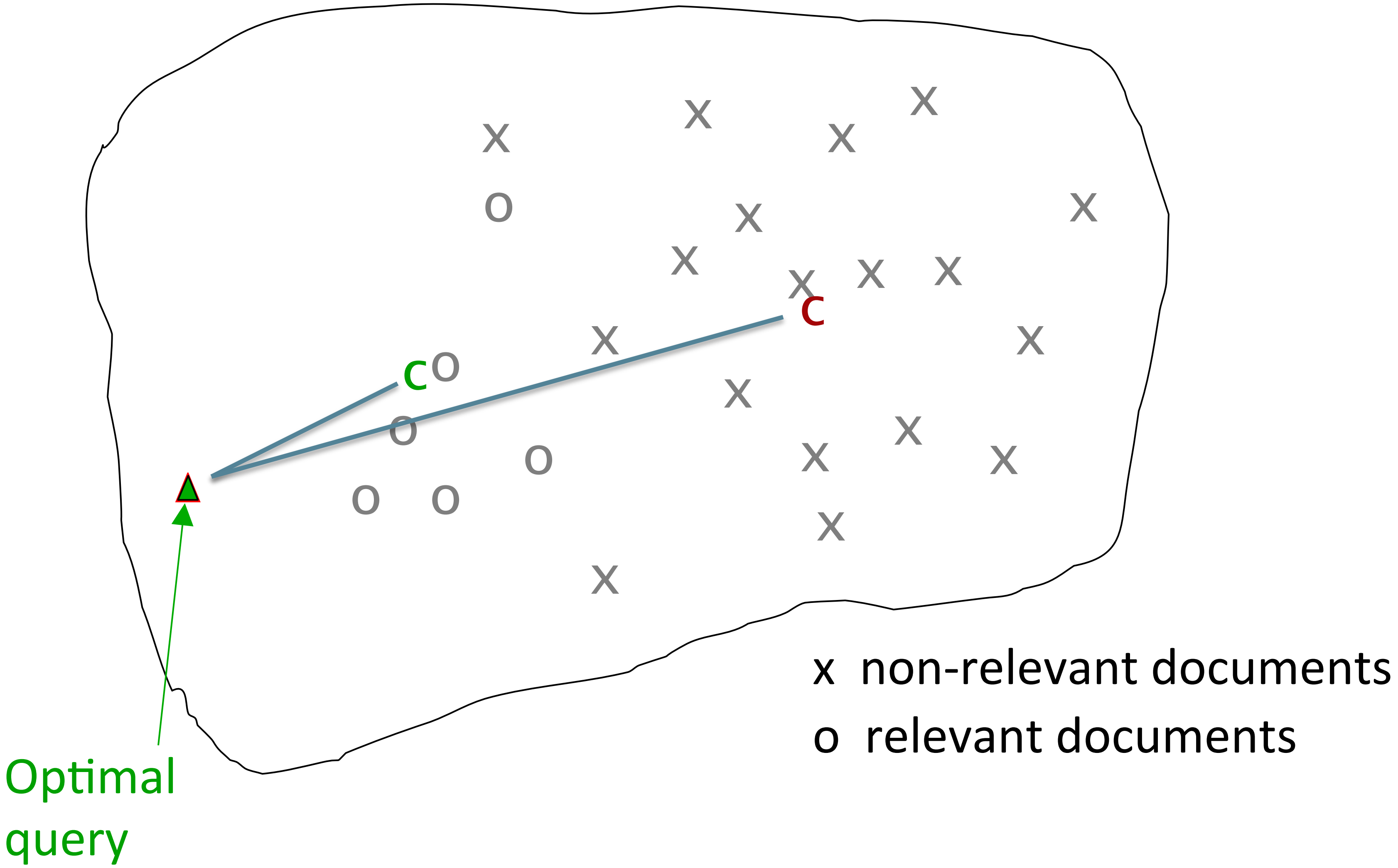
Where D is a set of documents.

Rocchio Algorithm

- Intuitively, we want to separate docs marked as relevant and non-relevant from each other
- The Rocchio algorithm uses the vector space model to pick a new query

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

The Theoretically Best Query



Rocchio (1971)

Popularized in the SMART system (Salton)

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 <u>known</u> relevant doc vectors
- D_{nr} = set of <u>known</u> irrelevant doc vectors

  ⚠ Different from C_r and C_{nr} as we only get judgments from a few documents

- $\{\alpha, \beta, \gamma\}$ = weights (hand-chosen or set empirically)

Weighting

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

- Tradeoff α vs. β/γ : What if we have only a few judged documents?
- Β vs. γ: Which is more valuable?
  - Many systems only allow positive feedback (γ=0). Why?

- Some weights in the query vector can go negative
  - So negative term weights are ignored (set to 0)

When does RF work?

Empirically, a round of RF is often very useful. Two rounds is sometimes marginally useful.

When does it work? When two assumptions hold:

1. User's initial query at least partially works.

2. (Non)-relevant documents are similar.

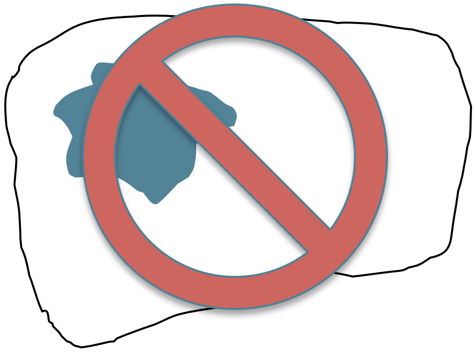
   or term distribution in non-relevant documents are sufficiently distinct from relevant documents

Violation of Assumption 1

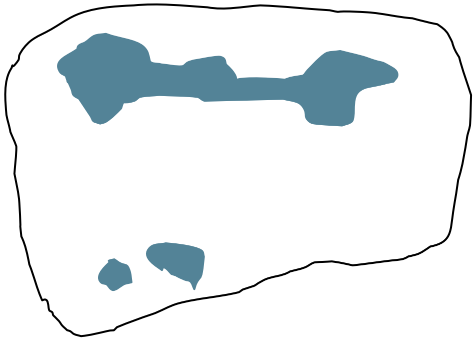
- User does not have sufficient initial knowledge.
- Examples:
  - Misspellings (but not *Brittany Speers*).
  - Cross-language information retrieval (hígado).
  - Mismatch of searcher’s vocabulary vs. collection vocabulary
    - Q: “laptop” but collection all uses “notebook”

Violation of Assumption 2

- There are several relevance prototypes.



- Examples:
  - Burma/Myanmar: change of name
  - Instances of a general concept
  - Pop stars that worked at Burger King



Relevance Feedback: Problems

- Long queries are inefficient for typical IR engine.
  - Long response times for user, as it deals with long queries.
  - Hack: reweight only a # of prominent terms, e.g., top 20.
- Users reluctant to provide explicit feedback
- Harder to understand why particular document was retrieved after RF

Evaluation of relevance feedback strategies

Use q_m and compute precision recall graph

1. Assess on all documents in the collection
   - Spectacular improvements, but … it's cheating!
   - Must evaluate with respect to documents not seen by user
2. Use documents in residual collection (set of documents minus those assessed relevant)
   - Measures usually then lower than for original query
   - But a more realistic evaluation
   - Relative performance can be validly compared

- Best: use two collections each with their own relevance assessments
  - q_o and user feedback from first collection
  - q_m run on second collection and measured

RF in Web search

- True evaluation of RF must also account for usability and time.
- Alternative: User revises and resubmits query.
- Users may prefer revision/resubmission to having to judge relevance of documents (more transparent)

- Some search engines offer a similar/related pages
  - Google (link-based), Altavista, Stanford WebBase
- Some don't use RF because it's hard to explain:
  - Alltheweb, Bing, Yahoo!
- Excite initially had true RF, but abandoned it due to lack of use.

Pseudo relevance feedback (PRF)

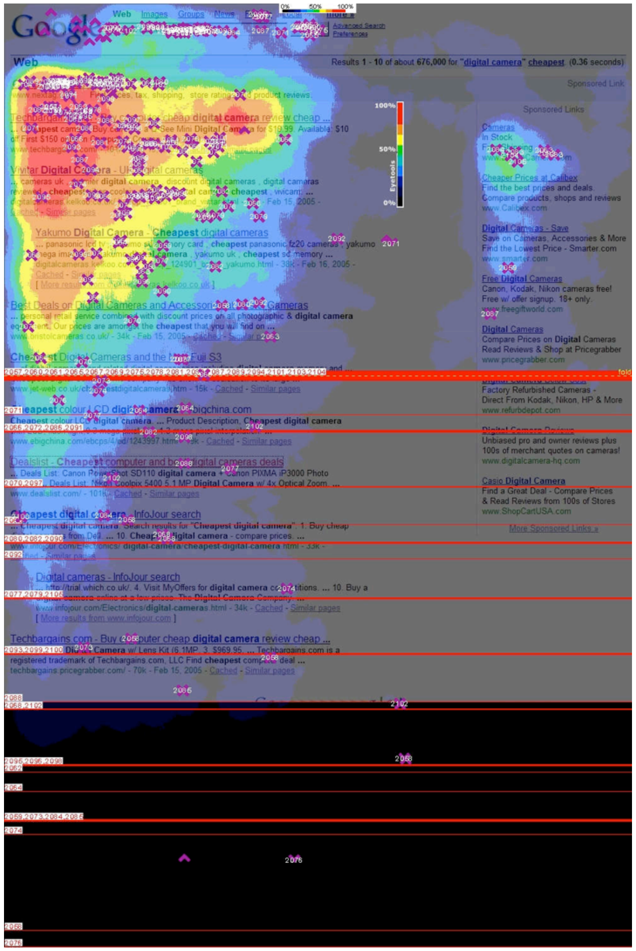
- Blind feedback automates the "manual" part of true RF, by assuming the top *k* is actually relevant.

- Algorithm:
  - Retrieve a ranked list of hits for the user's query
  - Assume that the top *k* documents are relevant.
  - Do relevance feedback

- Works very well on average
- But can go horribly wrong for some queries
- Several iterations can cause query drift

Indirect relevance feedback

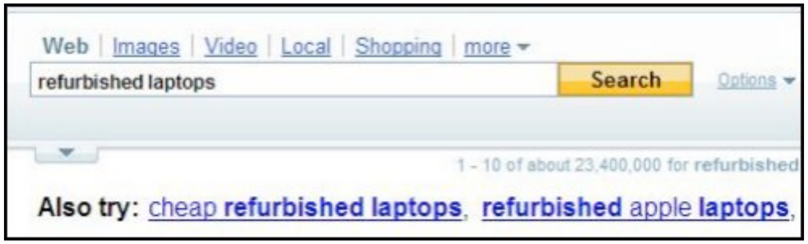
- DirectHit ranked documents that users looked at more often higher.
  - Clicked links are assumed relevant
- Globally: Not necessarily user or query specific.
  - Area of clickstream mining, related to computational advertising (W12)
- Handled as part of machine-learned ranking (Learning to Rank)



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



refurbished laptops

1 - 10 of about 23,400,000 for **refurbished**

Also try: cheap **refurbished laptops**, **refurbished** apple **laptops**,

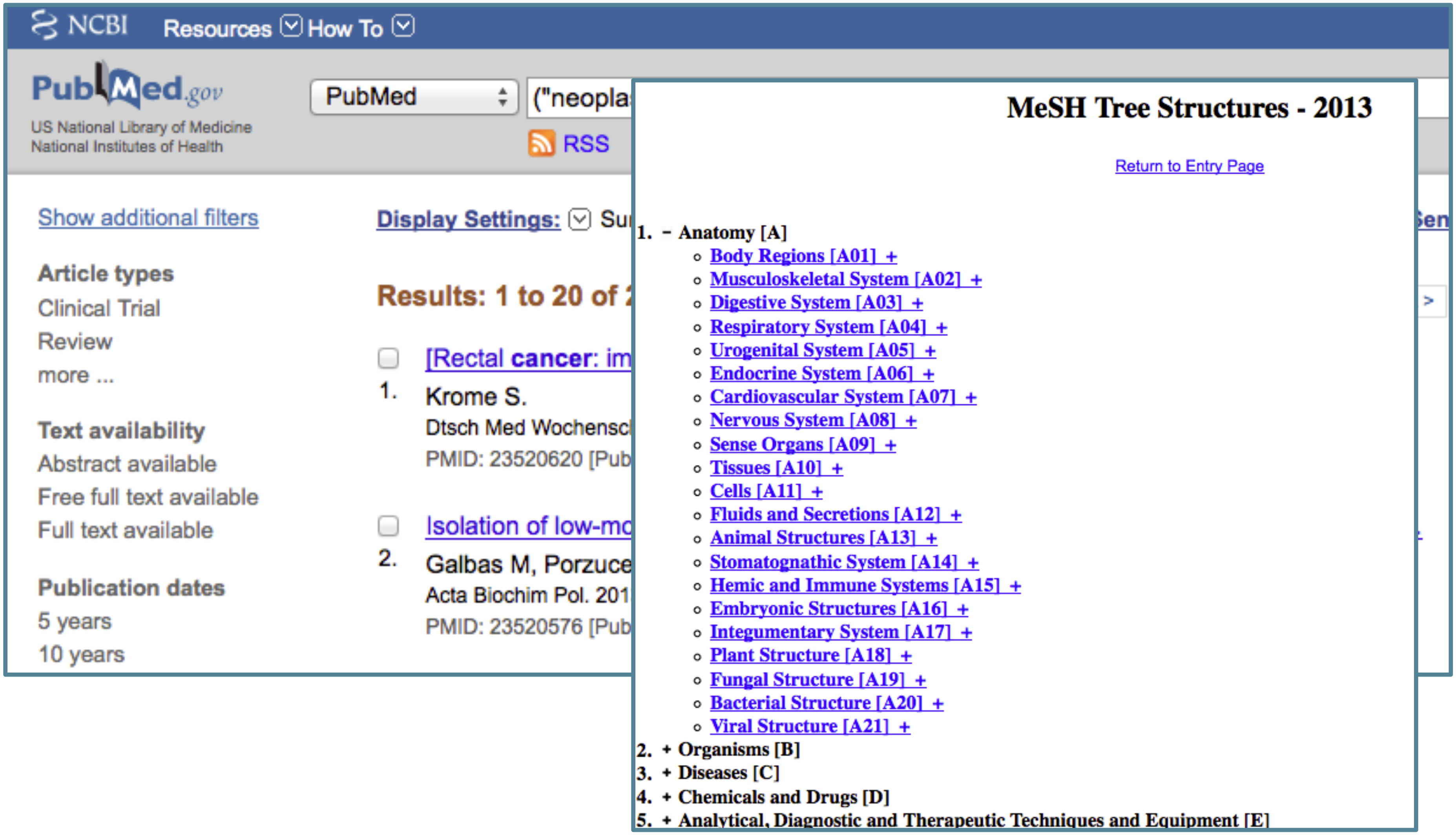
How do we augment the user query?

- Manual thesaurus
  - E.g. MedLine: physician, syn: doc, doctor, MD, medico
  - Can be query rather than just synonyms
- Global Analysis: (static; of all documents in collection)
  - Automatically derived thesaurus
  - Refinements based on query log mining

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
- Generally increases recall, but may decrease precision when terms are ambiguous.
  E.g., "interest rate" → "interest rate fascinate evaluate"

An example of thesaurii: MeSH



Article types
Clinical Trial
Review
more ...

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years

Results: 1 to 20 of

1. [Rectal cancer: im
Krome S.
Dtsch Med Wochensc
PMID: 23520620 [Pub

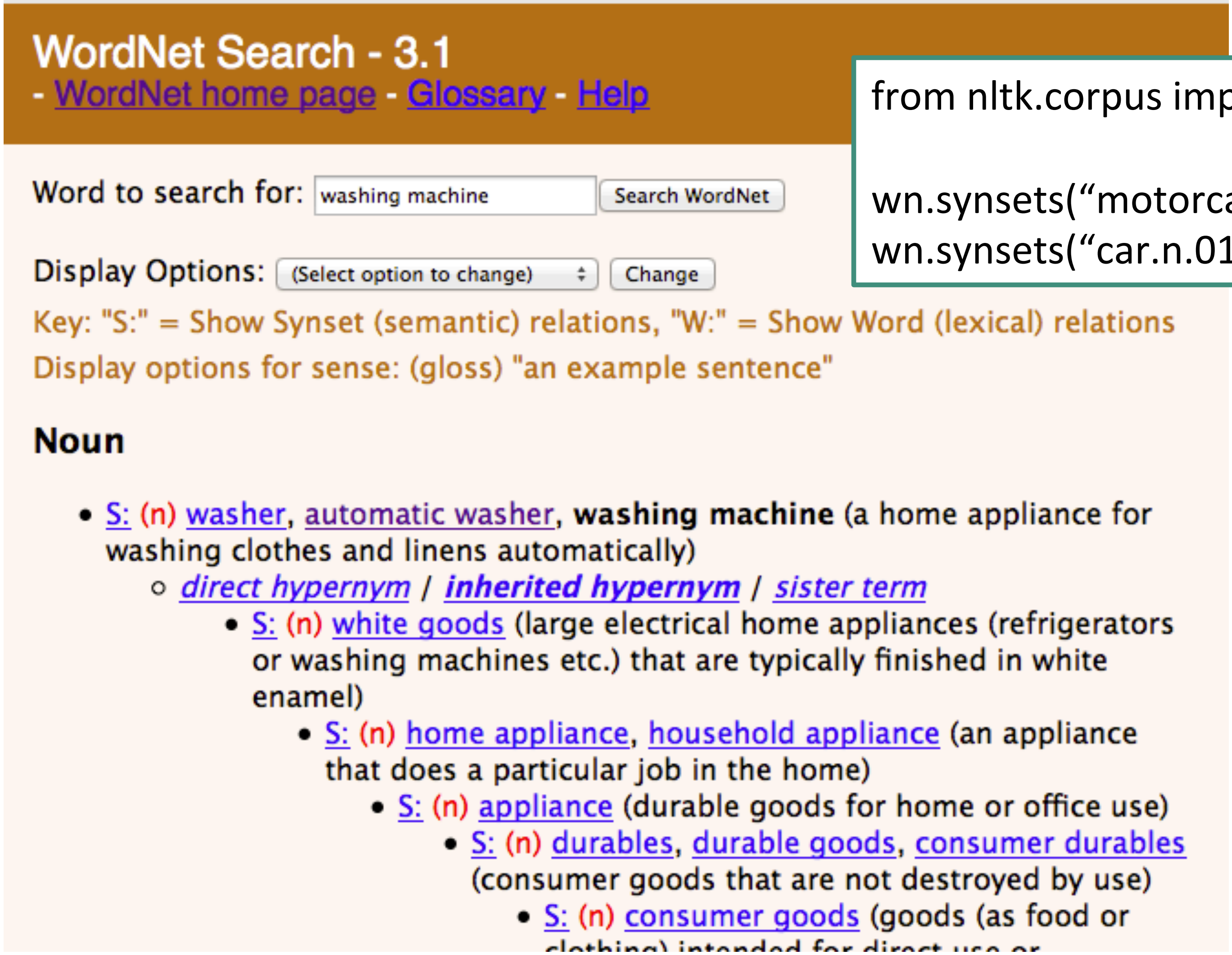
2. Isolation of low-mo
Galbas M, Porzuce
Acta Biochim Pol. 201
PMID: 23520576 [Pub

MeSH Tree Structures - 2013

Return to Entry Page

1. – Anatomy [A]
   - Body Regions [A01] +
   - Musculoskeletal System [A02] +
   - Digestive System [A03] +
   - Respiratory System [A04] +
   - Urogenital System [A05] +
   - Endocrine System [A06] +
   - Cardiovascular System [A07] +
   - Nervous System [A08] +
   - Sense Organs [A09] +
   - Tissues [A10] +
   - Cells [A11] +
   - Fluids and Secretions [A12] +
   - Animal Structures [A13] +
   - Stomatognathic System [A14] +
   - Hemic and Immune Systems [A15] +
   - Embryonic Structures [A16] +
   - Integumentary System [A17] +
   - Plant Structure [A18] +
   - Fungal Structure [A19] +
   - Bacterial Structure [A20] +
   - Viral Structure [A21] +
2. + Organisms [B]
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]

Princeton's WordNet



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: washing machine [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) washer, automatic washer, **washing machine** (a home appliance for washing clothes and linens automatically)
  - ○ *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) white goods (large electrical home appliances (refrigerators or washing machines etc.) that are typically finished in white enamel)
      - S: (n) home appliance, household appliance (an appliance that does a particular job in the home)
        - S: (n) appliance (durable goods for home or office use)
          - S: (n) durables, durable goods, consumer durables (consumer goods that are not destroyed by use)
            - S: (n) consumer goods (goods (as food or clothing) intended for direct use or

```
from nltk.corpus import wordnet as wn

wn.synsets(“motorcar”)
wn.synsets(“car.n.01”).lemma_names
```

Automatic Thesaurus Generation

You shall know a word by the company it keeps
– John R. Firth

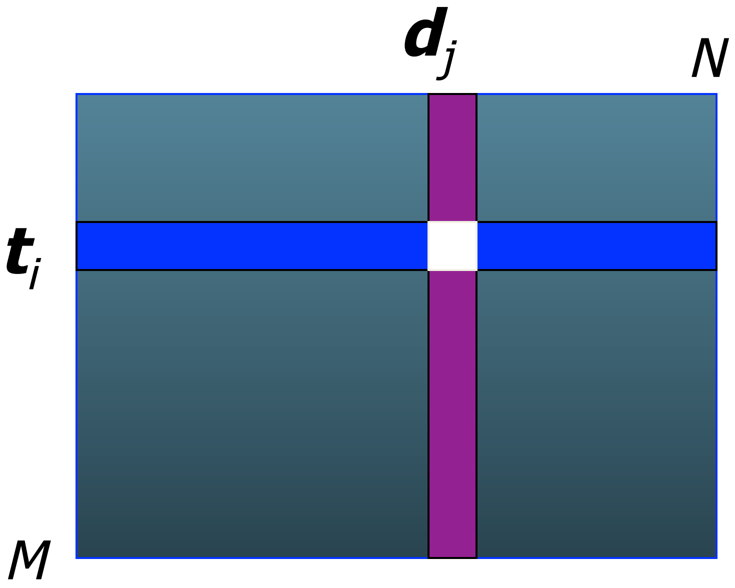
- You can “harvest”, “peel”, “eat” and “prepare” apples and pears, so apples and pears must be similar
- Generate a thesaurus by analyzing the documents
- Assumption: distributional similarity
- I.e., Two words are similar if they co-occur / share same grammatical relations with similar words.

Co-occurrences are more robust; grammatical relations are more accurate. Why?

Co-occurrence Thesaurus

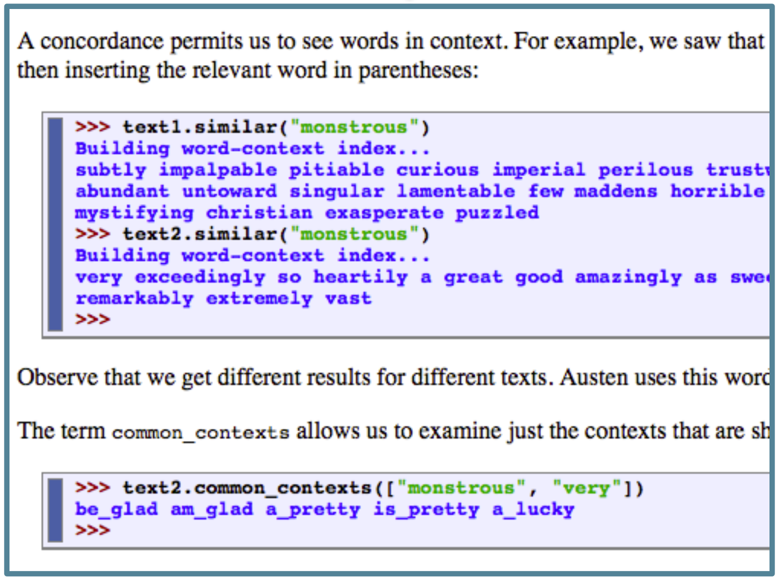
Simplest way to compute one is based on term-term similarities in $C = AA^T$ where A is term-document matrix.

- $w_{i,j}$ = (normalized) weight for (t_i,$\boldsymbol{d}_j$)



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

In NLTK. Did you forget?



A concordance permits us to see words in context. For example, we saw that then inserting the relevant word in parentheses:

```
>>> text1.similar("monstrous")
Building word-context index...
subtly impalpable pitiable curious imperial perilous trustw
abundant untoward singular lamentable few maddens horrible
mystifying christian exasperate puzzled
>>> text2.similar("monstrous")
Building word-context index...
very exceedingly so heartily a great good amazingly as swee
remarkably extremely vast
>>>
```

Observe that we get different results for different texts. Austen uses this word

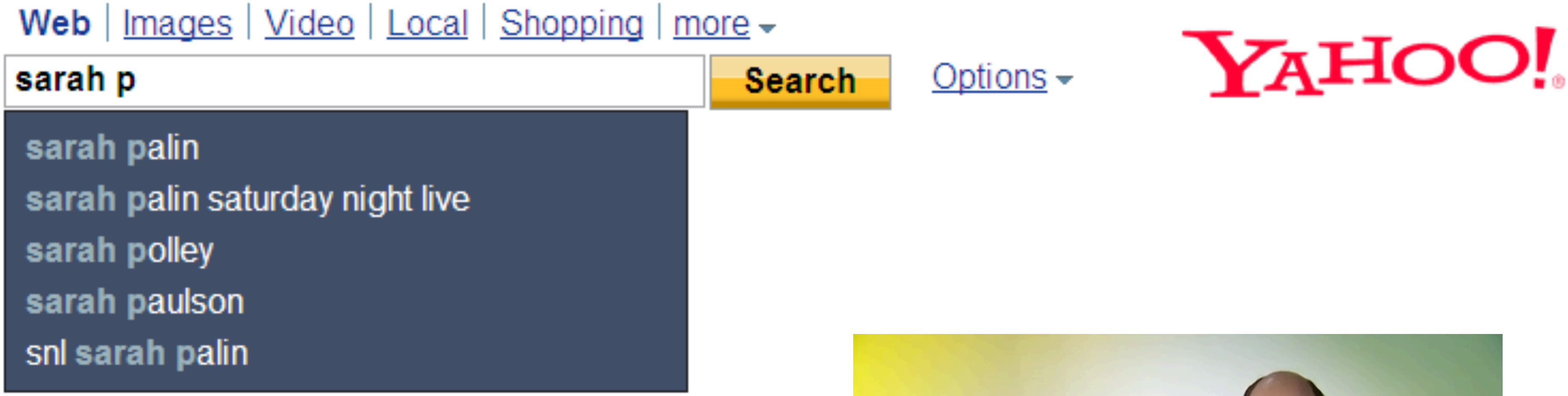
The term `common_contexts` allows us to examine just the contexts that are sh

```
>>> text2.common_contexts(["monstrous", "very"])
be_glad am_glad a_pretty is_pretty a_lucky
>>>
```

Automatic Thesaurus Generation: Problems

- Term ambiguity may introduce irrelevant statistically correlated terms.
  - “Apple computer” → “Apple red fruit computer”
- Problems:
  - False positives: Words deemed similar that are not (Especially opposites)
  - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

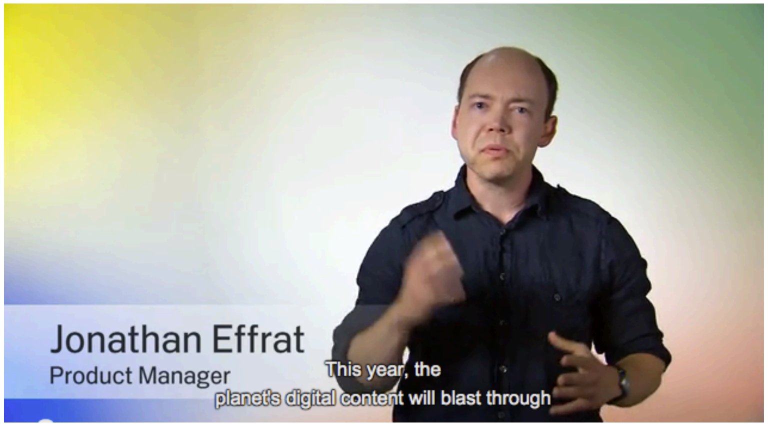
Implicit Query Expansion



sarah p

sarah palin
sarah palin saturday night live
sarah polley
sarah paulson
snl sarah palin

Would you expect such a feature to increase the query?



Summary

- Chapter 9 of IIR

1. Relevance Feedback “Documents”
2. Query Expansion “Terms”

Rocchio: Intuition: Maximize similarity with relevant and difference from non-relevant.

In the web context, clickstream and implicit feedback common

- Resources
  - MG Ch. 4.7 and MIR Ch. 5.2 – 5.4