Relevance Feedback

Dorota Glowacka

glowacka@cs.helsinki.fi

Can we improve recall in search?

- Two ways to improve recall:
  - relevance feedback
  - query expansion
- Example: you type the query ”aircraft” but the database contains only documents containing the word ”plane”.
- A simple IR system will not return these documents although they might be perefectly satisfactory for the user
- Aim: enable the IR system to return relevant documents even if there is no term match between the original query and the relevant document(s)

Query (re-)formulation

- No detailed knowledge of collection and retrieval environment
  - Difficult to formulate queries well designed for retrieval
  - Need many formulations of queries for good retrieval
- First formulation is usually a naive attempt to retrieve relevant information
- ”Word mismatch” problem
  - Some of the unretrieved relevant documents are indexed by a different set of terms compared to the query or other relevant documents
- The idea is when documents are initially retrieved:
  - They should be examined for relevance information
  - Then we can improve the query for retrieving additional relevant documents
- Query reformulation:
  - Expanding original query with new terms
  - Reweighing the terms in the (expanded) query

Term re-weighting without query expansion

A probabilistic model proposed by Robertson and Sparck-Jones (1976)

$$W_{ij} = \log \frac{\dfrac{r}{R-r}}{\dfrac{n-r}{(N-n)-(R-r)}}$$

Wij = the term weight for term i in query j

r = the number of relevant documents for query j having term i

R = the total number of relevant documents for query j

n = the number of documents in the collection having term i

N = the number of documents in the collection

Experimental results show that this term weighting produced somewhat better results than IDF measure alone

Term re-weighting without query expansion

Croft (1983) extended this weighting scheme as follows:

initial search $$W_{ijk} = (C + IDF_i) * f_{ik}$$

Feedback $$W_{ijk} = (C + \log \frac{p_{ij}(1-q_{ij})}{(1-p_{ij})q_{ij}}) * f_{ik}$$

W_{ijk} = weight for term *i* in query *j* and document *k*
IDF_i = IDF weight for term *i*
p_{ij} = probability of term *i* to be assigned within the set of relevant documents for query *j*
q_{ij} = probability that term *i* is assigned to the set of non-relevant docuemnts for query *j*
C = a constant to taylor the weighting for various document collections
$f_{ik} = K + (1 - K)(freq_{ik}/max\ freq_k)$
K = a constant to adjust the relative importance of the two weighting schemes
$freq_{ik}$ = frequency of term *i* in document *k*
$freq_k$ = frequency of any term in document *k*

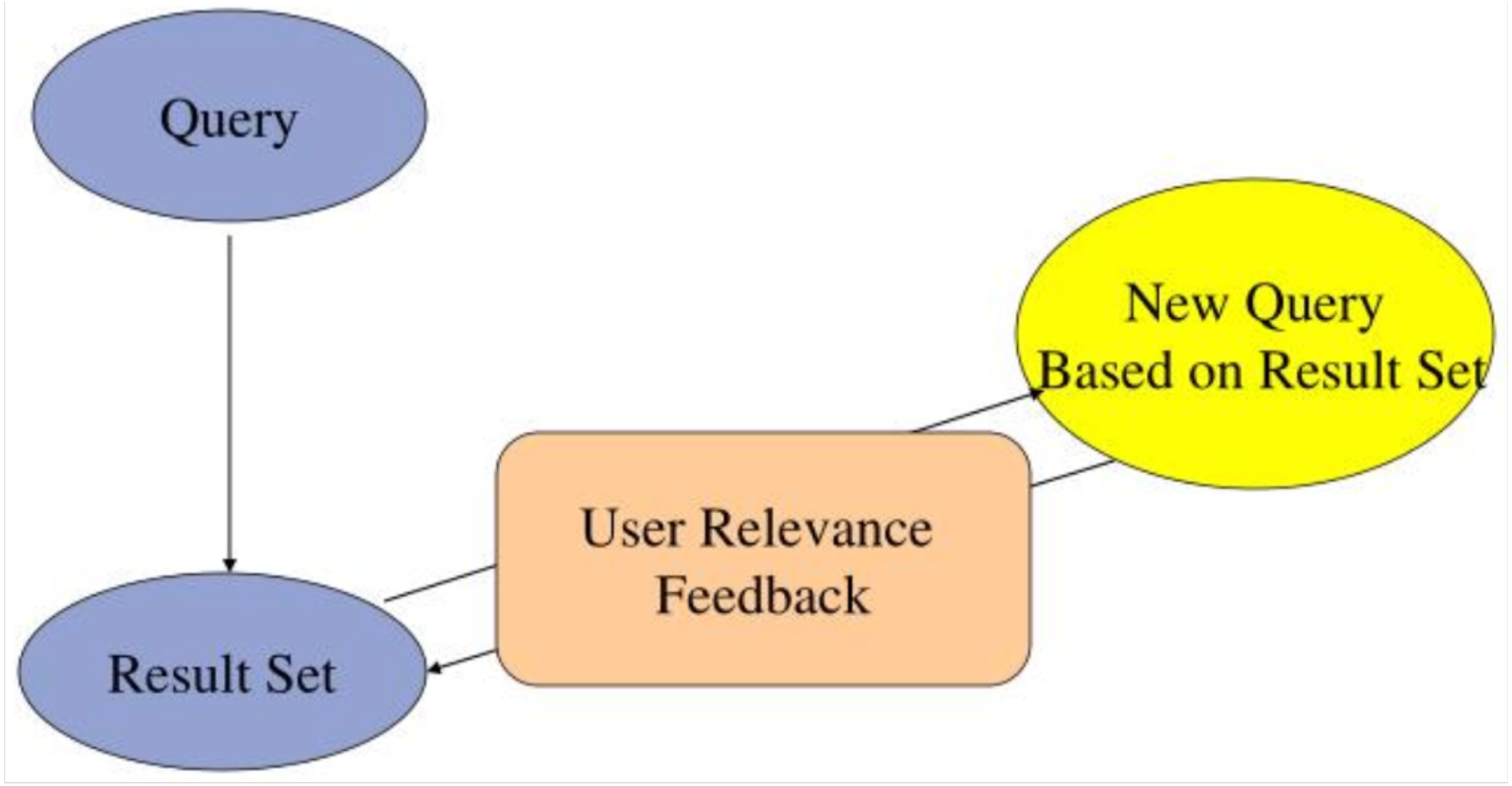
Query (re)-formulation: Three approaches

- **Relevance feedback**: based on feedback from users, e.g. Rocchio or Ide.

- **Local analysis** (also called **pseudo-relevance feedback**):
    - Approaches based on information derived from the set of initially retrieved documents (local set of documents)

- **Global analysis**
    - Approaches based on global information derived from the document collection

Conceptual Model of Relevance Feedback

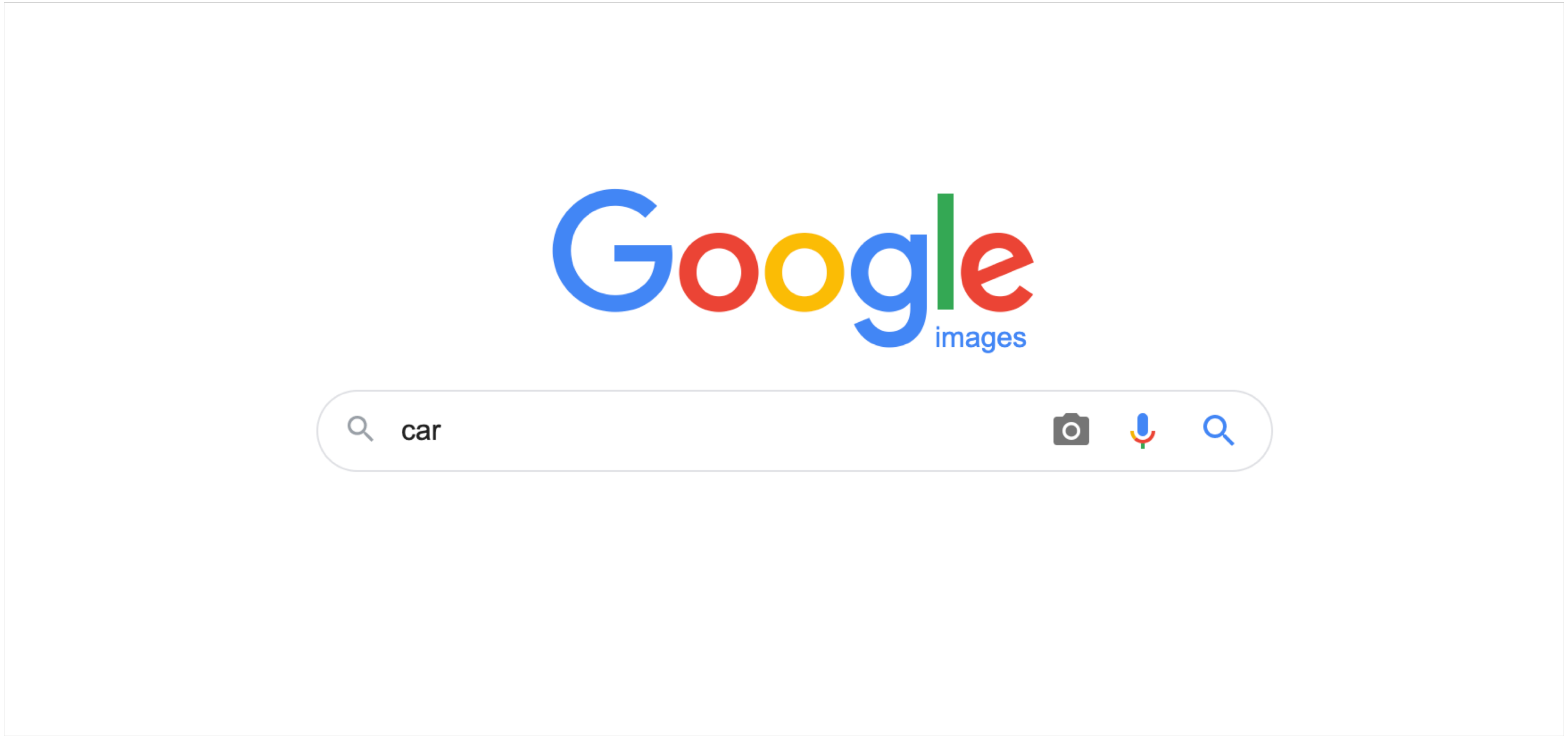


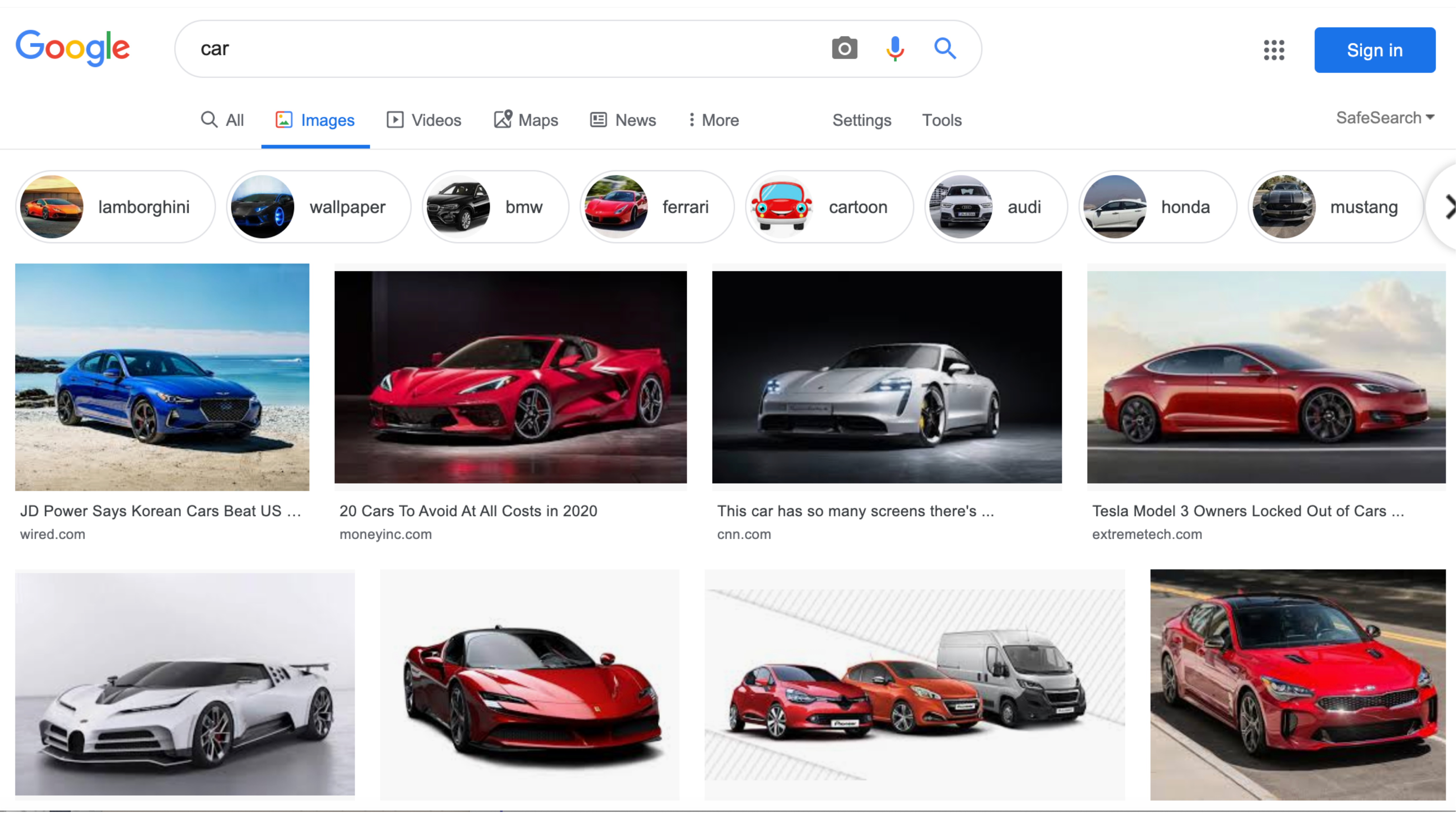
Relevance feedback

Relevance feedback can be viewed as an iterative cycle:

- User are presented with a list of retrieved documents.
- User marks documents that they consider relevant (or not relevant)
  - In practice only top 10-20 ranked documents are examined
  - The procedure is incremental: users look at one document at a time
- The relevance feedback algorithm selects important terms from documents assessed relevant by users.
- The relevance feedback algorithm emphasises the importance of these terms in a new query in the following ways:
  - Query expansion: add these terms to the query
  - Term reweighing: modify the term weights in the query
  - Query expansion + term reweighing
- The updated query is submitted to the system.
- If the user is satisfied with the new set of retrieved documents, then the relevance feedback process stops, otherwise the user marks more documents as relevant or not relevant

Relevance feedback example




Google
car
Sign in
All
Images
Videos
Maps
News
More
Settings
Tools
SafeSearch
lamborghini
wallpaper
bmw
ferrari
cartoon
audi
honda
mustang
JD Power Says Korean Cars Beat US ...
wired.com
20 Cars To Avoid At All Costs in 2020
moneyinc.com
This car has so many screens there's ...
cnn.com
Tesla Model 3 Owners Locked Out of Cars ...
extremetech.com

JD Power Says Korean Cars Beat US and ...
wired.com

20 Cars To Avoid At All Costs in 2020
moneyinc.com

This car has so many screens there's ...
cnn.com

Tesla Model 3 Owners Locked Out of Cars ...
extremetech.com

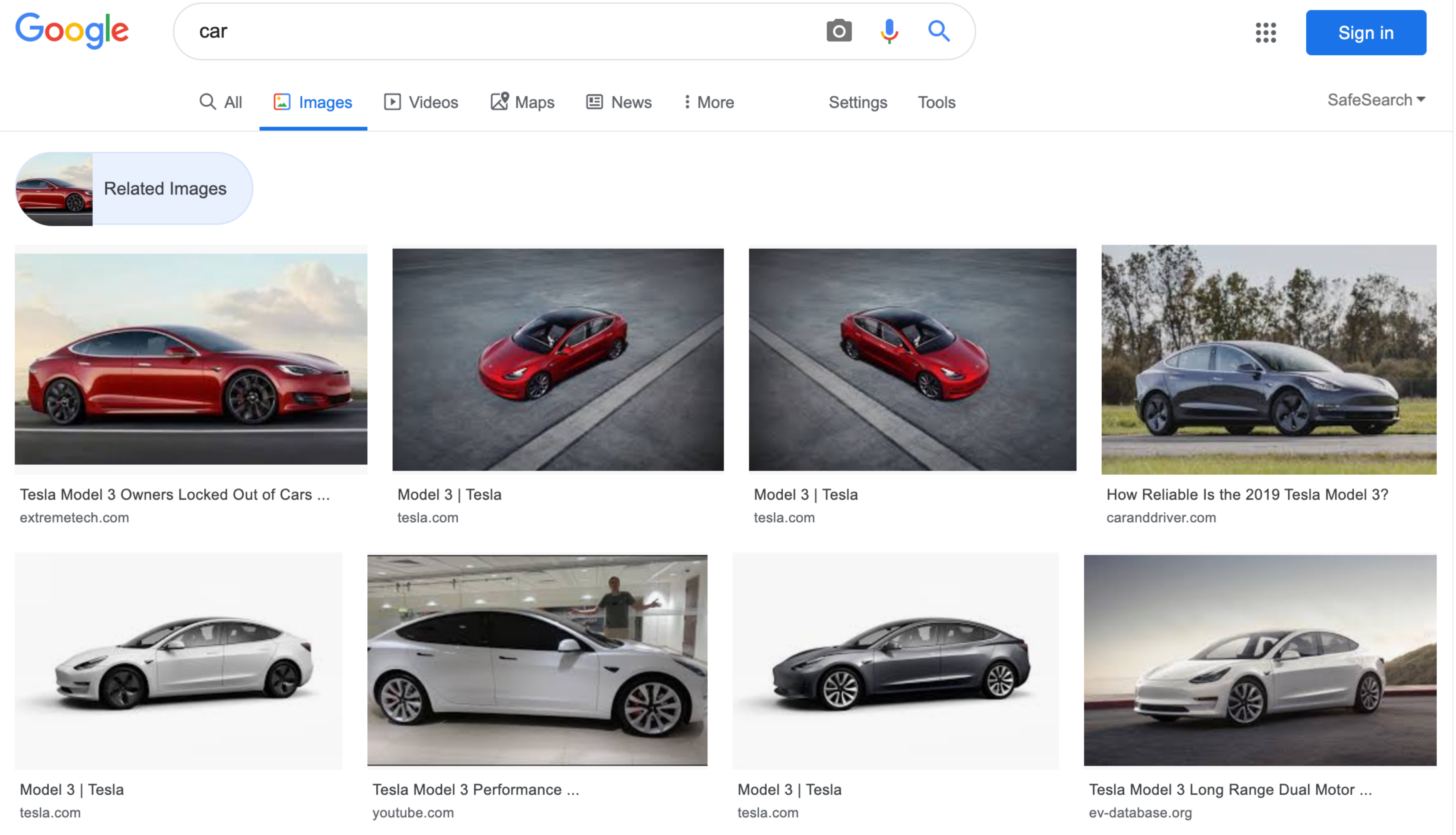
ExtremeTech

Tesla Model 3 Owners Locked Out of Cars Thanks to App Outage ...

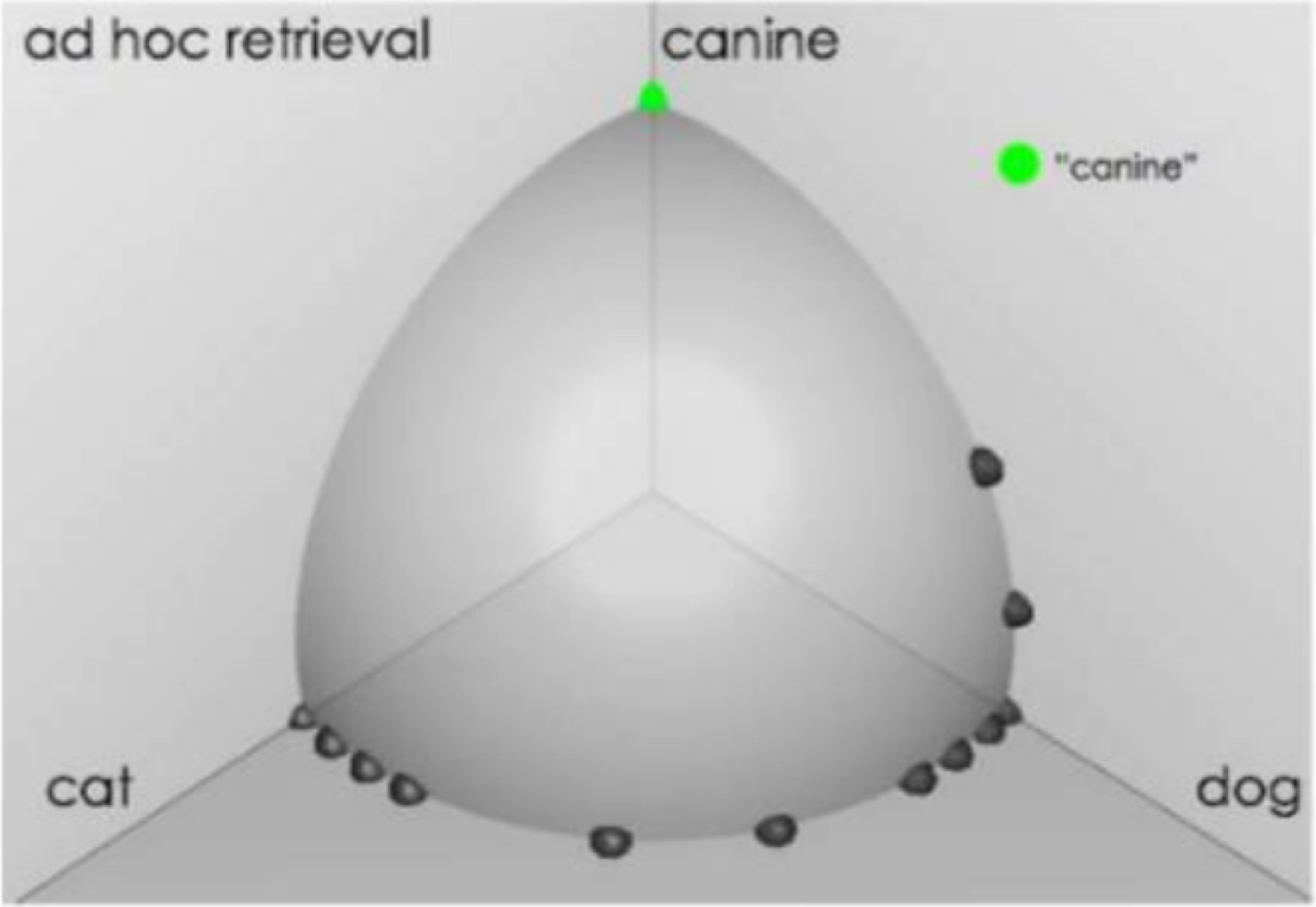
Images may be subject to copyright. Find out more

Related images

See more


Google
car
Sign in
All
Images
Videos
Maps
News
More
Settings
Tools
SafeSearch
Related Images
Tesla Model 3 Owners Locked Out of Cars ...
extremetech.com
Model 3 | Tesla
tesla.com
Model 3 | Tesla
tesla.com
How Reliable Is the 2019 Tesla Model 3?
caranddriver.com
Model 3 | Tesla
tesla.com
Tesla Model 3 Performance ...
youtube.com
Model 3 | Tesla
tesla.com
Tesla Model 3 Long Range Dual Motor ...
ev-database.org

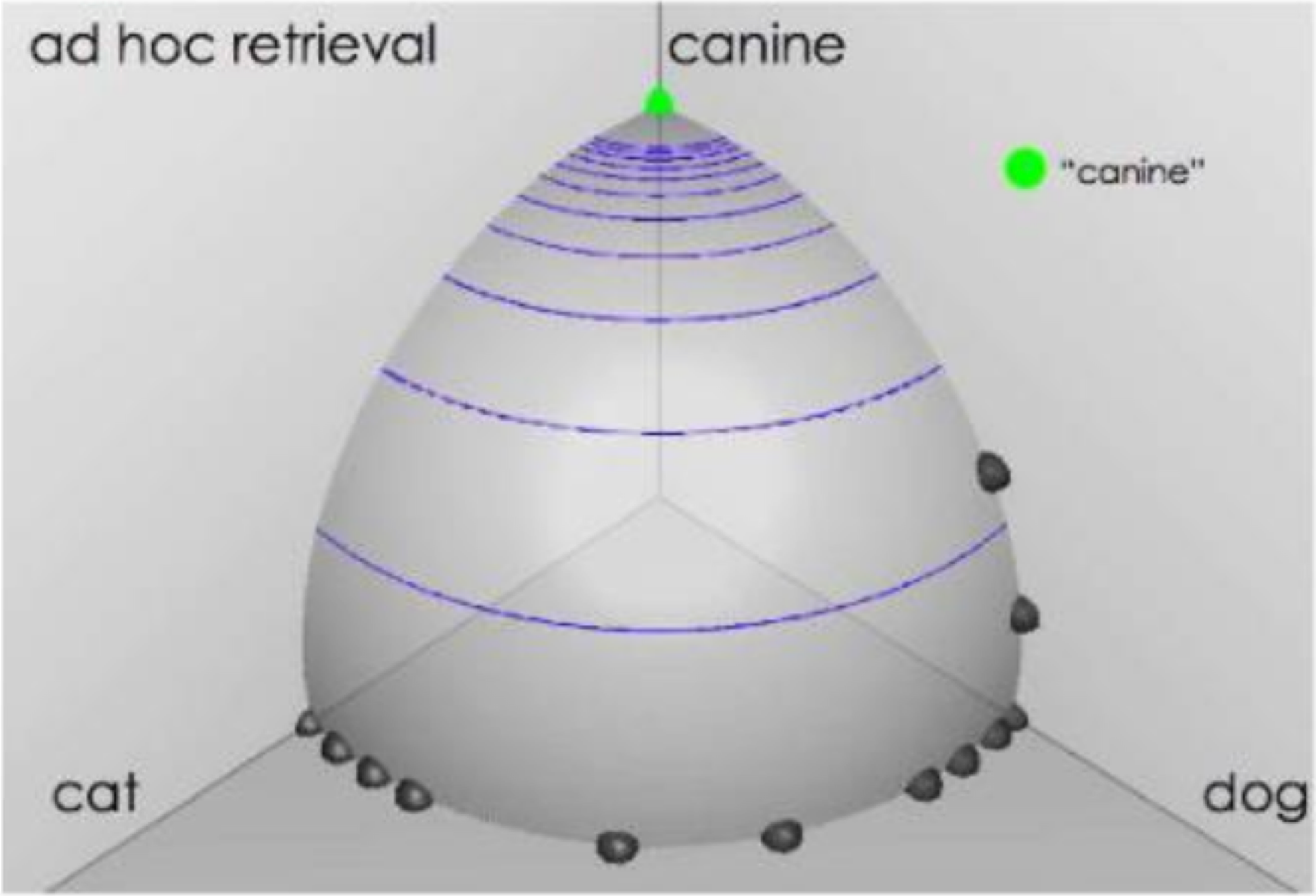
Vector space example: query ”canine”



Source:

Fernando Díaz

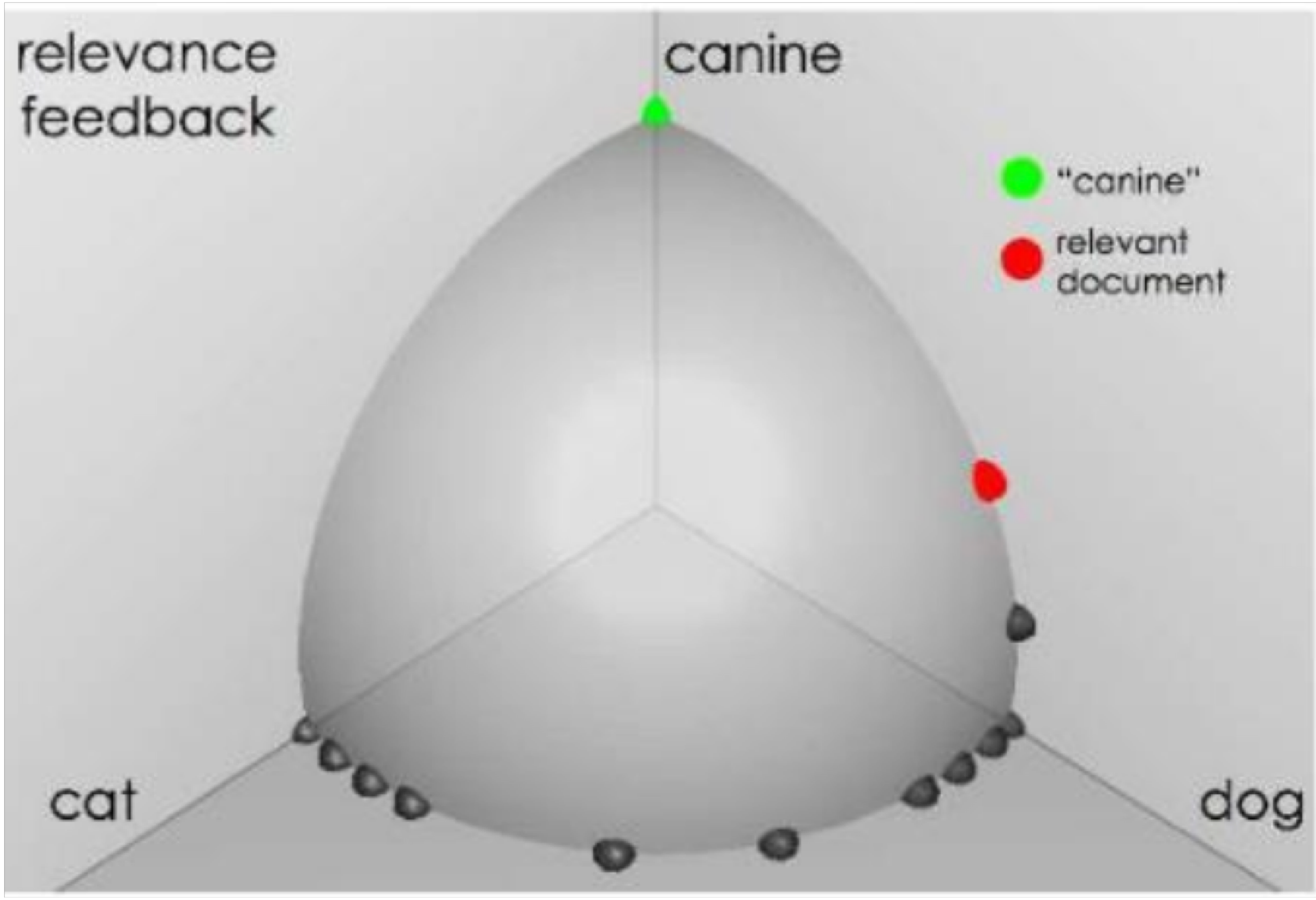
Similarity of documents to query ”canine”



Source:

Fernando Díaz

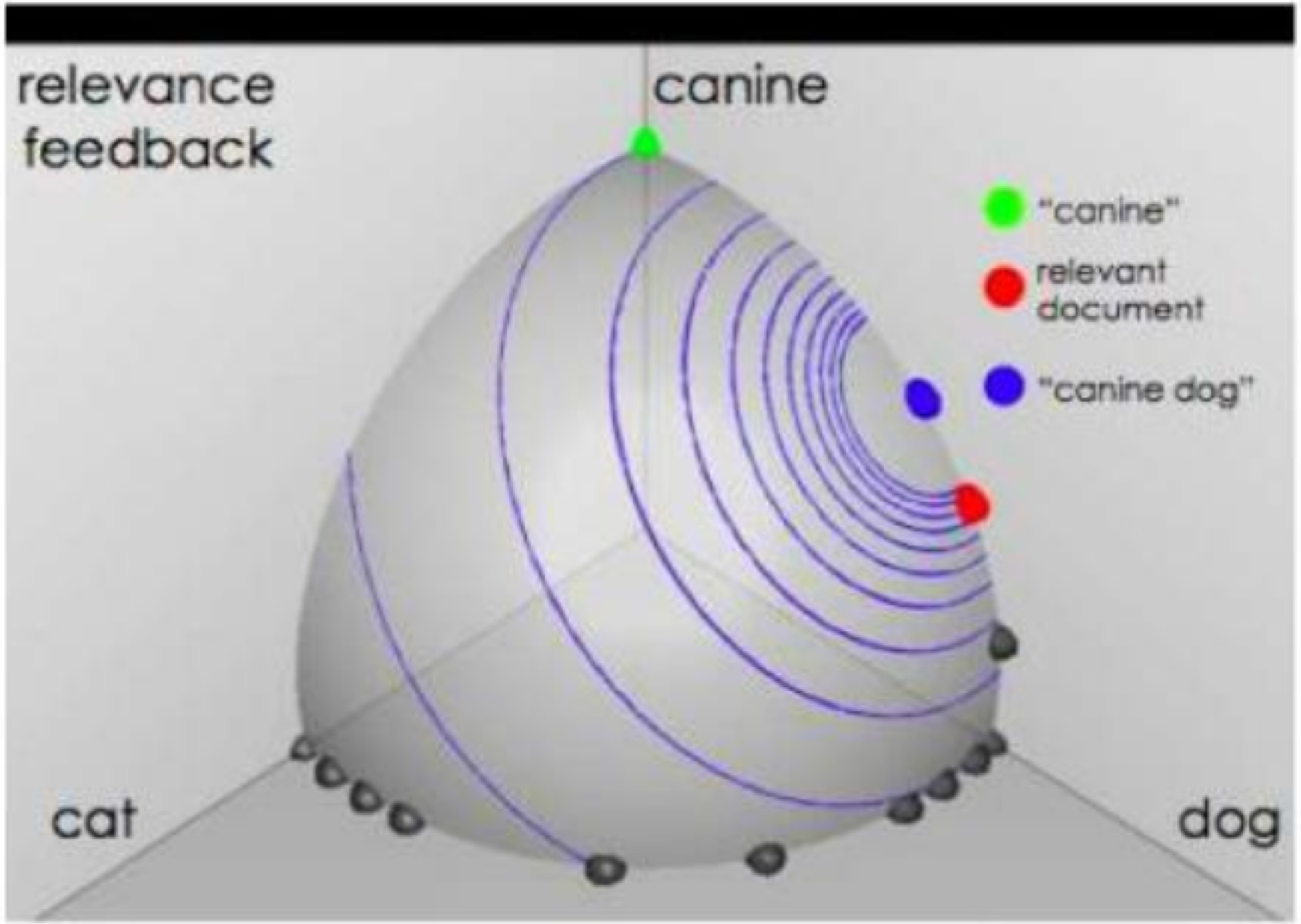
User feedback: Select relevant documents



Source:

Fernando Díaz

Results after relevance feedback



Source:

Fernando Díaz

Document search example

Initial query:

[new space satellite applications] Results for initial query: (r = rank)

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

User then marks relevant documents with "+".

Expanded query after relevance feedback

2.074	new	15.106	space
30.816	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

Compare to original

query: [new space satellite applications]

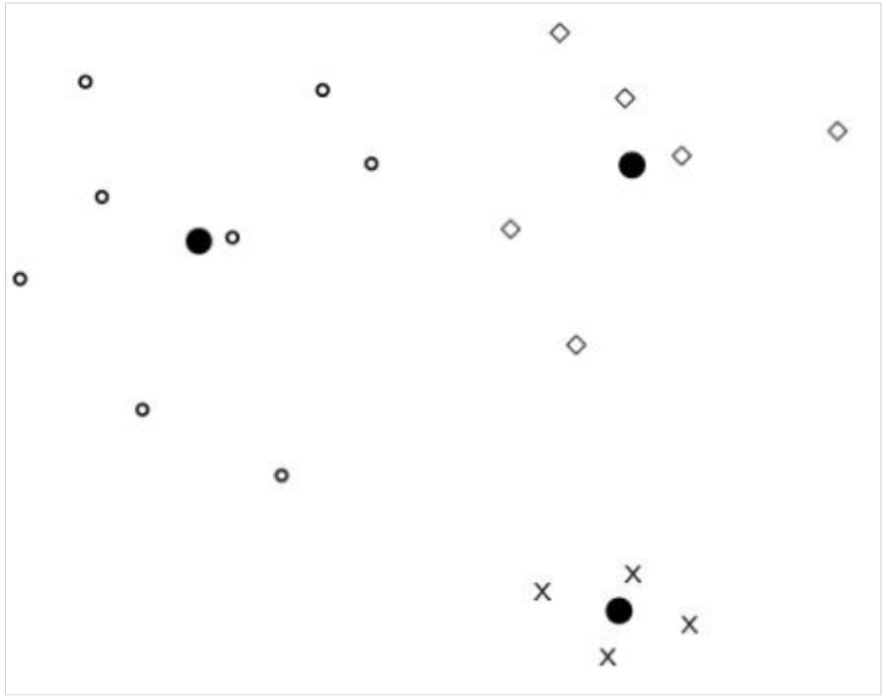
Results for expanded query

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

Key concept in relevance feedback: Centroid

- The centroid is the centre of mass of a set of points
- Documents are represented as points in a high-dimensional space
- We can compute centroids of documents

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



where *D* is a set of documents and $\vec{v}(d) = \vec{d}$ is the vector representing document *d*.

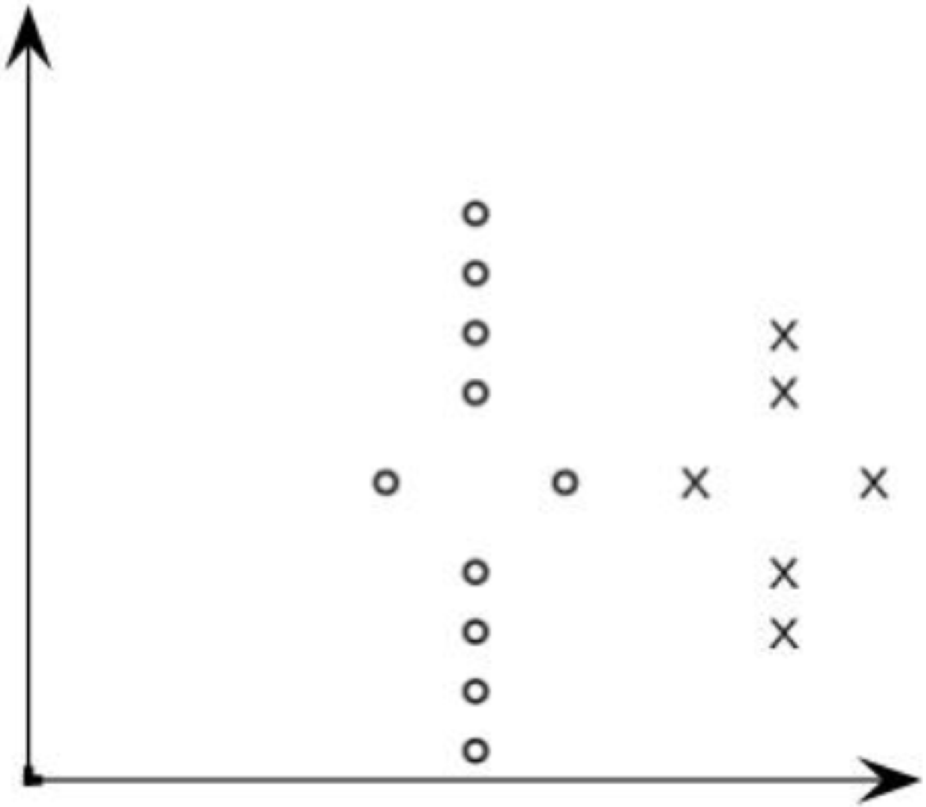
Rocchio Algorithm

- The Rocchio algorithm incorporates relevance feedback information into the vector space model.
- We want to maximize *sim(Q, Cr) – sim(Q, Cnr)*
- The optimal query vector for separating relevant and non-relevant documents (with cosine similarity):

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

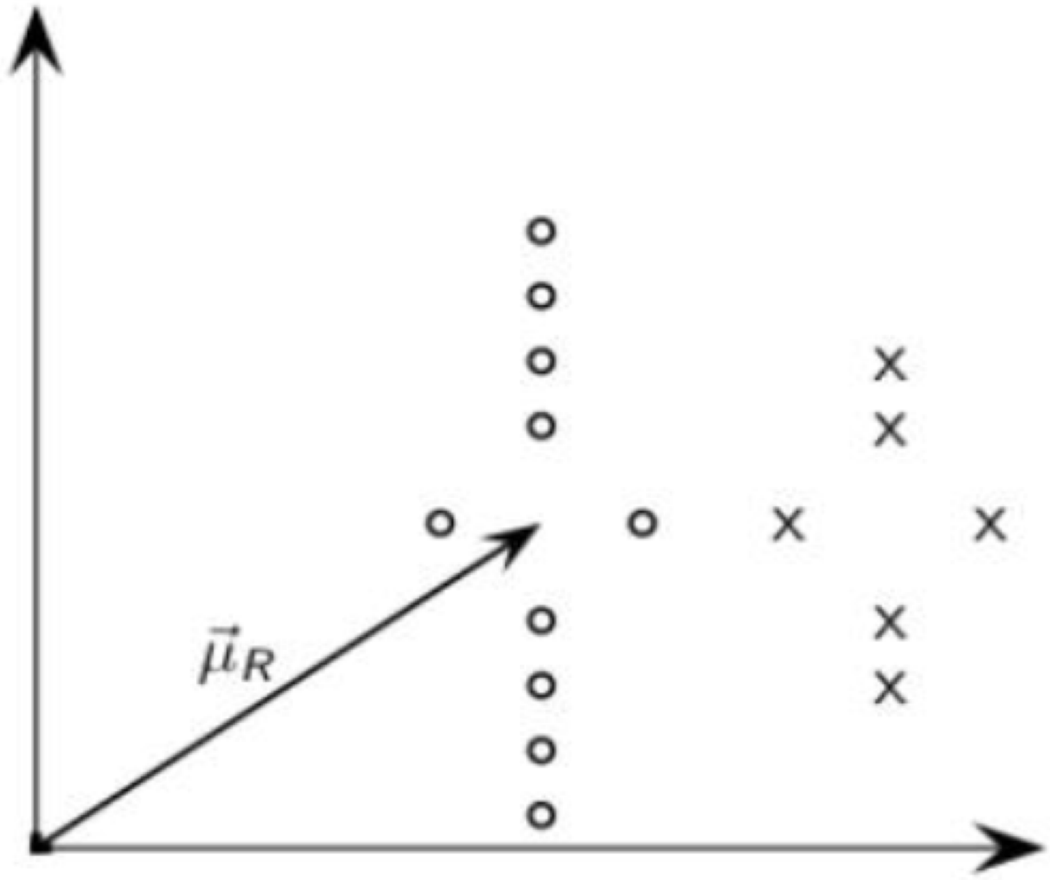
Q_{opt}= optimal query; C_r = set of relevant doc vectors; N = collection size

Computing Rocchio's vector



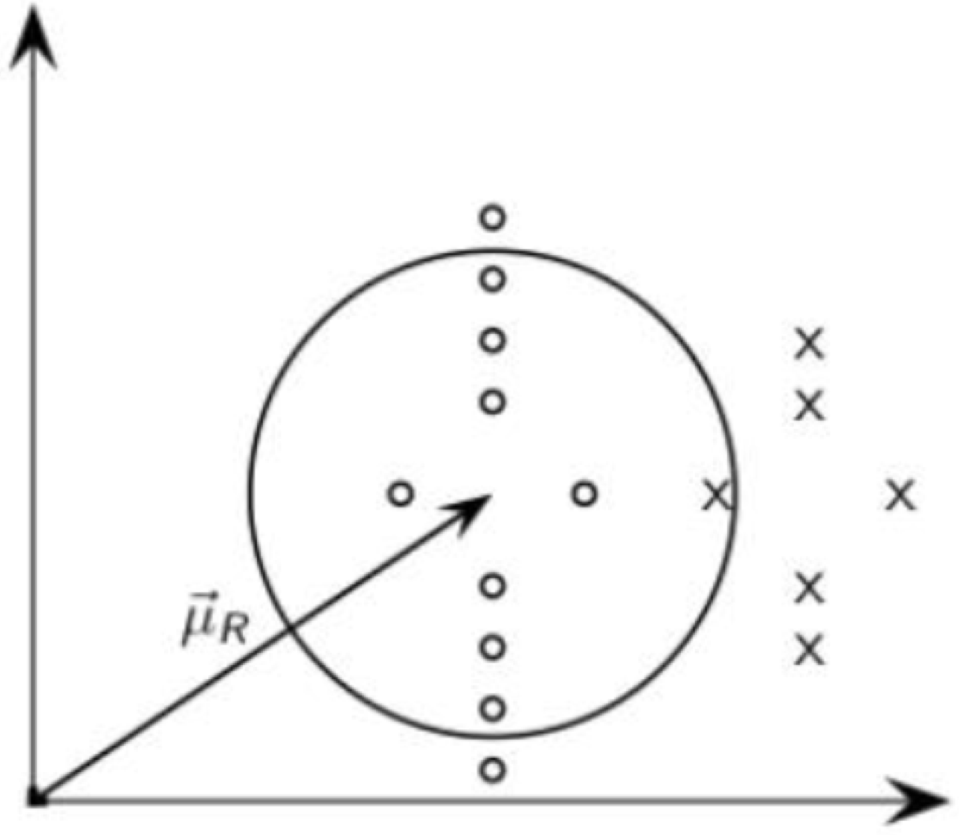
circles: relevant documents, Xs: nonrelevant documents

Rocchio algorithm illustrated



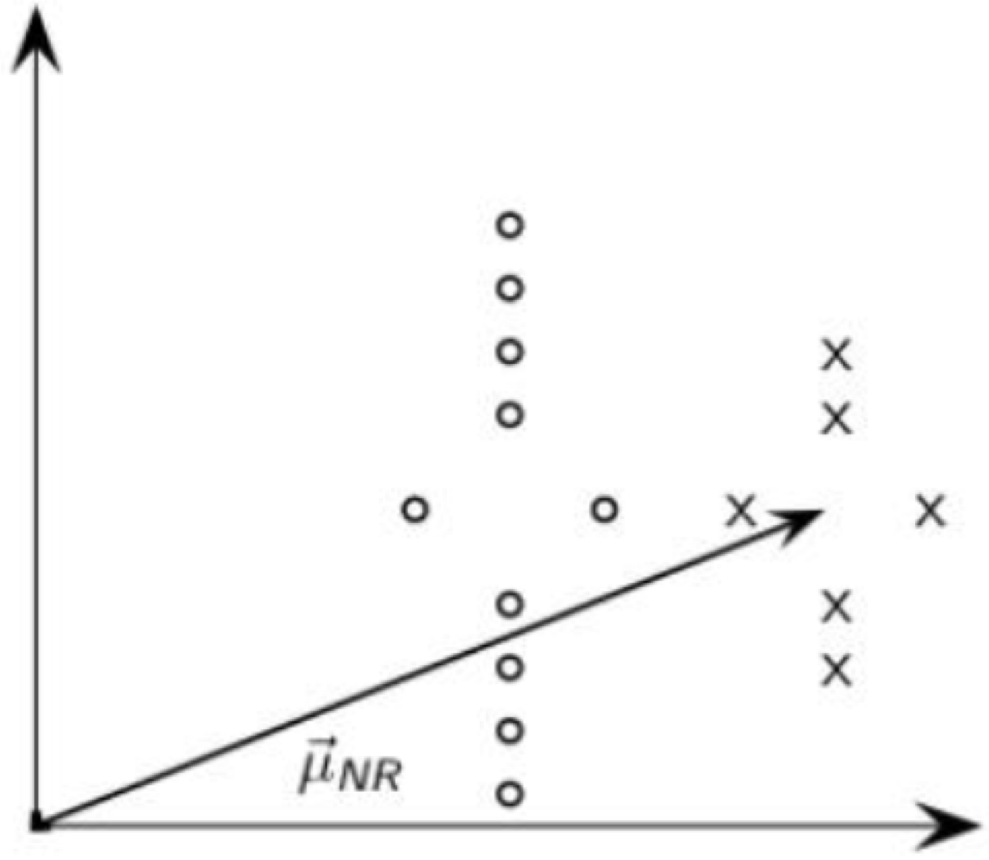
$\vec{\mu}_R$: centroid of relevant documents

Rocchio algorithm illustrated



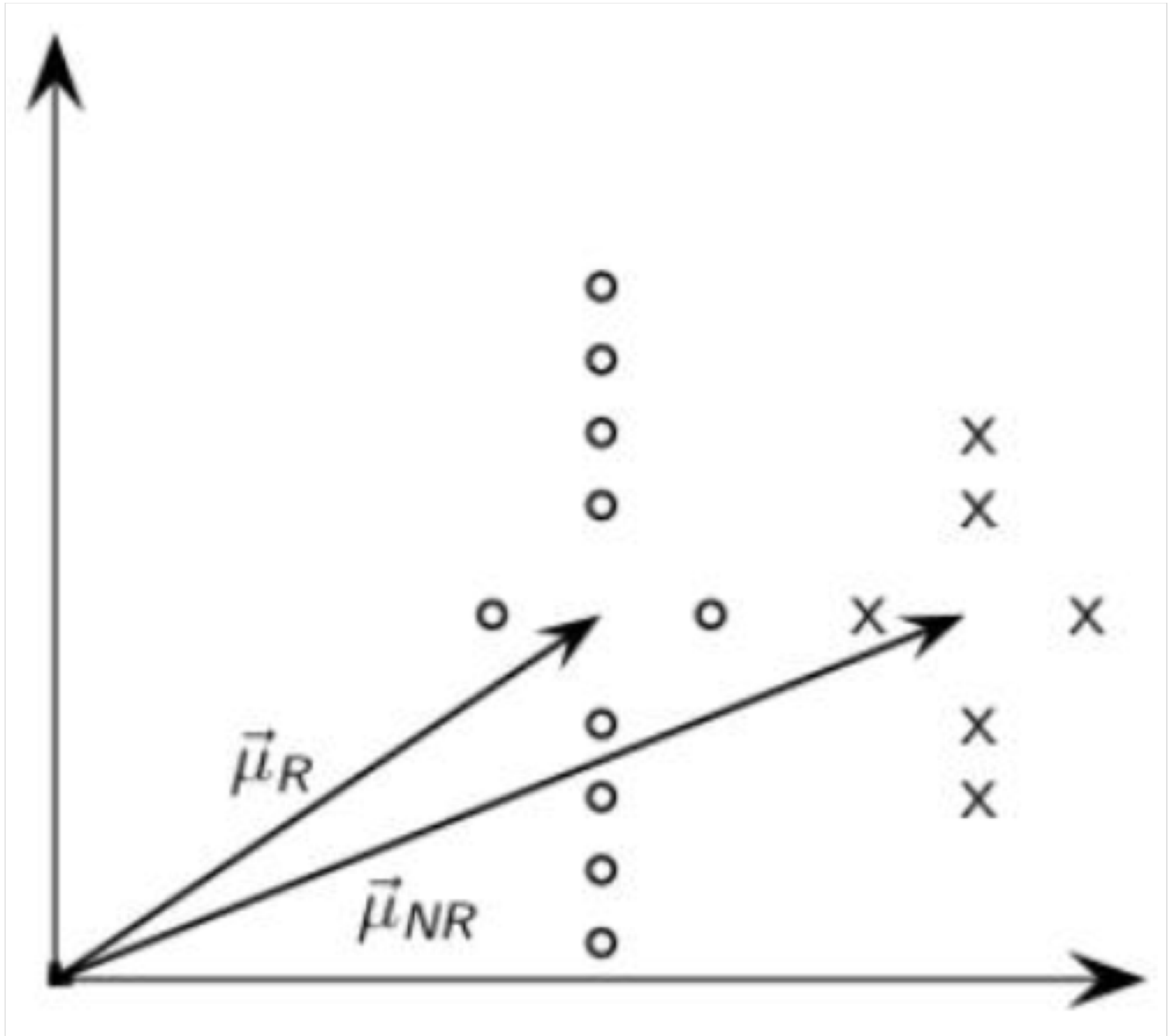
$\vec{\mu}_R$ does not separate relevant / nonrelevant.

Rocchio algorithm illustrated

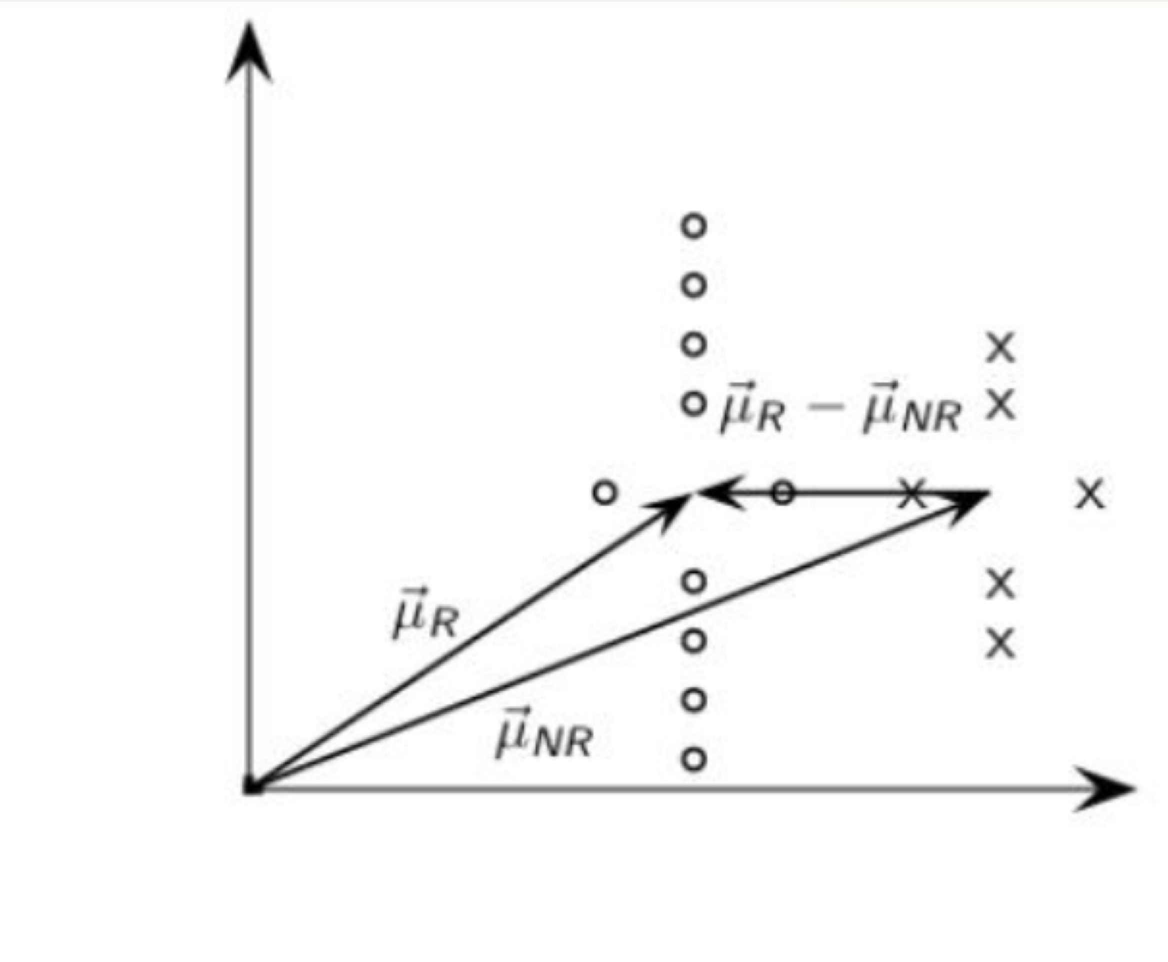


$\vec{\mu}_{NR}$: centroid of nonrelevant documents.

Rocchio algorithm illustrated

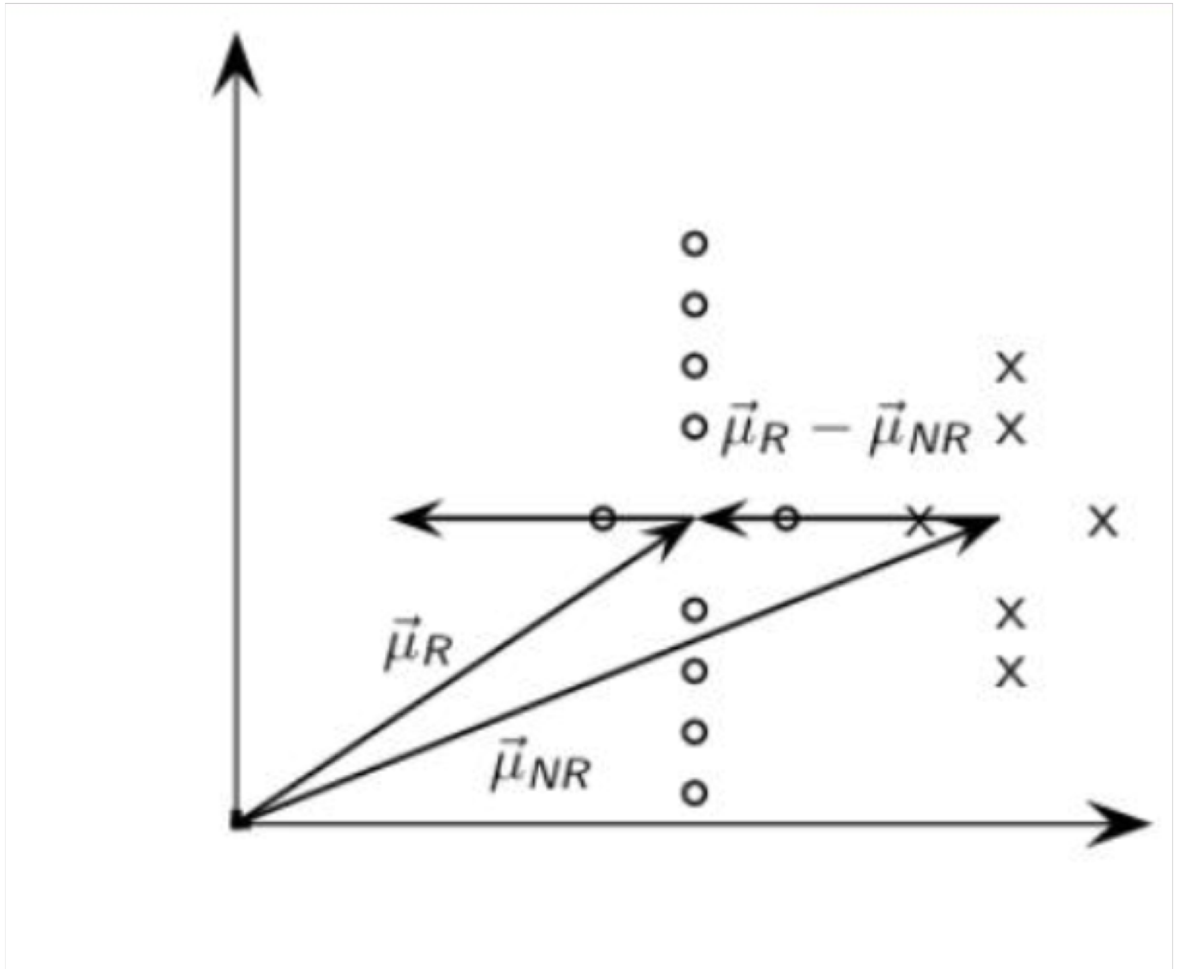


Rocchio algorithm illustrated



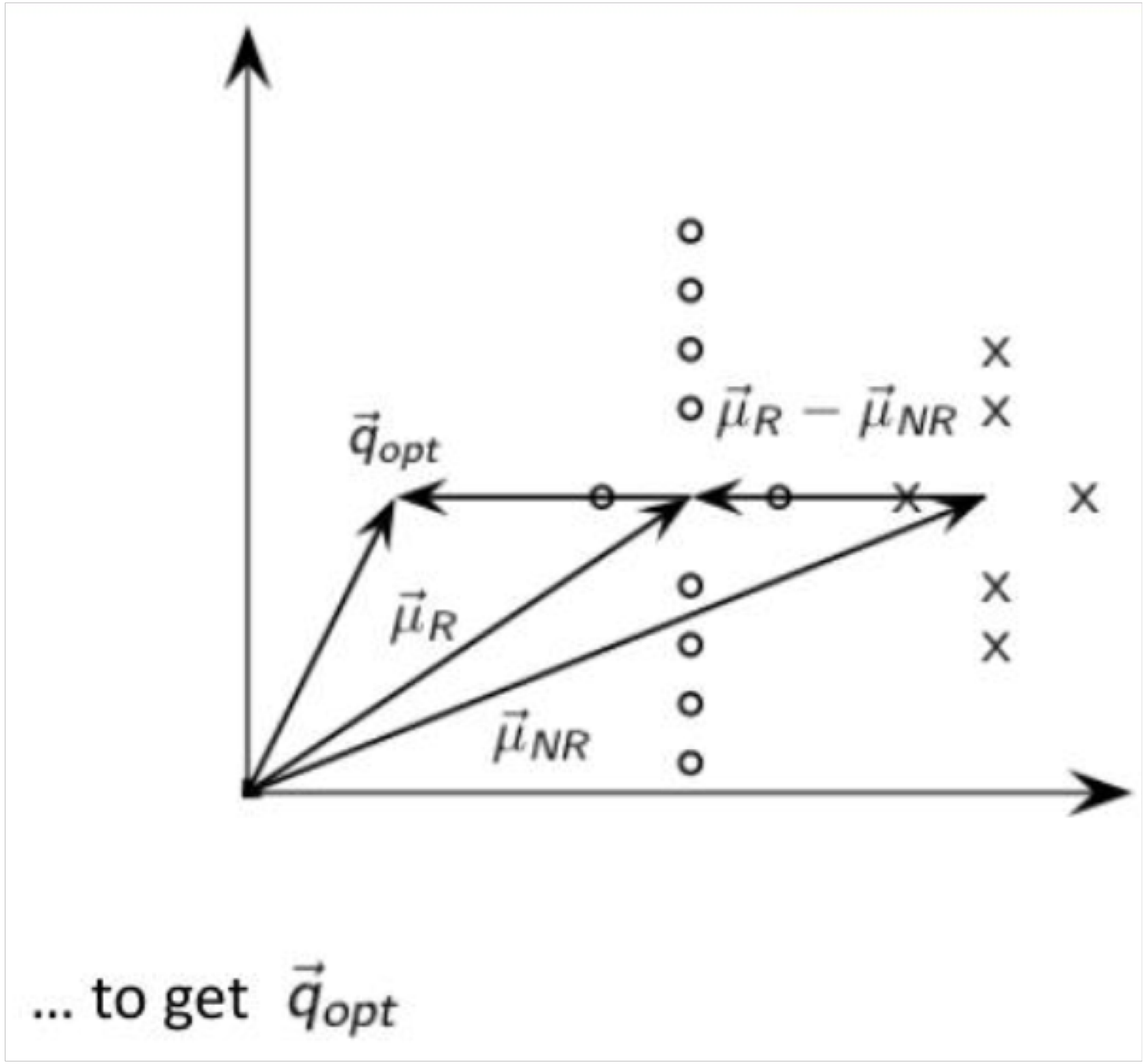
$\vec{\mu}_R$ - $\vec{\mu}_{NR}$: difference vector

Rocchio algorithm illustrated

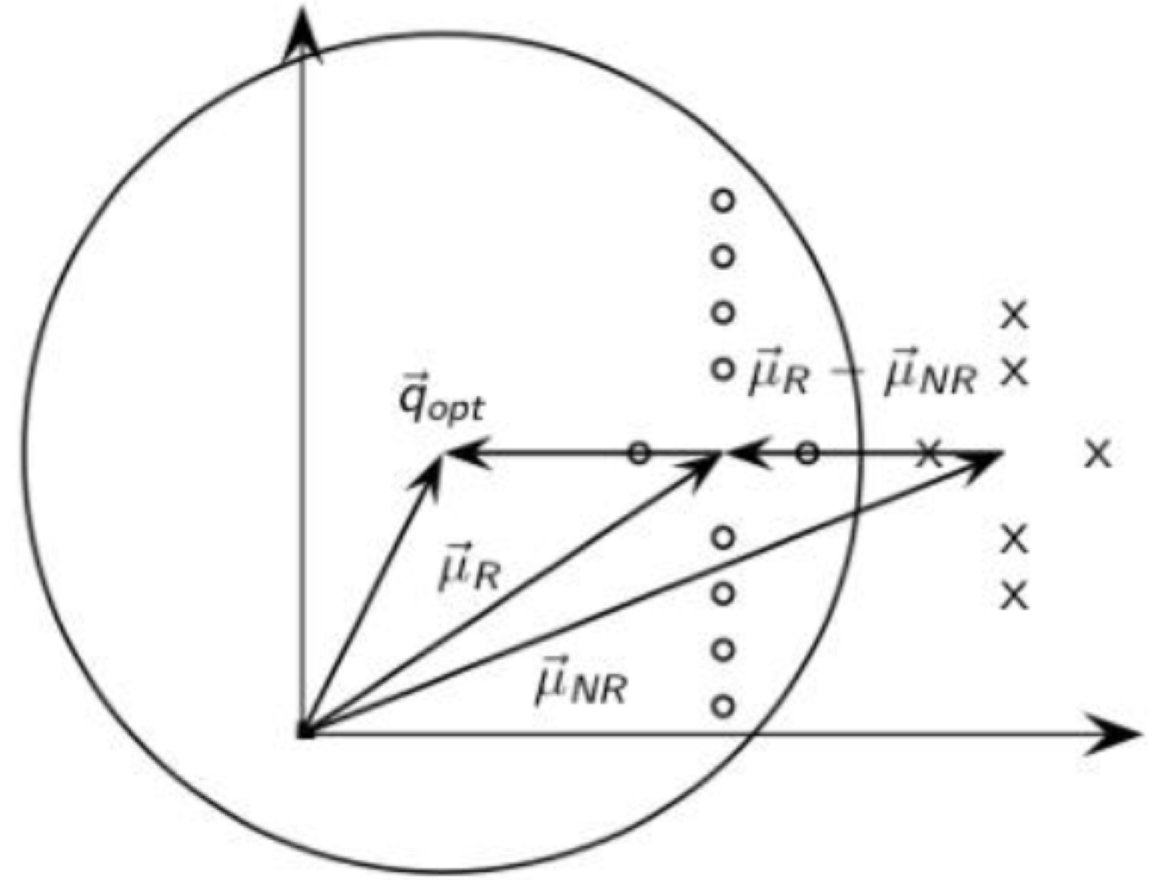


Add difference vector to $\vec{\mu}_R$ ...

Rocchio algorithm illustrated

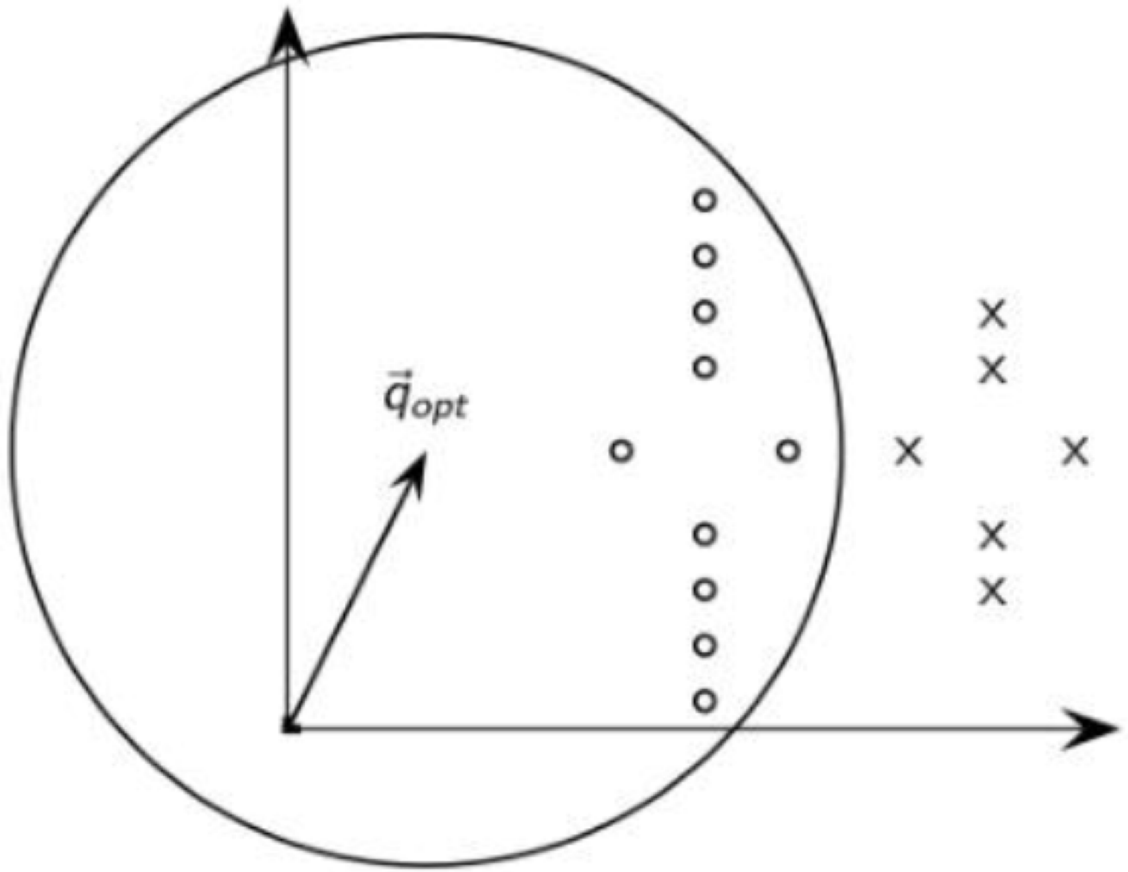


Rocchio algorithm illustrated



$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Rocchio algorithm illustrated



$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Rocchio 1971 Algorithm (SMART)

D_r : set of relevant and retrieved documents
D_n: set of non-relevant and retrieved documents

$$\vec{q}_{next} = \alpha \cdot \vec{q}_{prev} + \beta \cdot \frac{1}{|D_r|} \sum_{d_i \in D_r} \vec{d}_i - \gamma \cdot \frac{1}{|D_n|} \sum_{d_i \in D_n} \vec{d}_i$$

The factors α, β, γ control the effect of previous query, relevant documents and non-relevant documents on the new query

Rocchio 1971 Algorithm (SMART)

$$\vec{q}_{next} = \alpha \cdot \vec{q}_{prev} + \beta \cdot \frac{1}{|D_r|} \sum_{d_i \in D_r} \vec{d}_i - \gamma \cdot \frac{1}{|D_n|} \sum_{d_i \in D_n} \vec{d}_i$$

- Usually information in relevant documents more important than in non-relevant documents ($\gamma \ll \beta$).
- Positive relevance feedback ($\gamma = 0$) is when we only extract information from documents assessed relevant.
- α emphasises the importance of the original query ($\vec{q}_{prev}$).

Rocchio in practice

$$\vec{q}_{next} = \vec{q}_{prev} + \beta \cdot \frac{1}{|D_r|} \sum_{d_i \in D_r} \vec{d}_i - \gamma \cdot \frac{1}{|D_n|} \sum_{d_i \in D_n} \vec{d}_i$$

- $\alpha=1$
- Terms forming the reformulated query ($\vec{q}_{prev}$) are those:
  - in the original query,
  - that appear in more relevant documents that non-relevant documents
  - that appear in more than half of the relevant documents
- Negative weights ignored

Relevance feedback - Ide

- Ide developed three strategies extending Rocchio's approach:
  - Basic Rocchio's formula minus the normalization for the number of relevant and non-relevant documents
  - Allowed only feedback from relevant documents
  - Allowed limited negative feedback from only the highest ranked non-relevant document

$$\vec{q}_{next} = \alpha \cdot \vec{q}_{prev} + \beta \cdot \sum_{d_i \in D_r} \vec{d}_i - \gamma \cdot \sum_{d_i \in D_n} \vec{d}_i$$

Start with $\alpha = \beta = \gamma = 1$.

The cardinalities of the sets of relevant and non-relevant documents are not considered.

Issues with relevance feedback

- Increased burden on the user: users don’t like providing constant feedback; increased cognitive load
- Often users are not reliable in making relevance assessments, or do not make relevance assessments
- Partial relevance assessments (e.g very relevant or partially relevant): users don’t explicitly provide this type of information
- Why is a document relevant? Even if we get relevance feedback from the user, it is not always clear why positive/negative feedback was provided.

Relevance feedback: Evaluation

- Pick one of the evaluation measures from previous lectures, e.g. Precision@K
- Compute Prec@K for original query q_0
- Compute Prec@K for modified relevance feedback query q_1
- Fair evaluation must be on ”residual” collection, i.e. documents not yet judged by user
- Empirically, one round of relevance feedback is often very useful. Two rounds are marginally useful.

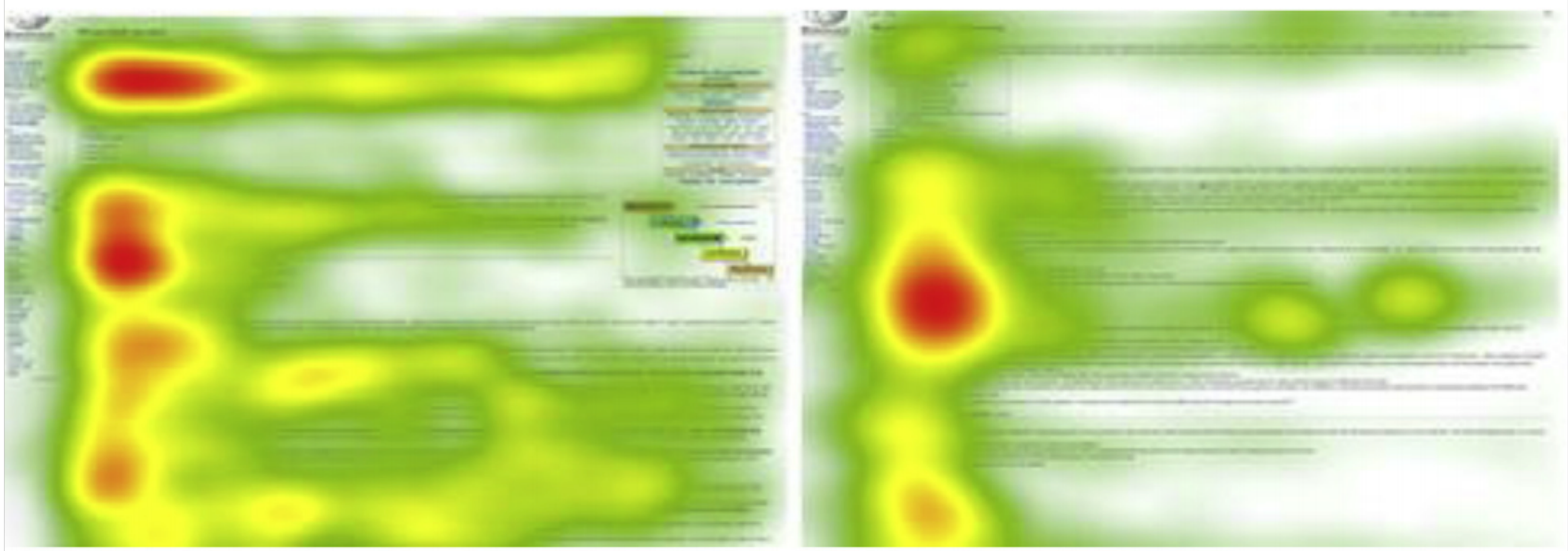
Relevance feedback methods

Parameters	Description
Dwell Time (DT):	This is the accumulated time in seconds spent by a user on an active page during browsing. It is also called reading time.
Distance of Mouse Movement (DMM)	The Euclidean distance of mouse movement is calculated by its X and Y coordinates on the monitor in every 100 ms.
Total Mouse Movement (TMM):	This is the total mouse movement calculated by its X and Y coordinates on the monitor. The count is incremented by one for X and Y as the mouse hovers.
Mean Mouse Velocity (MMV)	This is the total speed covered by the mouse on the monitor.
Number of Mouse Clicks (NMC)	This is the total amount of mouse clicks on a page. The number of mouse click is incremented every time the mouse is clicked by a user.
Amount of Scroll (AS)	Most web pages are longer in length than the monitor height. When readers are interested in a page, they scroll the page. The scrolling is normally done by either clicking or dragging the scroll bar. Any time a user clicks the scrollbar up or down, the count is incremented.
Number of Keystrokes (NK):	This is the total number of keystrokes on a document. This is incremented when the user strikes a key.
Amount of Copy (AC)	This is the number of times text is copied to the clipboard from a document. It is incremented by one any time text from a particular document is copied.
Mouse Duration Count (DC)	This is the total number of 100 ms intervals that occurs while the mouse is moved on the screen.
Time Stamp	This is the time and date in GMT when a document is loaded and when a document is closed.
URL	This is the http address of any web document visited by a user.
IP Address (IP)	This is the internet protocol address of a user. It represents the user's location.
Explicit Relevance Ratings (ER)	This is the actual rating of the web document by the user. The Firefox plugin attaches a six scale rating button on each of the webpages. After reading a webpage, the user rates it by clicking on any of six scale buttons where 5 – means very relevant, 4 – means more relevant, 3 – means moderate relevant, 2 – means slightly relevant, 1– means very low relevance, 0 – means not relevant.

S. Akuma, R. Iqbal, C. Jayne, F. Doctor. 2016. *Comparative analysis of relevance feedback methods based on two user studies.* Computers in Human Behavior 60. 138 – 146.

Relevance feedback: eye gaze

Parameters	Description
Total Fixation Duration (TFD):	This is the sum of duration of all individual fixations within a specific area of interest of a document. Individual fixation is between 250 ms and 300 ms.
Total Fixation Count (TFC)	This is number of times that a user fixates within a specific area of interests of a document.
Heat Map	This is a visualization technique that separates different levels of fixation intensity, it show areas that are more fixated to be denser than areas that are less fixated.

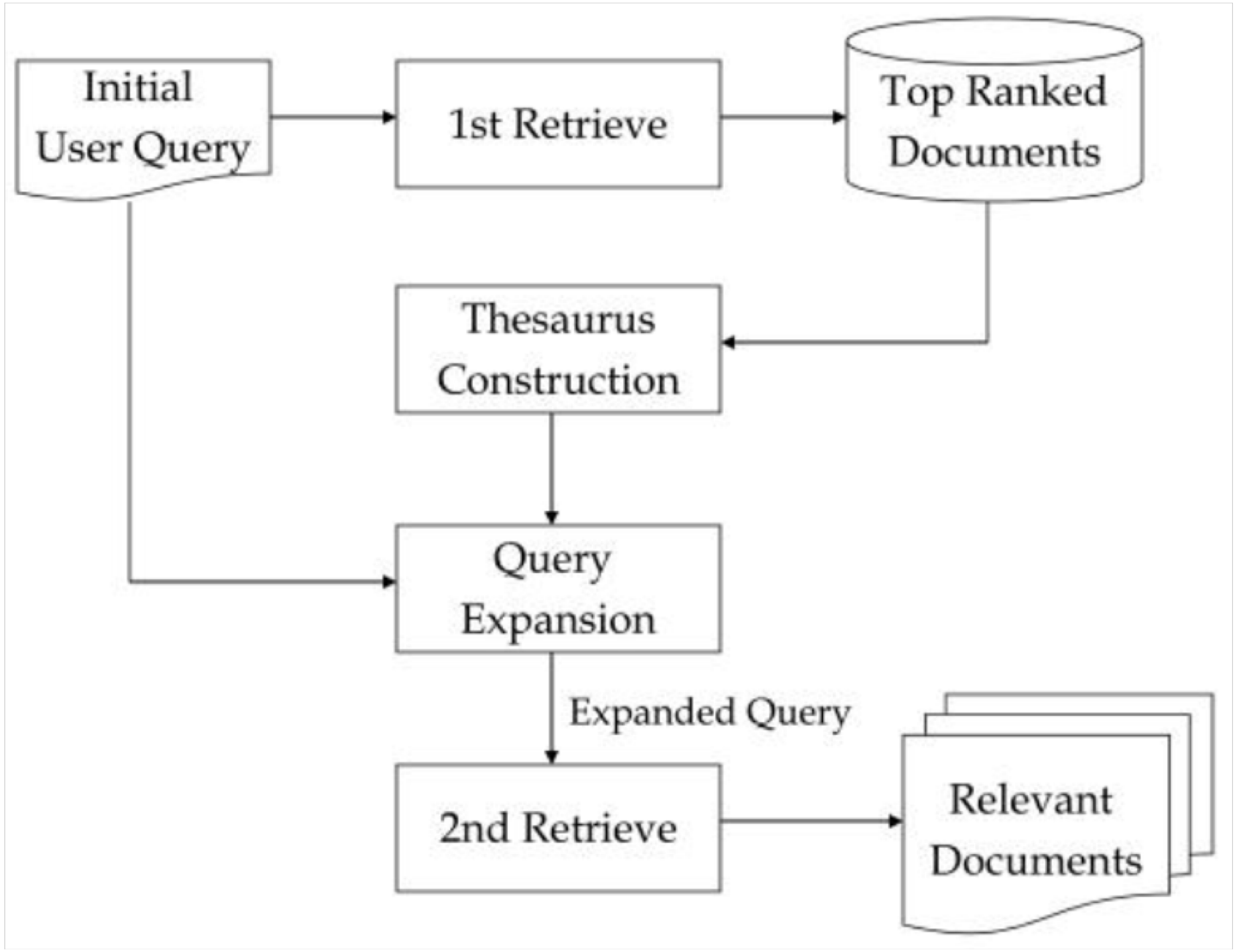


a) Highest mean fixation count b) Lowest mean fixation count

Local Analysis

- Examine documents retrieved for query to determine query expansion with no user assistance
- Two strategies are used to add terms to the query:
  - Local clustering (terms that are synonyms, stemming variations)
  - Local context analysis (terms close to query terms in text proximity of terms in text)
- Two issues:
  - Query drift: if top documents are not that relevant, the reformulated query may not reflect the user information need
  - Computation cost high since must be done at retrieval time (on-line)

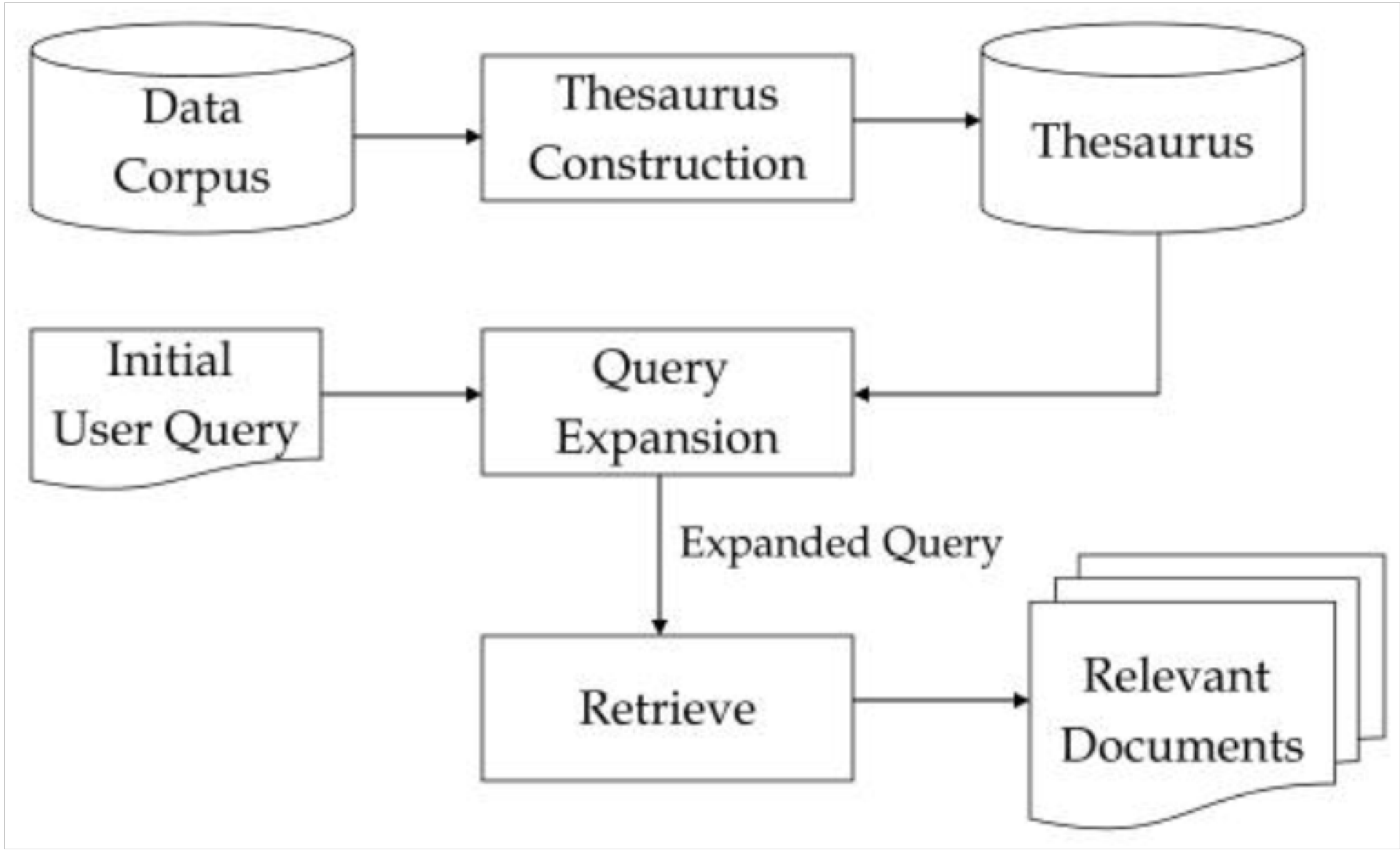
Local Analysis



Global Analysis

- Expand query using information from whole set of documents in collection
- No user assistance
- Make use of of a global thesaurus that is build based on the document collection.
- Two issues:
  - Approach to build thesaurus (e.g. term co-occurrence)
  - Approach to select terms for query expansion (e.g. the top 20 terms ranked according to IDF value)
- session analysis (queries used in same sessions as analyzed from logs) for query recommendation/suggestion

Global Analysis



Dictionary-based query expansion

- For each term *t* in the query, expand the query with words the thesaurus lists as semantically related *t*, e.g. *aircraft -> plane*
- Generally increases recall
- May significantly decrease precision, particularly with ambiguous terms, e.g. *interest rate vs. develop an interest in*
- Widely used in specialised search engines, e.g. scince, medicine, engineering
- It is expensive to create a manual thesaurus and maintain it over time.

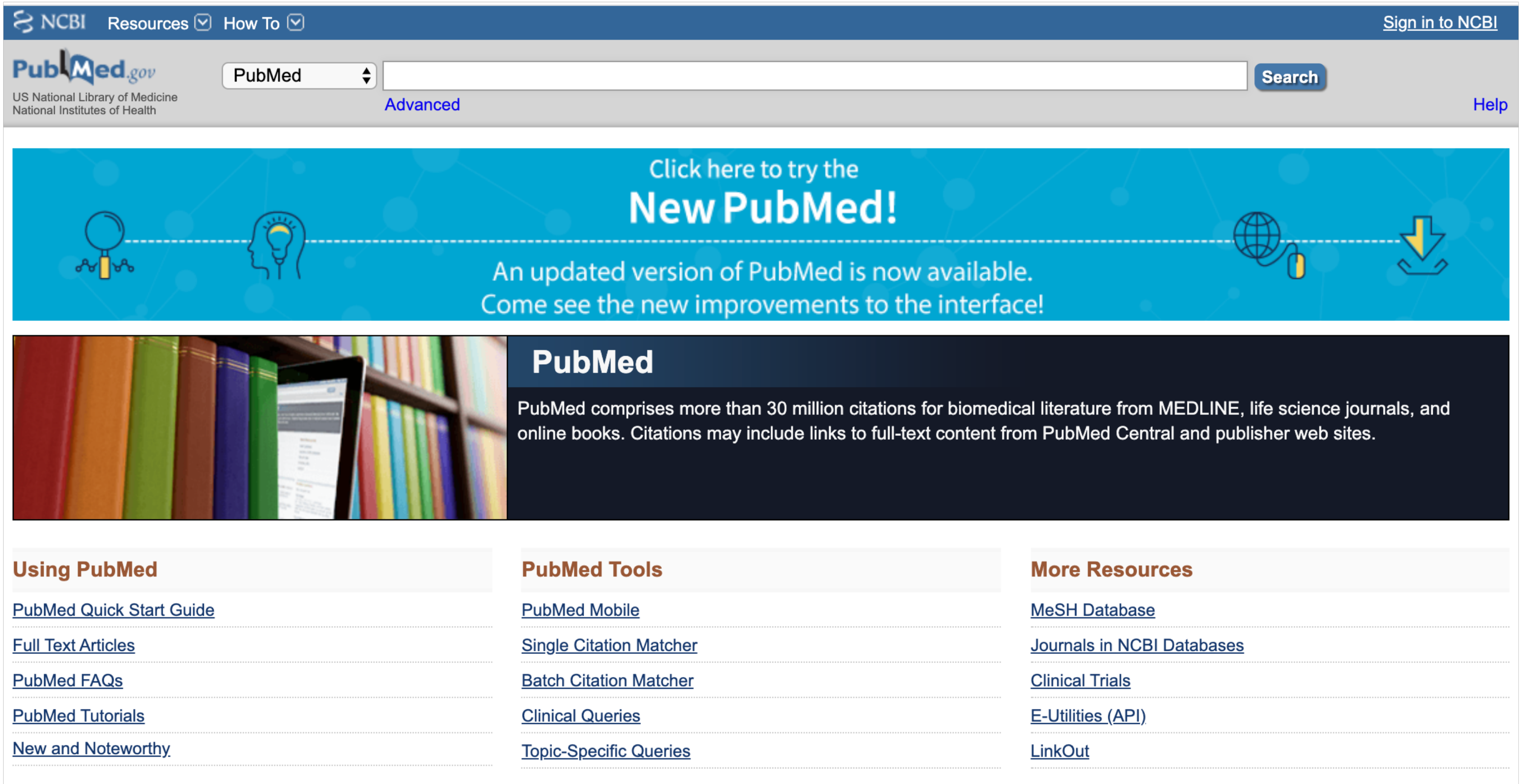
Dictionary-based Query Expansion

- Based on manual thesauri, e.g. WordNet
- In the expansion process, synonymous words of initial query terms are selected and assigned weights
- Disadvantages:
  - Manual thesaurus construction is labour intensive
  - A general thesaurus does not consistently improve retrieval perfomance

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V. Aj, Av	sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	powerful, powerless rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple tree, plant
Meronymy (part)	N	brim, hat gin, martini
Troponymy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
Note: N = Nouns, Aj = Adjectives, V = Verbs, Av = Adverbs		

WordNet example

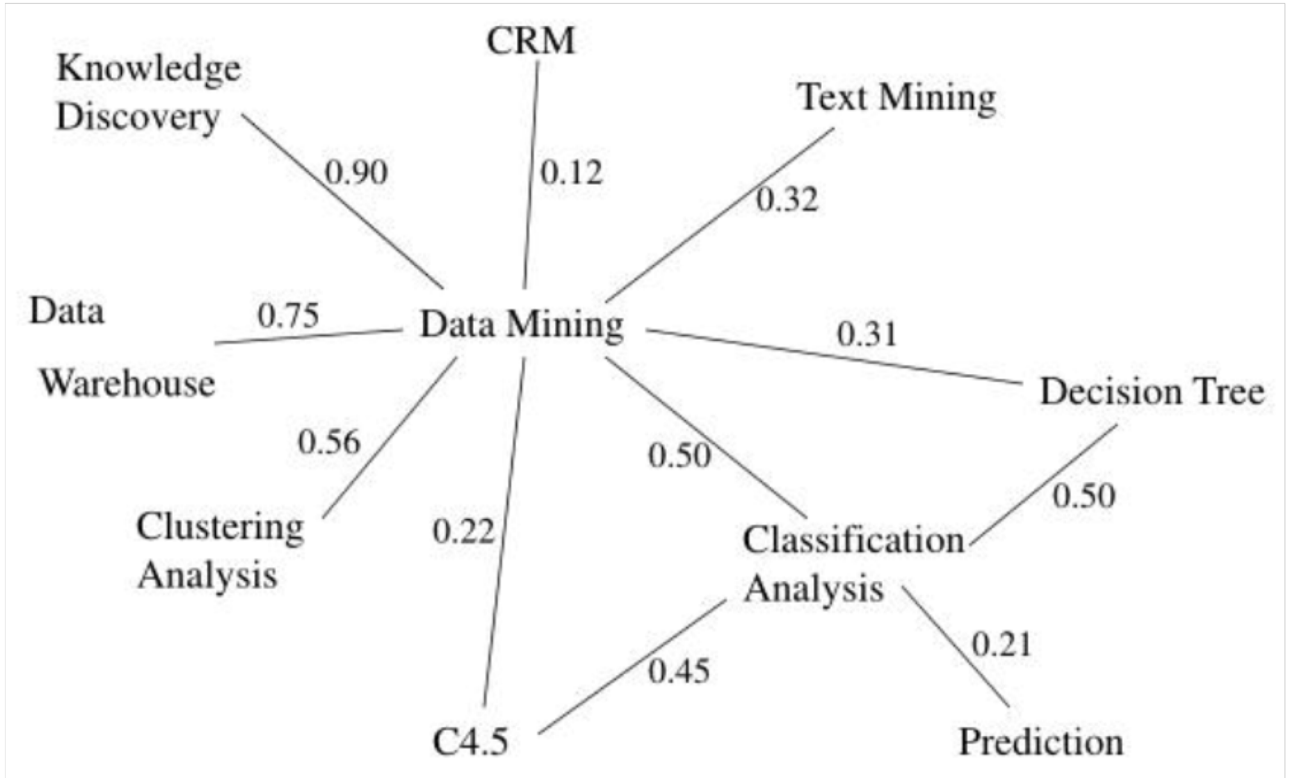
Example of manual thesaurus: PubMed



Automatic Thesauri Construction

Thesauri are constructed from the data corpus:

- Term co-occurence
- Traditional or a variant of TF-IDF
- Mining association rules



Automatic thesaurus generation

- Attempt to generate a thesaurus automatically by analysing the distribution of words in documents
- Fundamental notion: similarity between two words
- Two words are ***similar if they co-occur with similar words***
  - *car* and *motorcycle* are similar becuase both occur with *road, petrol, licence*
- 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-occurence is more robust while grammatical relations are more accurate

Automatic thesaurus construction: discussion

- Quality of term associations is usually an issue
- Term ambiguity may introduce irrelevant statistically correlated terms
  - Apple computer --> apples and computers

- Problems:
  - False positives: words deemed similar that are not
  - False negatives: words deemed dissimilar that are similar

- Since terms are highly correlated, expansion may not retrieve many additional documenst

Query expansion in search engines

- **Query logs** – main source of query expansion in search engines

- Example 1: after issuing the query *herbs*, users frequently search for *herbal remedies*
  - *herbal remedies* is a potential expansion of *herb*

- Example 2: user searching for *car pictures* frequently click on the same URL as users searching for *car photos*
  - *car photos* and *car pictures* are potential expansions of each other

Resources

- Chapters 9 and 14 of Introduction to Information Retrieval
- Chapter 5 of Modern Information Retrieval