
Chapter 5

Query Operations

Motivation - Feast or famine

- Queries return *either too few or too many results*
- Users are generally looking for *the best document* with a particular piece of information
- Users don't want to look through hundreds of documents to locate the information

⇒ Rank documents according to expected relevance!

Relevance Feedback

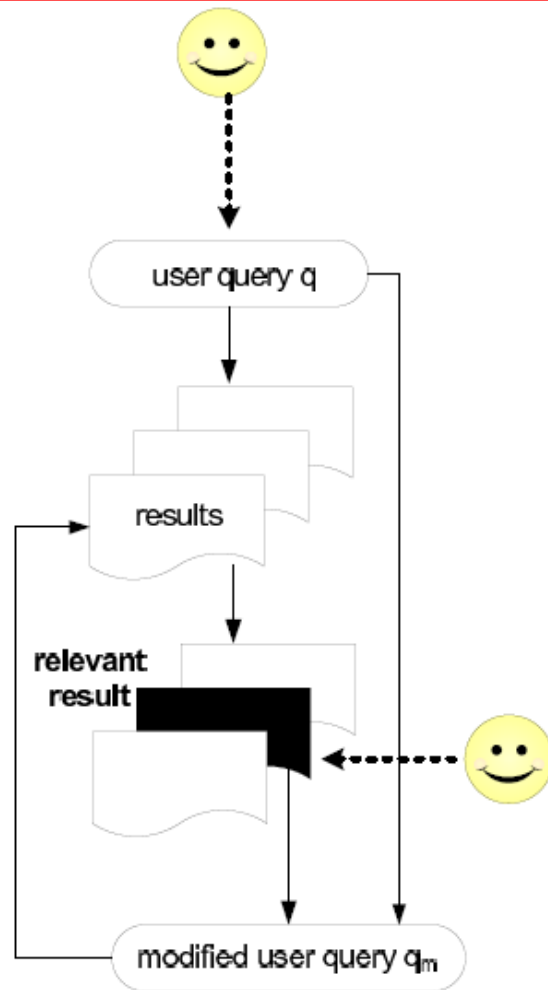
Queries

- Most queries are short
 - One to three words
- Many queries are ambiguous
 - “Saturn”
 - Saturn the planet?
 - Saturn the car?

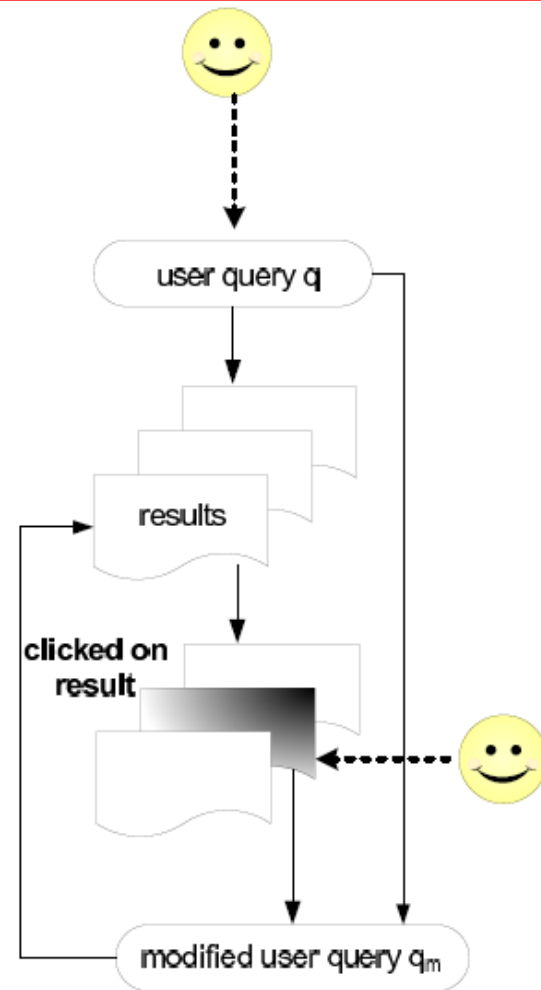
Relevance Feedback

- Two general approaches:
 - Create new queries with *user feedback* (*explicit feedback*)
 - Create new queries *automatically* (*implicit feedback*)
- Re-compute document weights with new information
- Expand or modify the query to more accurately reflect the user's desires

Explicit Feedback

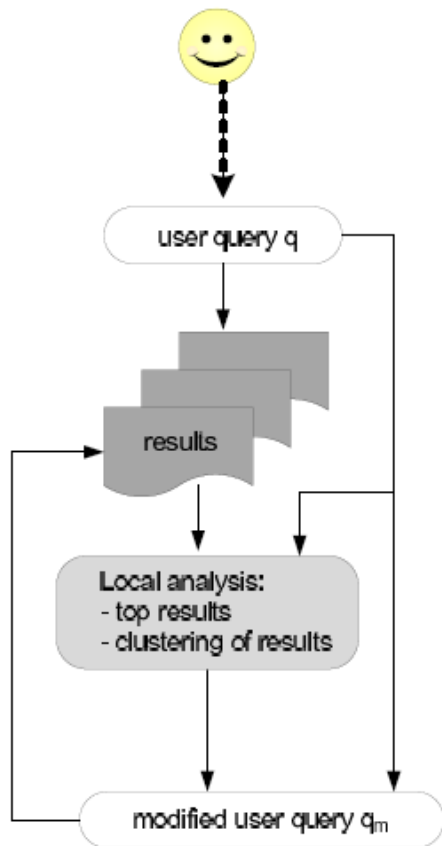


(a) **relevance feedback**

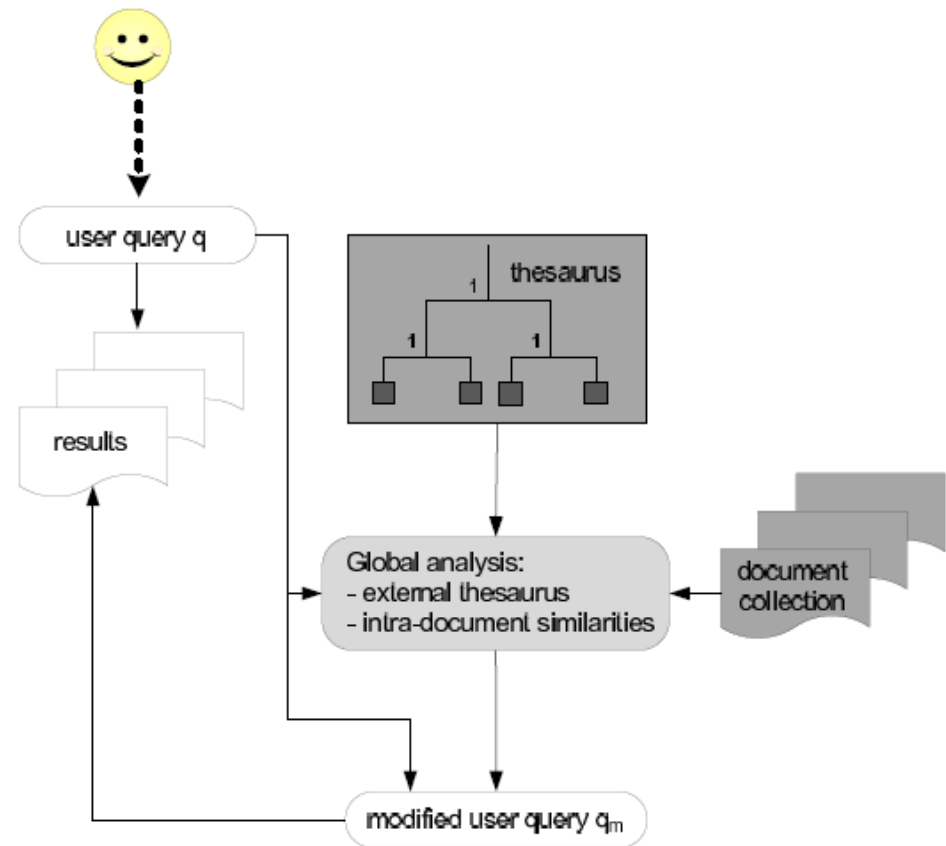


(b) **click feedback**

Implicit Feedback



(a) local analysis



(b) global analysis

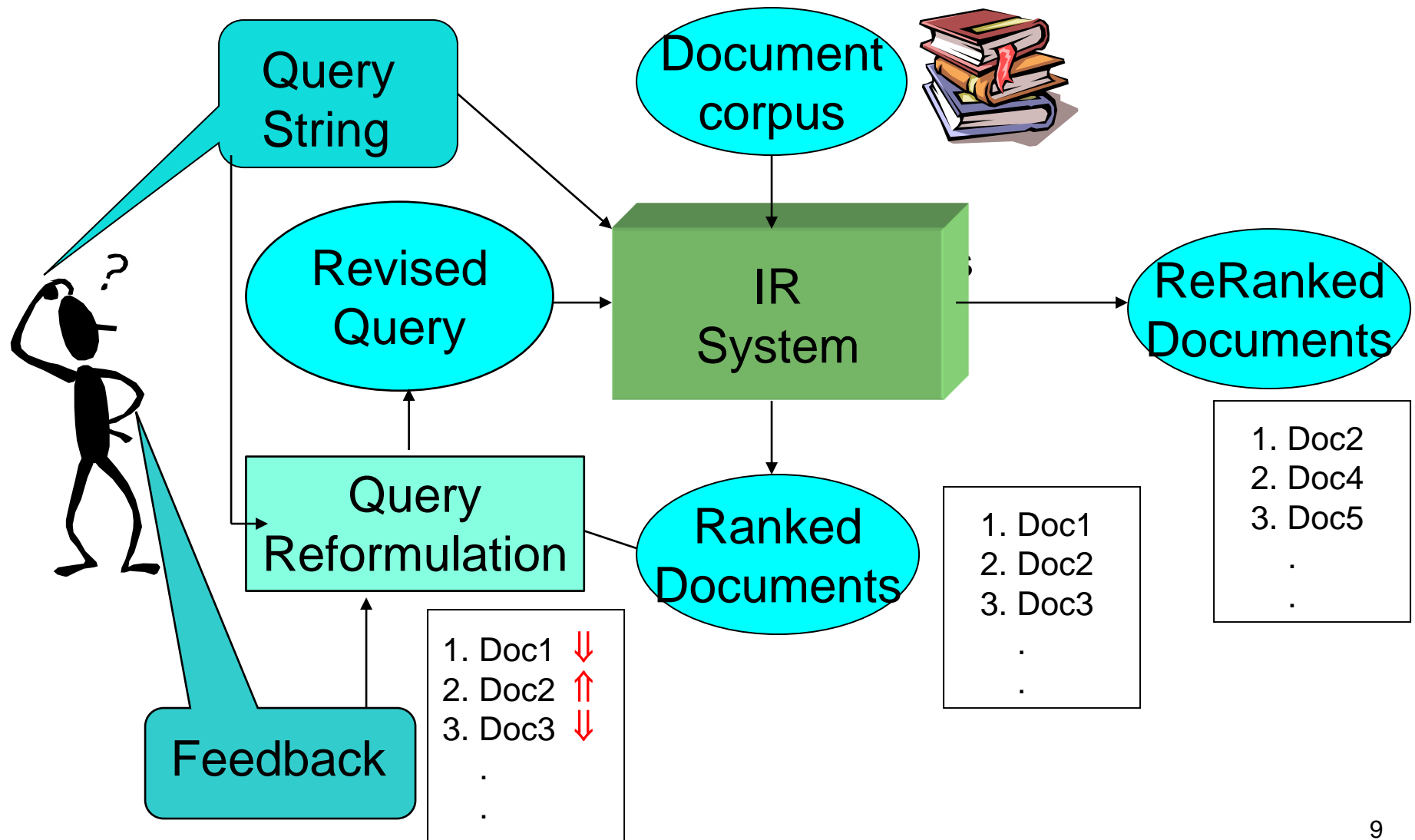
User Feedback

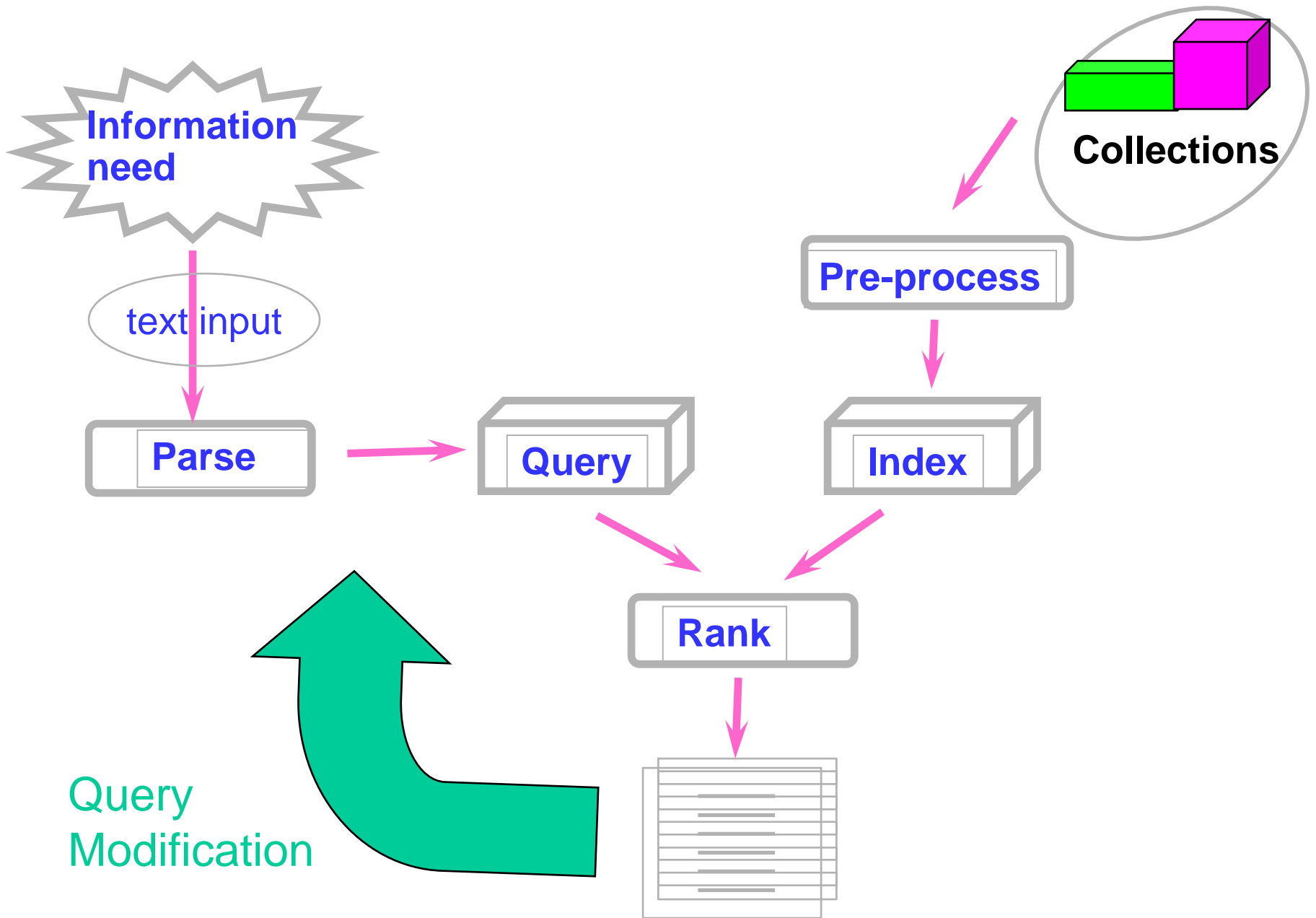
- After initial retrieval results are presented, allow the user to provide feedback on the relevance of one or more of the retrieved documents.
- Use this feedback information to reformulate the query.
- Produce new results based on reformulated query.
- Allows more interactive, ***multi-pass process***.

User Feedback

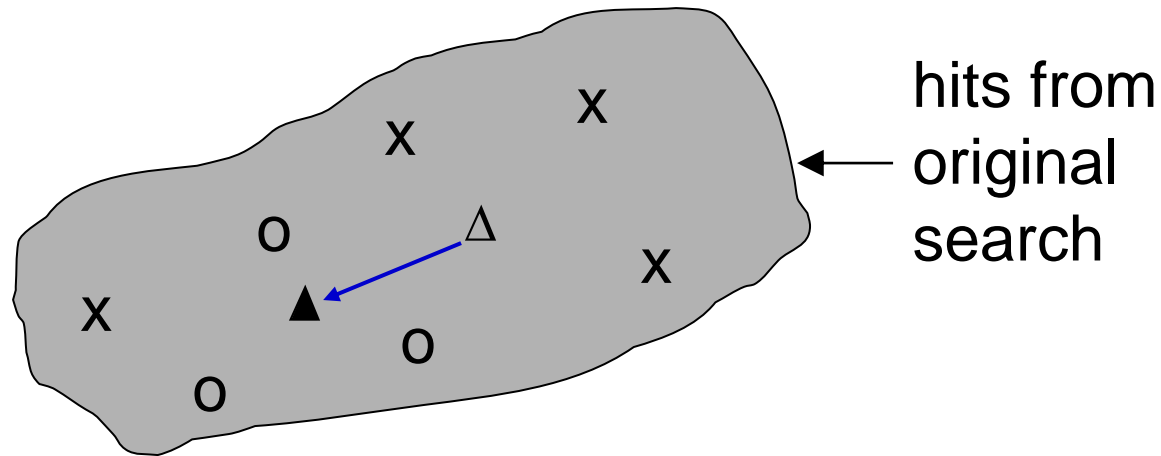
- ❑ The main idea consists of
 - selecting important terms from the documents that have been identified as relevant, and
 - enhancing the importance of these terms in a new query formulation

User Feedback Architecture





User Feedback (concept)



x documents identified as non-relevant

o documents identified as relevant

▲ original query
reformulated query

User Feedback (concept)

eResponder

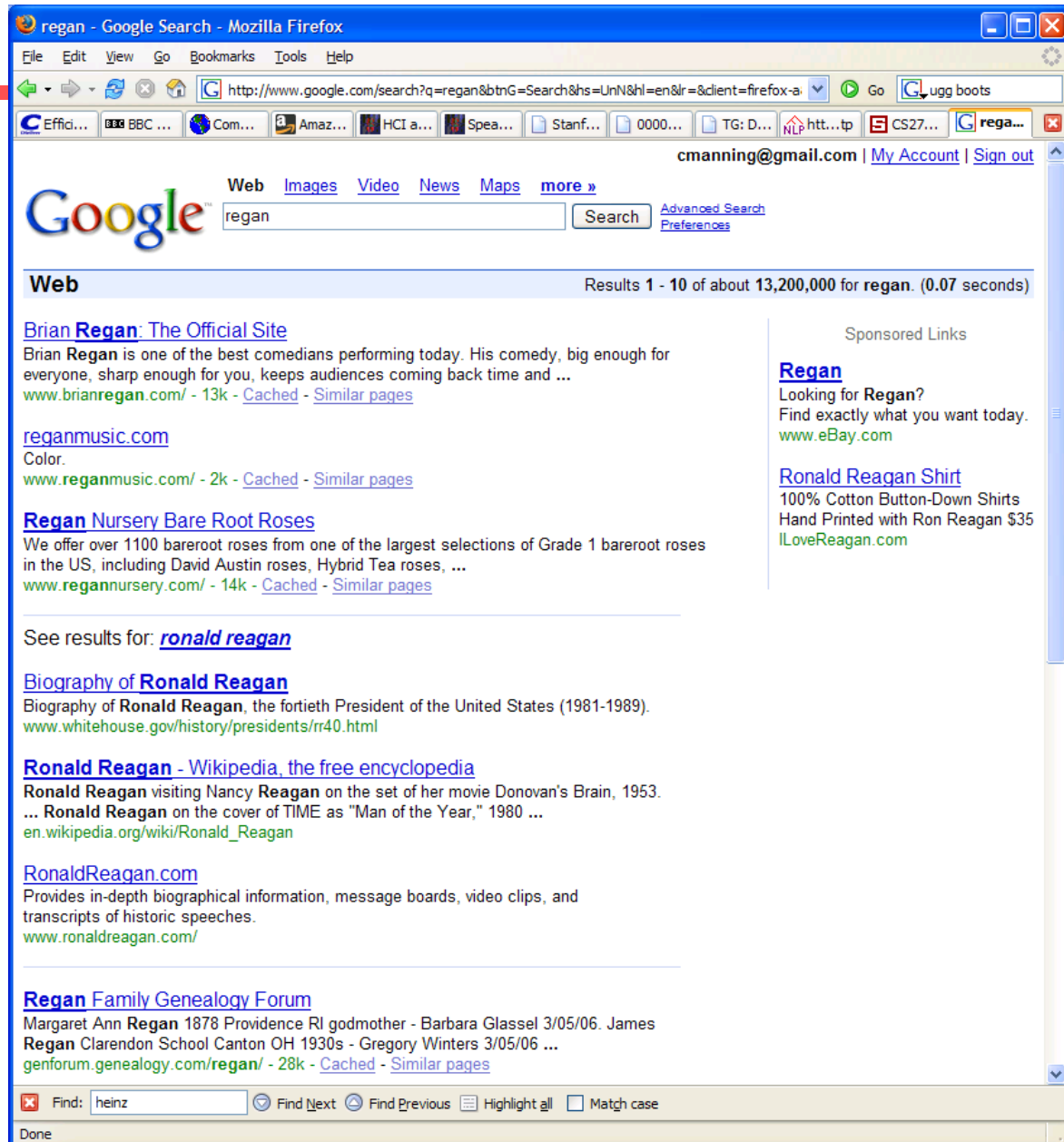
DB list: dbMapuccino, JavaFaq, HeartQA, NsfaskAScientist, QAdb, dbHelpNow

New question: How do I license Mapuccino?

Neutral/Relevant/Non-Relevant	Similar Questions	Q-similarity	A-relevance	AMR Score
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino licensing Hi, I would like to license Mapucc...	100.0	96.809	99.361
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: RE: Mapuccino Mr. Jacovi, Hello, and thank you for your...	85.886	96.857	88.08
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino - Licensing Information -Reply Hello Michael,	87.44	24.352	74.823
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Hi: Can you give me some more details on the evaluation and lic...	56.437	30.282	51.206
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: mapuccino Hello. I was just wondering if the Classes th...	35.545	96.809	47.798
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: quick question Hey Mapuccino! I like your product a lot,...	35.011	96.495	47.308
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	RE: InTRANet Mapuccino Hi Michal, Thanks for your reply. We wou...	57.447	0.0	45.958
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino Hello. I'm a Web content developer in the In...	37.667	77.297	45.593
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino I like your Mapuccino product and would like ...	36.949	78.264	45.212
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino Hello: I would like to use Mapuccino on the ...	35.767	71.328	42.879
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino Very interesting applet. Can it 1) be brought...	28.221	95.15	41.607
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Hi Michal, Thanks for your reply. We would be very interested in ...	51.086	0.0	40.869
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Mapuccino for masters thesis Hello, my name is Olaf Be...	39.536	40.99	39.827
<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	Subject: Re: mapuccino Hello again. I am able to use the mapucc...	39.094	31.659	37.607

Answer: Hello Joel, Thanks for your interest in Mapuccino. The version that you might have seen on the IBM Corporate Java page at <http://www.ibm.com/java/mapuccino> is a light version with limited features and is available only as an evaluation copy for you

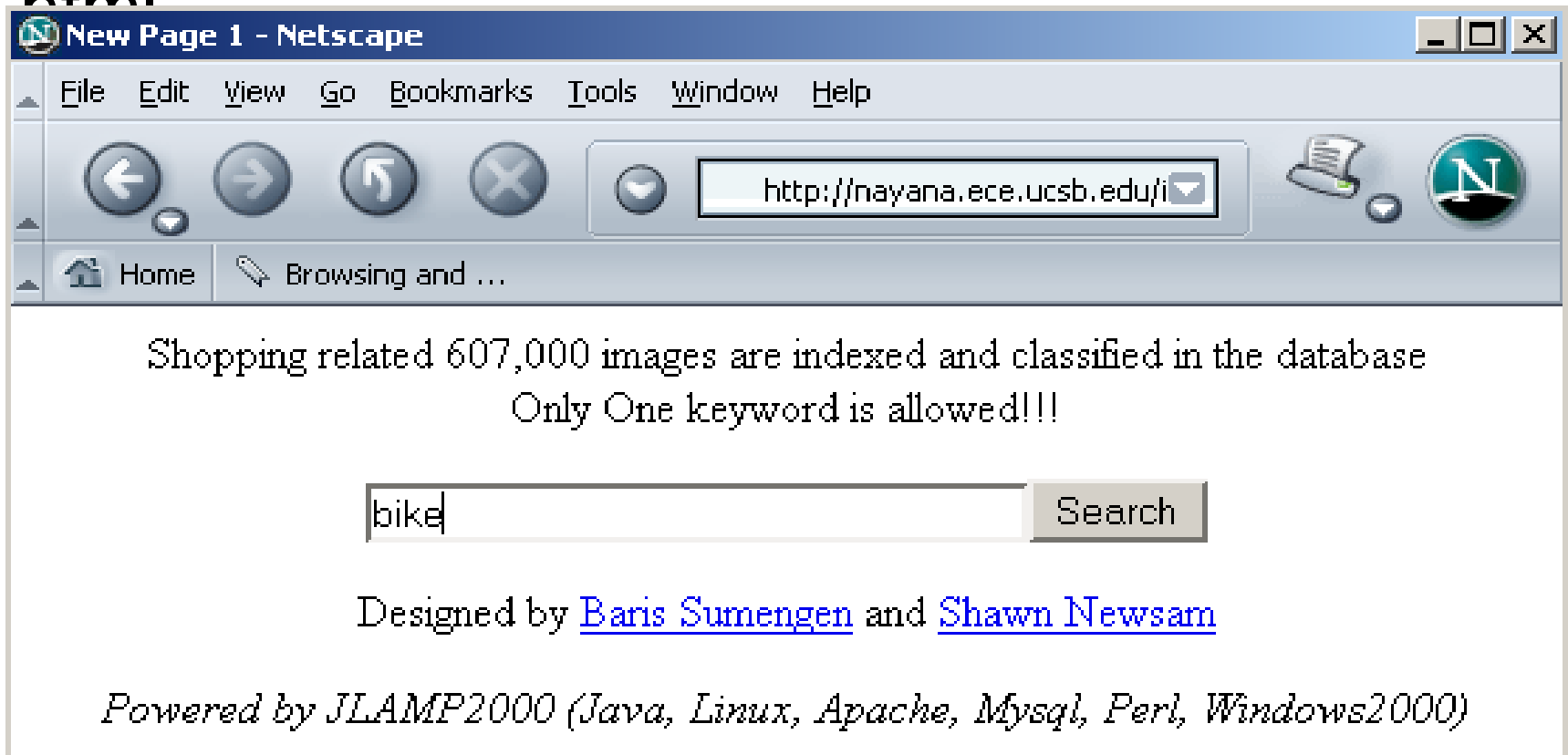
User Feedback (concept)








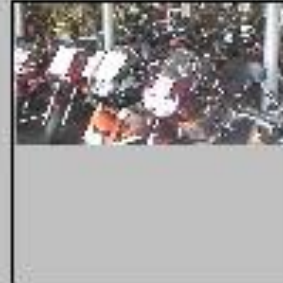






User Feedback: Example

- Image search engine


<http://nayana.ece.ucsb.edu/imsearch/imsearch.html>















Results for Initial Query

 <p>(144473, 16458) 0.0 0.0 0.0</p>	 <p>(144457, 252140) 0.0 0.0 0.0</p>	 <p>(144456, 262857) 0.0 0.0 0.0</p>	 <p>(144456, 262863) 0.0 0.0 0.0</p>	 <p>(144457, 252134) 0.0 0.0 0.0</p>	 <p>(144483, 265154) 0.0 0.0 0.0</p>
 <p>(144483, 264644) 0.0 0.0 0.0</p>	 <p>(144483, 265153) 0.0 0.0 0.0</p>	 <p>(144518, 257752) 0.0 0.0 0.0</p>	 <p>(144538, 525937) 0.0 0.0 0.0</p>	 <p>(144456, 249611) 0.0 0.0 0.0</p>	 <p>(144456, 250064) 0.0 0.0 0.0</p>








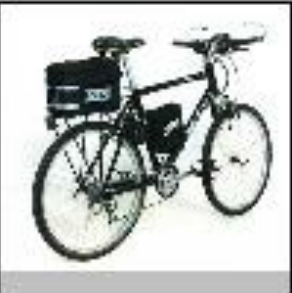




User Feedback



[Browse](#) [Search](#) [Prev](#) [Next](#) [Random](#)

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

Results after User Feedback

<div>BrowseSearchPrevNextRandom</div>					
					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4639 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391337	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

Query Reformulation

- Revise query to account for feedback:
 - **Query Expansion**: Add new terms to query from relevant documents.
 - **Term Reweighting**: Increase weight of terms in relevant documents and decrease weight of terms in irrelevant documents.
- Several algorithms for query reformulation.

Query Reformulation

- Change query vector using vector algebra.
- **Add** the vectors for the **relevant** documents to the query vector.
- **Subtract** the vectors for the **irrelevant** docs from the query vector.

Vector Space Re-Weighting

Rochio:

- $\mathbf{q}' = \alpha \mathbf{q} + (\beta/|\mathbf{D}_r|) \sum_{\mathbf{d}_i \in \mathbf{D}_r} \mathbf{d}_i - (\gamma/|\mathbf{D}_n|) \sum_{\mathbf{d}_i \in \mathbf{D}_n} \mathbf{d}_i$

Ide regular

- $\mathbf{q}' = \alpha \mathbf{q} + \beta \sum_{\mathbf{d}_i \in \mathbf{D}_r} \mathbf{d}_i - \gamma \sum_{\mathbf{d}_i \in \mathbf{D}_n} \mathbf{d}_i$

Ide Dec_hi

- $\mathbf{q}' = \alpha \mathbf{q} + \beta \sum_{\mathbf{d}_i \in \mathbf{D}_r} \mathbf{d}_i - \gamma \max_{\mathbf{d}_i \in \mathbf{D}_n} (\mathbf{d}_i)$

Rocchio Method

$$Q_1 = \alpha Q_0 + \frac{\beta}{n_1} \sum_{\forall d_j \in D_r} \vec{d}_j - \frac{\gamma}{n_2} \sum_{\forall d_j \in D_n} \vec{d}_j$$

where

Q_0 = the vector for the initial query

D_r = the set of relevant documents

D_n = the set of non - relevant documents

n_1 = the number of relevant documents chosen

n_2 = the number of non - relevant documents chosen

α, β and γ tune importance of relevant and nonrelevant terms

(in some studies best to set α to 1 β to 0.75 and γ to 0.25)

Example Rocchio Calculation

$$R_1 = (0.030, 0, 0, 0.025, 0.025, 0.050, 0, 0, 0.120)$$

Relevant
docs

$$R_2 = (0.020, 0.009, 0.020, 0.002, 0.050, 0.025, 0.100, 0.100, 0.120)$$

$$S_1 = (0.030, 0.010, 0.020, 0, 0.005, 0.025, 0, 0.020, 0)$$

Non-rel doc

$$Q = (0, 0, 0, 0, 0.500, 0, 0.450, 0, 0.950)$$

Original Query

$$\alpha = 1$$

$$\beta = 0.75$$

Constants

$$\gamma = 0.25$$

$$Q_{new} = \alpha \times Q + \left(\frac{\beta}{2} \times (R_1 + R_2) \right) - \left(\frac{\gamma}{1} \times S_1 \right)$$

Rocchio Calculation

Resulting feedback query

$$Q_{new} = (0.011, 0.000875, 0.002, 0.01, 0.527, 0.022, 0.488, 0.033, 1.04)$$

Rocchio Method - summary

- Rocchio automatically
 - re-weights terms
 - adds in new terms (from relevant docs)
 - have to be careful when dealing with negative terms
 - known to significantly improve results
- Quality
 - heavily dependent on test collection
 - heavily dependent on relevance quality

Ide Regular Method

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

$$\vec{q}_1 = \alpha \vec{q}_0 + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

α : Tunable weight for initial query.

β : Tunable weight for relevant documents.

γ : Tunable weight for irrelevant documents.

Relevance Feedback

$$\vec{q}_m = \vec{q} + \alpha \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \beta \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Original Query : (5,0,3,0,1)

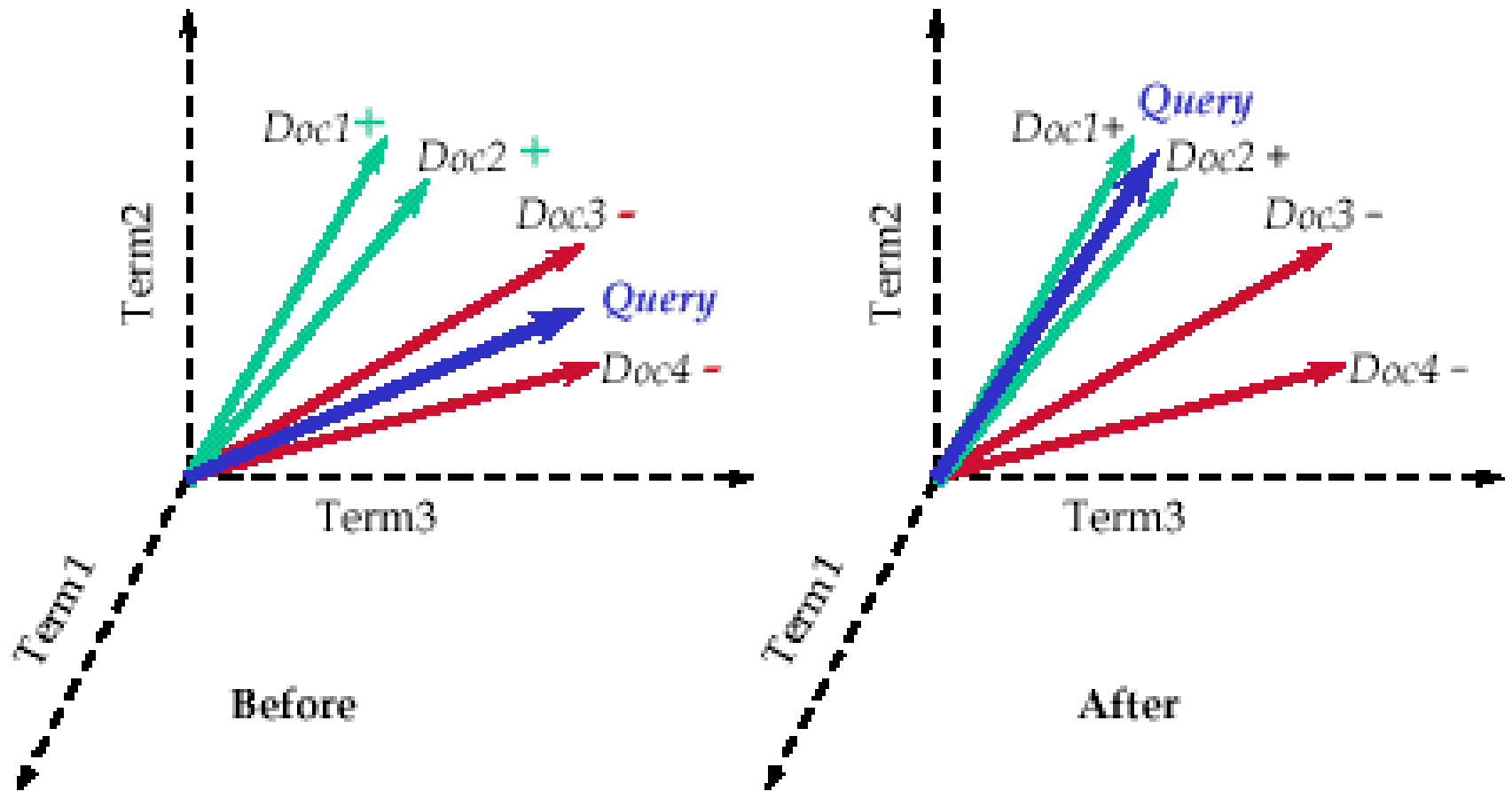
Document D₁ Relevant : (2,1,2,0,0)

Document D₂ Nonrelevant : (1,0,0,0,2)

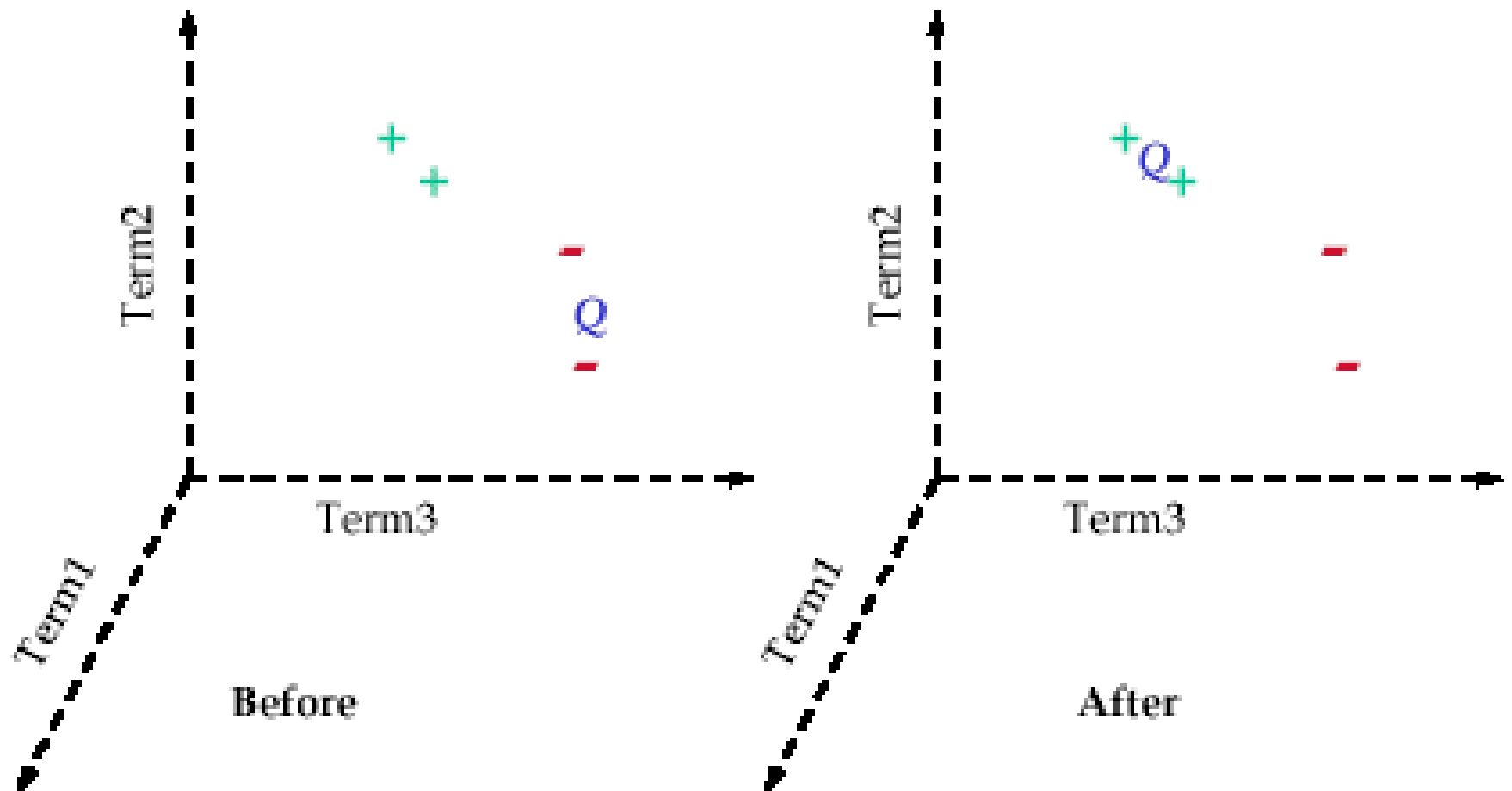
$\alpha = 0.50$ $\beta = 0.25$

$$\begin{aligned} q' &= q + 0.5D_1 - 0.25D_2 \\ &= (5,0,3,0,1) + 0.5(2,1,2,0,0) - 0.25(1,0,0,0,2) \\ &= (5.75, 0.50, 4.0, 0.0, 0.5) \end{aligned}$$

Relevance Feedback

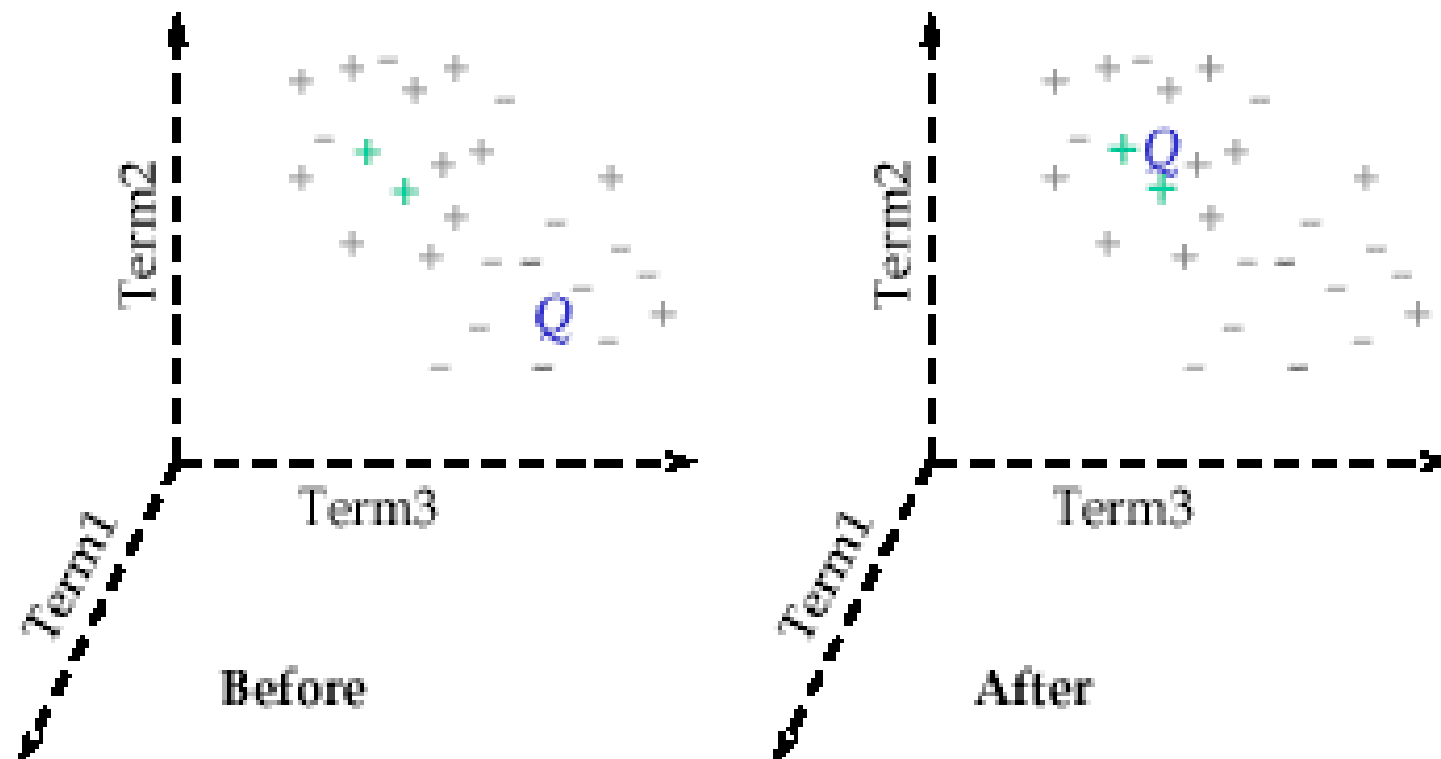


Relevance Feedback



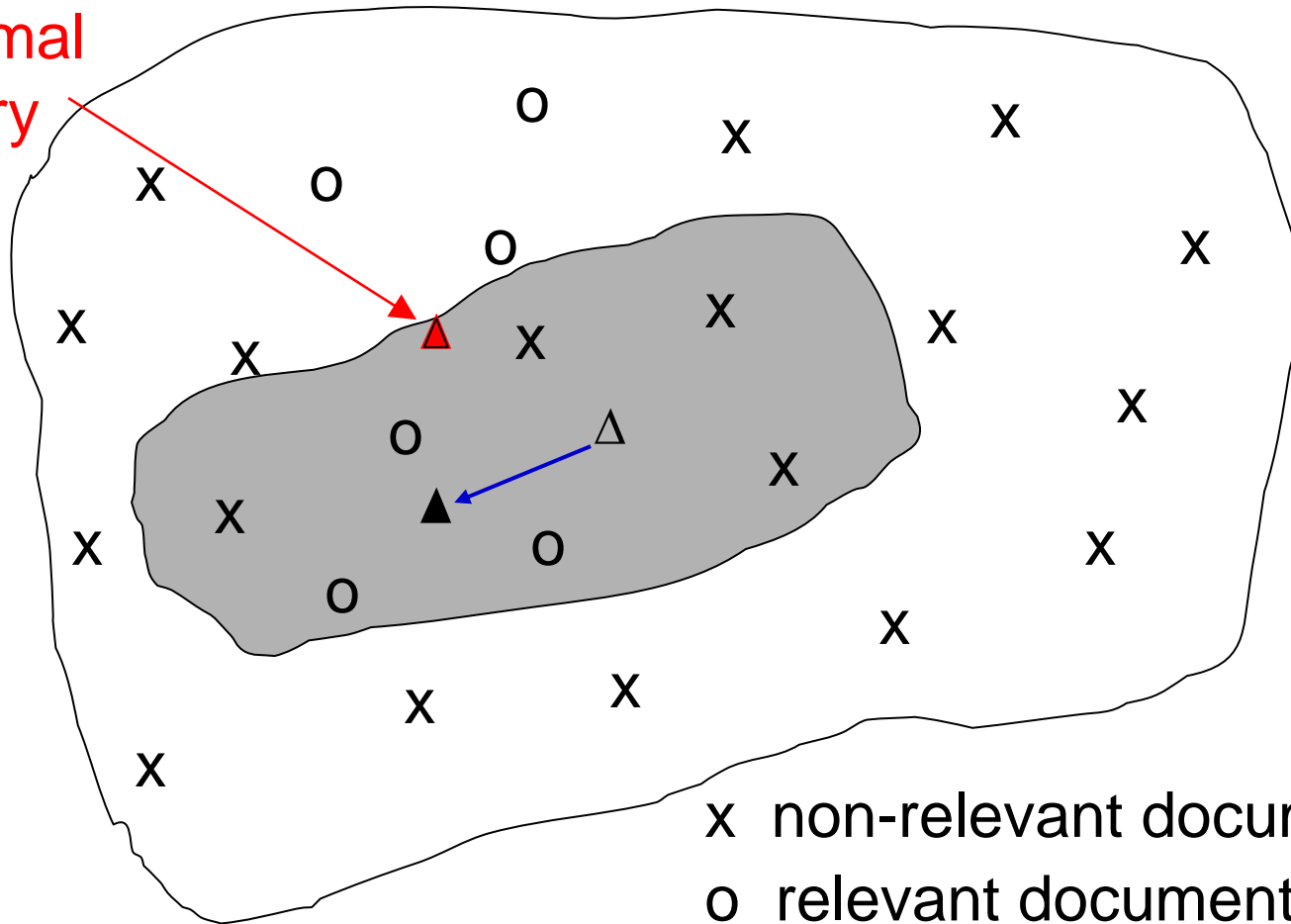
Relevance Feedback

How can relevance feedback save time if a person has to read documents?



Difficulties with Relevance Feedback

optimal
query



*Hits from
the initial
query are
contained
in the gray
shaded
area*

- x non-relevant documents
- o relevant documents
- △ original query
- ▲ reformulated query

Vector Space Re-Weighting

- The initial query vector \mathbf{q}_0 will have non-zero weights only for terms appearing in the query
- The query vector update process can add weight to terms that don't appear in the original query
- Some terms can **end up** having **negative** weight!
 - E.g., if you want to find information on the planet Saturn, “car” could have a negative weight...

Automatically (Implicit)

- **Automatic Global Analysis**
- **Automatic Local Analysis**

Automatic Global Analysis

- A thesaurus-like structure
- Short history
 - Until the beginning of the 1990s, global analysis was considered to be a technique which failed to yield consistent improvements in retrieval performance with **general collections**
 - This perception has changed with the appearance of modern procedures for **global analysis**

Query Expansion based on a Similarity Thesaurus

- **Idea by Qiu and Frei [1993]**
 - Similarity thesaurus is based on term to term relationships rather than on a matrix of co-occurrence
 - Terms for expansion are selected based on their similarity to the whole query rather than on their similarities to individual query terms
- **Definition**
 - N : total number of documents in the collection
 - t : total number of terms in the collection
 - $tf_{i,j}$: occurrence **frequency of term k_i** in the document d_j
 - t_j : the number of distinct index terms in the document d_j
 - itf_j : the inverse **term frequency** for document d_j

$$itf_j = \log \frac{t}{t_j}$$

Term weighting vs. Term concept space

	K_1	K_2	K_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
\vdots	\vdots	\vdots		\vdots
\vdots	\vdots	\vdots		\vdots
D_n	w_{1n}	w_{2n}	...	w_{tn}

	D_1	D_2	D_n
K_1	w_{11}	w_{12}	...	w_{1n}
K_2	w_{21}	w_{22}	...	w_{2n}
\vdots	\vdots	\vdots		\vdots
\vdots	\vdots	\vdots		\vdots
K_t	w_{t1}	w_{t2}	...	w_{tn}

$$w_{i,j} = \frac{(0.5 + 0.5 \frac{tf_{i,j}}{\max_k \{tf_{k,j}\}})idf_i}{\sqrt{\sum_{k=1}^t (0.5 + 0.5 \frac{tf_{k,j}}{\max_k \{tf_{k,j}\}})^2 idf_k^2}}$$

$$idf_i = \log \frac{N}{n_i}$$

$$w_{i,j} = \frac{(0.5 + 0.5 \frac{tf_{i,j}}{\max_k \{tf_{i,k}\}})itf_j}{\sqrt{\sum_{k=1}^N (0.5 + 0.5 \frac{tf_{i,k}}{\max_k \{tf_{i,k}\}})^2 itf_k^2}}$$

$$itf_j = \log \frac{t}{t_j}$$

Similarity Thesaurus

- Each term is associated with a vector

$$\vec{k}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,N})$$

- where $w_{i,j}$ is a weight associated to the index-document pair

$$w_{i,j} = \frac{(0.5 + 0.5 \frac{tf_{i,j}}{\max_k \{tf_{i,k}\}}) itf_j}{\sqrt{\sum_{k=1}^N (0.5 + 0.5 \frac{tf_{i,k}}{\max_k \{tf_{i,k}\}})^2 itf_k^2}}$$

- The *relationship between two terms k_u and k_v* is

$$c_{u,v} = \vec{k}_u \bullet \vec{k}_v = \sum_{j=1}^N w_{u,j} \times w_{v,j}$$

Query Expansion Procedure with Similarity Thesaurus

1. Represent the query in the concept space by using the representation of the index terms

$$\vec{q} = \sum_{k_u \in q} w_{u,q} \vec{k}_u$$

2. Compute the similarity $\text{sim}(q, k_v)$ between each term k_v and the whole query

$$\text{sim}(q, k_v) = \vec{q} \bullet \vec{k}_v = \left(\sum_{k_u \in q} w_{u,q} \vec{k}_u \right) \bullet \vec{k}_v = \sum_{k_u \in Q} w_{u,q} \times c_{u,v}$$

3. Expand the query with the top r ranked terms according to $\text{sim}(q, k_v)$

$$w_{v,q'} = \frac{\text{sim}(q, k_v)}{\sum_{k_u \in q} w_{u,q}}$$

Query Expansion based on a Similarity Thesaurus

- A document d_j is represented term-concept space by
$$\vec{d}_j = \sum_{k_v \in d_j} w_{v,j} \times \vec{k}_v$$
- If the original query q is expanded to include all the t index terms, then the similarity $\text{sim}(q, d_j)$ between the document d_j and the query q can be computed as

$$\text{sim}(\vec{q}, \vec{d}_j) = \left(\sum_{k_u \in q} w_{u,q} \times \vec{k}_u \right) \bullet \left(\sum_{k_v \in d_j} w_{v,j} \times \vec{k}_v \right)$$

$$\text{sim}(\vec{q}, \vec{d}_j) = \sum_{k_v \in d_j} \sum_{k_u \in q} w_{v,j} \times w_{u,q} \times c_{u,v}$$

- which is similar to the generalized vector space model

Automatic Global Analysis Example

$$\begin{pmatrix}
 K_1 & D_1 & D_2 & \dots & D_n \\
 W_{11} & W_{12} & \dots & W_{1n} \\
 K_2 & W_{21} & W_{22} & \dots & W_{2n} \\
 \vdots & \vdots & \vdots & & \vdots \\
 \vdots & \vdots & \vdots & & \vdots \\
 K_t & W_{t1} & W_{t2} & \dots & W_{tn}
 \end{pmatrix}$$

$$w_{i,j} = \frac{\left(0.5 + 0.5 \frac{f_{i,j}}{\max_j(f_{i,j})}\right) itf_j}{\sqrt{\sum_{l=1}^N \left(0.5 + 0.5 \frac{f_{i,l}}{\max_l(f_{i,l})}\right)^2 itf_l^2}}$$

$$itf_j = \log \frac{t}{t_j}$$

Automatic Global Analysis Example

The relationship between two terms

C	1	2	3	...	m
1	C1,1	C1,2	C1,3		C1,m
2	C2,1	C2,2	C2,3	...	C2,m
3	C3,1	C3,2	C3,3	...	C3,m
...					
n	Cn,1	Cn,2	Cn,3	...	Cn,m

$$c_{u,v} = \vec{k}_u \bullet \vec{k}_v = \sum_{j=1}^N w_{u,j} \times w_{v,j}$$

Ex.

$$C_{1,3} = w_{1,1} * w_{3,1} + w_{1,2} * w_{3,2} + w_{1,3} * w_{3,3} + \dots + w_{1,n} * w_{3,n}$$

Automatic Global Analysis Example

Original Query

$$q = w_{1,q}K_1 + w_{2,q}K_2 + w_{3,q}K_3 + \dots + w_{n,q}K_n$$

- compute a similarity $\text{sim}(q, kv)$ between each term kv correlated to the query terms and the whole query q

$$\text{sim}(q, k_v) = \vec{q} \cdot \vec{k}_v = \sum_{k_u \in q} w_{u,q} \times c_{u,v}$$

EX.

$$\text{sim}(q, k_3) = w_{1,q} * c_{1,3} + w_{2,q} * c_{2,3} + w_{3,q} * c_{3,3} + \dots + w_{n,q} * c_{n,3}$$

Automatic Global Analysis Example

Arrange $\text{sim}(q, k_t)$

Ex.

$$\text{sim}(q, k_1) = 0.53$$

$$\text{sim}(q, k_2) = 0.36$$

$$\text{sim}(q, k_3) = 3.98$$

$$\text{sim}(q, k_4) = 1.87$$

$$\text{sim}(q, k_3)$$

$$\text{sim}(q, k_4)$$

$$\text{sim}(q, k_1)$$

$$\text{sim}(q, k_2)$$



$$\text{sim}(q, k_2)$$

$$\text{sim}(q, k_4)$$

$$\text{sim}(q, k_3)$$

$$\text{sim}(q, k_1)$$



Original Query

$$q = K_1 + K_4$$

New Query

$$q = K_1 + K_3 + K_4$$

New Query

$$q = K_1 + K_2 + K_3 + K_4$$

Automatic Global Analysis Example

Compute new weight terms for query

Original Query

$$q = w_{1,q}K_1 + w_{2,q}K_2 + w_{3,q}K_3 + \dots + w_{n,q}K_n$$

$$w_{v,q'} = \frac{\text{sim}(q, k_v)}{\sum_{k_u \in q} w_{u,q}}$$

Ex.

$$w_{3,q'} = \frac{\text{sim}(q, k_3)}{(w_{1,q} + w_{2,q} + w_{3,q} + \dots + w_{n,q})}$$

$$w_{1,q'} = 2.6$$

$$w_{3,q'} = 5.4$$

$$w_{4,q'} = 4.8$$

New Query

$$q = 2.6K_1 + 5.4K_3 + 4.8K_4$$

Automatic Global Analysis Example

Compute $\text{sim}(q, d_j)$ for new relevance document

$$\text{sim}(q, d_j) \propto \sum_{k_v \in d_j} \sum_{k_u \in q} w_{i,j} \times w_{u,q} \times c_{u,v}$$

$$\begin{aligned} \text{sim}(q, d_2) = & w_{1,2} * w_{1,q} * c_{1,1} + w_{1,2} * w_{1,q} * c_{1,2} + w_{1,2} * w_{1,q} * c_{1,3} + \dots + w_{1,2} * w_{1,q} * c_{1,m} + \\ & w_{2,2} * w_{2,q} * c_{2,1} + w_{2,2} * w_{2,q} * c_{2,2} + w_{2,2} * w_{2,q} * c_{2,3} + \dots + w_{2,2} * w_{2,q} * c_{2,m} + \\ & w_{3,2} * w_{3,q} * c_{3,1} + w_{3,2} * w_{3,q} * c_{3,2} + w_{3,2} * w_{3,q} * c_{3,3} + \dots + w_{3,2} * w_{3,q} * c_{3,m} + \\ & \dots \dots \dots \\ & w_{n,2} * w_{n,q} * c_{n,1} + w_{n,2} * w_{n,q} * c_{n,2} + w_{n,2} * w_{n,q} * c_{n,3} + \dots + w_{n,2} * w_{n,q} * c_{n,m} \end{aligned}$$

$$\begin{aligned} \text{sim}(q, d_2) = & w_{1,2} * w_{1,q} * (c_{1,1} + c_{1,2} + c_{1,3} + \dots + c_{1,m}) + \\ & w_{2,2} * w_{2,q} * (c_{2,1} + c_{2,2} + c_{2,3} + \dots + c_{2,m}) + \\ & w_{3,2} * w_{3,q} * (c_{3,1} + c_{3,2} + c_{3,3} + \dots + c_{3,m}) + \\ & \dots \dots \dots \\ & w_{n,2} * w_{n,q} * (c_{n,1} + c_{n,2} + c_{n,3} + \dots + c_{n,m}) \end{aligned}$$

Automatic Global Analysis Example

Example

$D_1 = A, B, B, A, A, C$

$D_2 = D, D, C$

$D_3 = B, E, E$

$D_4 = D, E, A$

Query = 2.3A+ C

$$w_{i,j} = \frac{\left(0.5 + 0.5 \frac{f_{i,j}}{\max_j(f_{i,j})}\right) itf_j}{\sqrt{\sum_{l=1}^N \left(0.5 + 0.5 \frac{f_{i,l}}{\max_l(f_{i,l})}\right)^2 itf_l^2}}$$

$$itf_j = \log \frac{t}{t_j}$$

Automatic Global Analysis Example

Example

$D_1 = A, B, B, A, A, C$

$D_2 = D, D, C$

$D_3 = B, E, E$

$D_4 = A, D, E$

Query = 2.3A + C

Term = 5

$$itf_j = \log \frac{t}{t_j}$$

$$itf_4 = \log \frac{5}{3} = 0.222$$

Key/Doc	D1	D2	D3	D4
A	3	0	0	1
B	2	0	1	0
C	1	1	0	0
D	0	2	0	1
E	0	0	2	1
Max	3	2	2	1
t_j	3	2	2	3
itf(Doc)	0.222	0.398	0.398	0.222

Automatic Global Analysis Example

	D1	D2	D3	D4
A	3	0	0	1
B	2	0	1	0
C	1	1	0	0
D	0	2	0	1
E	0	0	2	1
Max	3	2	2	1
tj	3	2	2	3
itf	0.222	0.398	0.398	0.222

$$w_{i,j} = \frac{\left(0.5 + 0.5 \frac{f_{i,j}}{\max_j(f_{i,j})}\right) itf_j}{\sqrt{\sum_{l=1}^N \left(0.5 + 0.5 \frac{f_{i,j}}{\max_l(f_{i,l})}\right)^2 itf_l^2}}$$

$$w_{1,3} = \frac{\left(0.5 + 0.5 \frac{f_{1,3}}{\max(f_{d3})}\right) itf_3}{\sqrt{\left(0.5 + 0.5 \frac{f_{1,1}}{\max(f_{d1})}\right)^2 itf_1^2 + \left(0.5 + 0.5 \frac{f_{1,2}}{\max(f_{d2})}\right)^2 itf_2^2 + \left(0.5 + 0.5 \frac{f_{1,3}}{\max(f_{d3})}\right)^2 itf_3^2 + \left(0.5 + 0.5 \frac{f_{1,4}}{\max(f_{d4})}\right)^2 itf_4^2}}$$

$$(0.5 + 0.5 * \frac{0}{2}) 0.398$$

$$w_{1,3} = \frac{(0.5 + 0.5 * \frac{3}{3})^2 0.222^2 + (0.5 + 0.5 * \frac{0}{2})^2 0.398^2 + (0.5 + 0.5 * \frac{0}{2})^2 0.398^2 + (0.5 + 0.5 * \frac{1}{1})^2 0.222^2}{\sqrt{\quad}}$$

$$w_{1,3} = 1.509$$

Automatic Global Analysis Example

Term Weight

W	D ₁	D ₂	D ₃	D ₄
A	1.683	1.509	1.509	1.683
B	1.228	1.322	1.983	0.737
C	0.996	2.010	1.340	0.747
D	0.598	2.146	1.073	1.197
E	0.598	1.073	2.146	1.197



$$c_{u,v} = \vec{k}_u \bullet \vec{k}_v = \sum_{j=1}^N w_{u,j} \times w_{v,j}$$

$$C_{1,3} = w_{1,1} * w_{3,1} + w_{1,2} * w_{3,2} + w_{1,3} * w_{3,3} + w_{1,4} * w_{3,4}$$

$$= 1.683 * 0.996 + 1.509 * 2.010 + 1.509 * 1.340 + 1.683 * 0.747$$

$$= 7.987$$

Automatic Global Analysis Example

The relationship between two terms

C	A	B	C	D	E
A	10.218	8.293	7.987	7.879	7.879
B	8.293	7.728	7.085	6.581	7.290
C	7.987	7.085	7.383	7.241	6.522
D	7.879	6.581	7.241	7.548	6.397
E	7.879	7.290	6.522	6.397	7.548

Automatic Global Analysis Example

term similarity

C	A	B	C	D	E	Sim(q,K _i)	
A	10.218	8.293	7.987	7.879	7.879	31.487	←
B	8.293	7.728	7.085	6.581	7.290	26.159	←
C	7.987	7.085	7.383	7.241	6.522	25.753	←
D	7.879	6.581	7.241	7.548	6.397	25.362	
E	7.879	7.290	6.522	6.397	7.548	24.643	
q	2.3	0	1	0	0		

ADD K₂ to Query

$$\text{sim}(q, k_v) = \vec{q} \cdot \vec{k}_v = \sum_{k_u \in q} w_{u,q} \times c_{u,v}$$

$$\begin{aligned} \text{sim}(q, k_3) &= w_{1,q} * c_{1,3} + w_{2,q} * c_{2,3} + w_{3,q} * c_{3,3} + w_{4,q} * c_{4,3} + w_{5,q} * c_{5,3} \\ &= 2.3 * 7.987 + 1 * 7.383 = 25.753 \end{aligned}$$

Automatic Global Analysis Example

Recompute term similarity

C	A	B	C	D	E	Sim(q,K _i)
A	10.218	8.293	7.987	7.879	7.879	39.780
B	8.293	7.728	7.085	6.581	7.290	33.887
C	7.987	7.085	7.383	7.241	6.522	32.838
D	7.879	6.581	7.241	7.548	6.397	31.942
E	7.879	7.290	6.522	6.397	7.548	31.933
q	2.3	1	1	0	0	

$$\begin{aligned}\text{sim}(q, k_3) &= w_{1,q} * C_{1,3} + w_{2,q} * C_{2,3} + w_{3,q} * C_{3,3} + w_{4,q} * C_{4,3} + w_{5,q} * C_{5,3} \\ &= 2.3 * 7.987 + 1 * 7.085 + 1 * 7.383 = 32.838\end{aligned}$$

Automatic Global Analysis Example

Compute new weight terms for query

Original Query

$q = 2.3K_1 + K_2 + K_3$ Sum query weight = $2.3 + 1 + 1 = 4.3$

$$w_{v,q'} = \frac{\text{sim}(q, k_v)}{\sum_{k_u \in q} w_{u,q}}$$

$$w_{1,q'} = 39.780 / 4.3 = 9.251$$

$$w_{2,q'} = 33.887 / 4.3 = 7.881$$

$$w_{3,q'} = 32.838 / 4.3 = 7.637$$

	A	B	C	D	E
q'	9.251	7.881	7.637	-	-

Automatic Global Analysis Example

Arrange Relevance

$$q' = 9.251A + 7.881B + 7.637C$$

W	D1	D2	D3	D4
A	1.683	1.509	1.509	1.683
B	1.228	1.322	1.983	0.737
C	0.996	2.010	1.340	0.747
D	0.598	2.146	1.073	1.197
E	0.598	1.073	2.146	1.197

C	A	B	C	D	E
A	10.22	8.293	7.987	7.879	7.879
B	8.293	7.728	7.085	6.581	7.290
C	7.987	7.085	7.383	7.241	6.522
D	7.879	6.581	7.241	7.548	6.397
E	7.879	7.290	6.522	6.397	7.548

$$\text{sim}(q, d_j) \propto \sum_{k_v \in d_j} \sum_{k_u \in q} w_{i,j} \times w_{u,q} \times c_{u,v}$$

$$\begin{aligned} w_{1,2} &= 1.509 & w_{1,q} &= 9.251 \\ w_{2,2} &= 1.322 & w_{2,q} &= 7.881 \\ w_{3,2} &= 2.010 & w_{3,q} &= 7.637 \\ w_{4,2} &= 2.146 & w_{4,q} &= 0 \\ w_{5,2} &= 1.073 & w_{5,q} &= 0 \end{aligned}$$

$$\begin{aligned} \text{sim}(q, d_2) &= \\ &w_{1,2} * w_{1,q} * (c_{1,1} + c_{1,2} + c_{1,3} + c_{1,4} + c_{1,5}) + \\ &w_{2,2} * w_{2,q} * (c_{2,1} + c_{2,2} + c_{2,3} + c_{2,4} + c_{2,5}) + \\ &w_{3,2} * w_{3,q} * (c_{3,1} + c_{3,2} + c_{3,3} + c_{3,4} + c_{3,5}) + \\ &w_{4,2} * w_{4,q} * (c_{4,1} + c_{4,2} + c_{4,3} + c_{4,4} + c_{4,5}) + \\ &w_{5,2} * w_{5,q} * (c_{5,1} + c_{5,2} + c_{5,3} + c_{5,4} + c_{5,5}) \end{aligned}$$

$$\text{sim}(q, d_2) = 1531.123$$

Automatic Global Analysis Example

Arrange Relevance

$$q' = 9.251A + 7.881B + 7.637C$$

	D_1	D_2	D_3	D_4
$\text{Sim}(q, d_j)$	1,291.282	1,531.123	1,538.429	1,079.324

Answer = D_3, D_2, D_1, D_4

Automatically (Implicit)

- **Automatic Global Analysis**
- **Automatic Local Analysis**

Automatic Local analysis

- Basic concept
 - Expanding the query with terms correlated to the query terms
 - The correlated terms are presented in the local clusters built from *the local document set*

Automatic Local Analysis

- Definition
 - local document set D_l : the set of *documents retrieved* by a query
 - local vocabulary V_l : the set of *all distinct words* in D_l
 - stemmed vocabulary S_l : the set of *all distinct stems* derived from V_l
- Building local clusters
 - association clusters
 - metric clusters
 - scalar clusters

Association Clusters

- idea
 - Based on the co-occurrence of stems (or terms) **inside documents**
- association matrix
 - $f_{s_i, j}$: the frequency of a **stem** s_i in a document $d_j (\in D_I)$
 - $m = (f_{s_i, j})$: an association matrix with $|S_I|$ rows and $|D_I|$ columns
 - $s = mm$: a local **stem-stem** association matrix

Association Clusters

- Idea

- co-occurrence of **stems** (or terms) inside documents (**frequency of stems in doc**)

$$c(k_u, k_v) = \sum_{j=1}^{|D|} f_{u,j} \times f_{v,j}$$

- $f_{u,j}$: the frequency of a **stem** k_u in a document d_j
- local association cluster for a stem k_u
 - the set of k largest values $c(k_u, k_v)$
- given a query q , find clusters for the $|q|$ query terms

- normalized form $s(k_u, k_v) = \frac{c(k_u, k_v)}{c(k_u, k_u) + c(k_v, k_v) - c(k_u, k_v)}$

Metric Clusters

- Idea
 - consider the **distance between two terms** in the same cluster
- Definition
 - $V(k_u)$: the set of keywords which have the same stem form as k_u
 - distance $r(k_i, k_j)$ =the number of words between term k_i and k_j
 - normalized form

$$c(k_u, k_v) = \sum_{i \in V(k_u)} \sum_{j \in V(k_v)} \frac{1}{r(k_i, k_j)}$$

$$s(k_u, k_v) = \frac{c(k_u, k_v)}{|V(k_u)| \times |V(k_v)|}$$

Scalar Clusters

- Idea
 - two stems with similar neighborhoods have some **synonymity** relationships
- *Definition*
 - $c_{u,v} = c(k_u, k_v)$
 - vectors of correlation values for stem k_u and k_v
$$\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t}) \qquad \vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$$
 - scalar association matrix
$$S_{u,v} = \frac{\vec{s}_u \bullet \vec{s}_v}{|\vec{s}_u| \times |\vec{s}_v|}$$
 - scalar clusters
 - the set of k largest values of scalar association

Association Clusters

- Idea

- **co-occurrence** of stems (or terms) inside documents (**frequency of stems in doc**)

$$c(k_u, k_v) = \sum_{j=1}^{|D|} f_{u,j} \times f_{v,j}$$

- $f_{u,j}$: the frequency of a stem k_u in a document d_j
- local association cluster for a stem k_u
 - the set of k largest values $c(k_u, k_v)$
- given a query q , find clusters for the $|q|$ query terms

- normalized form $s(k_u, k_v) = \frac{c(k_u, k_v)}{c(k_u, k_u) + c(k_v, k_v) - c(k_u, k_v)}$

Association Clusters

$$c_{u,v} = \sum_{dj \in Dl} f_{su,j} \times f_{sv,j} : \text{a correlation between the stems } s_u \text{ and } s_v$$

an element in \overrightarrow{mm}^t

$$S_{u,v} = C_{u,v} : \text{ *unnormalized matrix* }$$

$$s_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}} : \text{ *normalized matrix* }$$

$s_u(n)$: local association cluster around the stem s_u

$\left\{ \begin{array}{l} \text{Take } u\text{-th row} \\ \text{Return the set of } n \text{ **largest values** } s_{u,v} \text{ (} u \neq v \text{)} \end{array} \right.$

Association Clusters Example

$q = A+B$

$\{B,D,C\} \rightarrow A$

$d_1 = A,A,B,D$

$d_2 = B,A,C,C,D$

$d_3 = A,B$

$d_4 = B,C,D$

$d_5 = D$

$d_6 = A,B,D$

$d_7 = B,B,A$

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
A	2	1	1	0	0	1	1
B	1	1	1	1	0	1	2
C	0	2	0	1	0	0	0
D	1	1	0	1	1	1	0



$$C_{u,v} = \sum_{dj \in D_I} f_{s_u,j} \times f_{s_v,j}$$

$$\begin{aligned}
 C_{1,4} &= (f_{1,1} * f_{4,1}) + (f_{1,2} * f_{4,2}) + (f_{1,3} * f_{4,3}) + (f_{1,4} * f_{4,4}) + (f_{1,5} * f_{4,5}) + (f_{1,6} * f_{4,6}) + (f_{1,7} * f_{4,7}) \\
 &= 2*1 + 1*1 + 1*0 + 0*1 + 0*1 + 1*1 + 1*0 \\
 &= 4
 \end{aligned}$$

Association Clusters Example

Correlation Matrix (C)

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	8	7	2	4
<i>B</i>	7	9	3	4
<i>C</i>	2	3	5	3
<i>D</i>	4	4	3	5

Association Clusters Example

Other way to compute the Correlation Matrix

$$C = m m^t$$

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
A	2	1	1	0	0	1	1
B	1	1	1	1	0	1	2
C	0	2	0	1	0	0	0
D	1	1	0	1	1	1	0

m

	A	B	C	D
d_1	2	1	0	1
d_2	1	1	2	1
d_3	1	1	0	0
d_4	0	1	1	1
d_5	0	0	0	1
d_6	1	1	0	1
d_7	1	2	0	0

m^t

$$\begin{aligned}
 C_{1,4} &= (m_{1,1} * m_{1,4}^t) + (m_{1,2} * m_{2,4}^t) + (m_{1,3} * m_{3,4}^t) + (m_{1,4} * m_{4,4}^t) + (m_{1,5} * m_{5,4}^t) + \\
 &\quad (m_{1,6} * m_{6,4}^t) + (m_{1,7} * m_{7,4}^t) \\
 &= 2*1 + 1*1 + 1*0 + 0*1 + 0*1 + 1*1 + 1*0 \\
 &= 4
 \end{aligned}$$

Association Clusters Example

Correlation Matrix (C)

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	8	7	2	4
<i>B</i>	7	9	3	4
<i>C</i>	2	3	5	3
<i>D</i>	4	4	3	5

Association Clusters Example

Normalized Correlation Matrix (S)

$$s_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}}$$

$$s_{1,2} = \frac{c_{1,2}}{c_{1,1} + c_{2,2} - c_{1,2}} = \frac{7}{8+9-7} = 0.70$$

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	8	7	2	4
<i>B</i>	7	9	3	4
<i>C</i>	2	3	5	3
<i>D</i>	4	4	3	5

Association Clusters Example

Normalized Correlation Matrix

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	1	0.70	0.18	0.44
<i>B</i>	0.70	1	0.27	0.40
<i>C</i>	0.18	0.27	1	0.43
<i>D</i>	0.44	0.40	0.43	1

Take u-th row

Return the set of n **largest values** $s_{u,v}$ ($u \neq v$)

Term Relation

1. $\{A, B\}$
2. $\{B, A\}$
3. $\{C, D\}$
4. $\{D, A\}$

Original Query

$$q = A + B$$

New Query

$$\begin{aligned} q' &= (A + 0.7B) + (0.7A + B) \\ &= 1.7A + 1.7B \\ &= A + B \end{aligned}$$

Association Clusters Example (other case)

Normalized Correlation Matrix

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>A</i>	1	0.70	0.18	0.44
<i>B</i>	0.70	1	0.85	0.63
<i>C</i>	0.18	0.85	1	0.63
<i>D</i>	0.44	0.63	0.63	1

Term Relation

1. $\{A,B\}$
2. $\{B,C\}$
3. $\{C,B\}$
4. $\{D,B,C\}$

New Query

Original Query

$$q = A+B$$



$$\begin{aligned} q' &= (A+0.7B)+(B+0.85C) \\ &= A+1.7B+0.85C \end{aligned}$$

New Query

Original Query

$$q = C+2D$$



$$\begin{aligned} q' &= (0.85B+C)+2*(0.63B+0.63C+D) \\ &= 2.11B+2.26C+2D \end{aligned}$$

Metric Clusters

- Idea
 - consider the **distance between two terms** in the same cluster
- Definition
 - $V(k_u)$: the set of keywords which have the same stem form as k_u
 - distance $r(k_i, k_j)$ = the number of words between term k_i and k_j
 - normalized form

$$c(k_u, k_v) = \sum_{i \in V(k_u)} \sum_{j \in V(k_v)} \frac{1}{r(k_i, k_j)}$$

$$s(k_u, k_v) = \frac{c(k_u, k_v)}{|V(k_u)| \times |V(k_v)|}$$

Metric Clusters

$s_{u,v} = c_{u,v}$: unnormalized matrix

$s_{u,v} = \frac{c_{u,v}}{|V(s_u)| \times |V(s_v)|}$: normalized matrix

$s_u(n)$: local metric cluster around the stem s_u

$\left\{ \begin{array}{l} \text{Take } u\text{-th row} \\ \text{Return the set of } n \text{ **largest values** } s_{u,v} \text{ (} u \neq v \text{)} \end{array} \right.$

Metric Clusters Example

$$q = A + 2D$$

$$k_n = A, B, C, D, E, F$$

A, B, C base on S_1 stem

D, E base on S_2 stem

F base on S_3 stem

Then

$$V(S_1) = \{A, B, C\}$$

$$V(S_2) = \{D, E\}$$

$$V(S_3) = \{F\}$$

ระยะ ห่าง	A	B	C	D	E	F
A	0	5	∞	∞	1	2
B	5	0	3	2	1	1
C	∞	3	0	3	4	∞
D	∞	2	3	0	∞	5
E	1	1	4	∞	0	1
F	2	1	∞	5	1	0

Metric Clusters Example

ระยะ ห่าง	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	0	5	∞	∞	1	2
<i>B</i>	5	0	3	2	1	1
<i>C</i>	∞	3	0	3	4	∞
<i>D</i>	∞	2	3	0	∞	5
<i>E</i>	1	1	4	∞	0	1
<i>F</i>	2	1	∞	5	1	0



	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	-	0.20	0	0	1	0.50
<i>B</i>	0.20	-	0.33	0.50	1	1
<i>C</i>	0	0.33	-	0.33	0.25	0
<i>D</i>	0	0.50	0.33	-	0	0.20
<i>E</i>	1	1	0.25	0	-	1
<i>F</i>	0.50	1	0	0.20	1	-

Metric Clusters Example

$$V(S_1) = \{A, B, C\}$$

$$V(S_2) = \{D, E\}$$

$$V(S_3) = \{F\}$$

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	-	0.20	0	0	1	0.50
<i>B</i>	0.20	-	0.33	0.50	1	1
<i>C</i>	0	0.33	-	0.33	0.25	0
<i>D</i>	0	0.50	0.33	-	0	0.20
<i>E</i>	1	1	0.25	0	-	1
<i>F</i>	0.50	1	0	0.20	1	-

$$c_{u,v} = \sum_{ki \in V(su)} \sum_{kj \in V(sv)} \frac{1}{r(k_i, k_j)}$$

$$\begin{aligned}
 c_{1,2} &= c(A, D) + c(A, E) + c(B, D) + c(B, E) + c(C, D) + c(C, E) \\
 &= 0 + 1 + 0.50 + 1 + 0.33 + 0.25 \\
 &= 3.08
 \end{aligned}$$

Metric Clusters Example

Correlation Matrix (C)

	S_1	S_2	S_3
S_1	0	3.08	1.50
S_2	3.08	0	1.20
S_3	1.50	1.20	0

Metric Clusters Example

Normalized Correlation Matrix (S)

	S_1	S_2	S_3
S_1	0	3.08	1.50
S_2	3.08	0	1.20
S_3	1.50	1.20	0

$V(S_1) = \{A, B, C\} = 3$

$V(S_2) = \{D, E\} = 2$

$V(S_3) = \{F\} = 1$

$$s_{u,v} = \frac{c_{u,v}}{|V(s_u)| \times |V(s_v)|}$$

$$s_{2,3} = \frac{c_{2,3}}{|V(s_2)| \times |V(s_3)|} = \frac{1.2}{2 \times 1} = 0.6$$

Metric Clusters Example

Normalized Correlation Matrix (S)

	S_1	S_2	S_3
S_1	0	0.51	0.50
S_2	0.51	0	0.60
S_3	0.50	0.60	0

Stem Relation

1. $\{S_1, S_2\}$
2. $\{S_2, S_3\}$
3. $\{S_3, S_2\}$

Original Query

$$q = A + 2D$$

New Query

$$\begin{aligned} q' &= (S_1 + 0.51S_2) + 2*(S_2 + 0.60S_3) \\ &= S_1 + 2.51S_2 + 1.2S_3 \end{aligned}$$

Scalar Clusters

- Idea
 - two stems with similar neighborhoods have some **synonymity relationships**

- *Definition*

- $c_{u,v} = c(k_u, k_v)$
- vectors of correlation values for stem k_u and k_v

Database , Math , Set
Tree, Water , Fertilizer
Flower, Letter , Lover

$$\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t}) \quad \vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$$

- scalar association matrix

$$S_{u,v} = \frac{\vec{s}_u \bullet \vec{s}_v}{|\vec{s}_u| \times |\vec{s}_v|}$$

- scalar clusters
 - the set of k **largest values** of scalar association

Scalar Clusters

$$\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t})$$

$$\vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$$

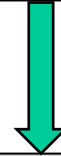
$$\vec{s}_1 = (c_{1,1}, c_{1,2}, \dots, c_{1,t})$$

$$\vec{s}_3 = (c_{3,1}, c_{3,2}, \dots, c_{3,t})$$

$$\vec{s}_1 = (c_{1,1}, c_{1,2}, c_{1,3}) = (5, 6, 1) \rightarrow C_{\text{Database, Algebra}}, C_{\text{Database, Math}}, C_{\text{Database, Set}}$$

$$\vec{s}_2 = (c_{2,1}, c_{2,2}, c_{2,3}) = (6, 9, 0) \rightarrow C_{\text{AI, Algebra}}, C_{\text{AI, Math}}, C_{\text{AI, Set}}$$

$$\vec{s}_3 = (c_{3,1}, c_{3,2}, c_{3,3}) = (1, 0, 2) \rightarrow C_{\text{Network, Algebra}}, C_{\text{Network, Math}}, C_{\text{Network, Set}}$$



Network = {Set(2), Algebra (1), Math(0)} ***idea

Scalar Clusters

$$\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t})$$

$$\vec{s}_1 = (c_{1,1}, c_{1,2}, \dots, c_{1,t})$$

$$\vec{s}_1 = (c_{1,1}, c_{1,2}, c_{1,3}) = (5, 6, 1)$$

$$\vec{s}_2 = (c_{2,1}, c_{2,2}, c_{2,3}) = (6, 9, 0)$$

$$\vec{s}_3 = (c_{3,1}, c_{3,2}, c_{3,3}) = (1, 0, 2)$$

$$|S_1| = \sqrt{25 + 36 + 1} = 7.874$$

$$|S_2| = \sqrt{36 + 81 + 0} = 10.817$$

$$|S_3| = \sqrt{1 + 0 + 4} = 2.236$$

$$\vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$$

$$\vec{s}_3 = (c_{3,1}, c_{3,2}, \dots, c_{3,t})$$

$$S_{u,v} = \frac{\vec{s}_u \bullet \vec{s}_v}{|\vec{s}_u| \times |\vec{s}_v|}$$

$$S_{1,3} = \frac{\vec{s}_1 \bullet \vec{s}_3}{|\vec{s}_1| \times |\vec{s}_3|}$$

$$S_{1,3} = \frac{7}{7.874 \times 2.236} = 0.398$$

Scalar Clusters

Normalized Correlation Matrix (S)

<i>S</i>	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃
<i>S</i> ₁	1	0.986	0.398
<i>S</i> ₂	0.986	1	0.248
<i>S</i> ₃	0.398	0.248	1

Stem Relation

1. $\{S_1, S_2\}$
2. $\{S_2, S_1\}$
3. $\{S_3, S_1\}$

Original Query

$$q = 3S_1 + S_3$$



Database

Network

New Query

$$\begin{aligned} q' &= 3*(S_1 + 0.986S_2) + (0.398S_1 + S_3) \\ &= 3.398S_1 + 2.958S_2 + S_3 \end{aligned}$$

Discussion

- Query expansion
 - useful
 - little explored technique
- Trends and research issues
 - The combination of local analysis, global analysis, visual displays, and interactive interfaces is also a current and important research problem