

# Chapter 5

User can ដ្ឋានអំពីលក្ខណៈ  
Query Operations ទារម្រោគនាក់

## Query Operations

វគ្គ ឱ្យ

- ដ្ឋានអំពីលក្ខណៈ user
- ដ្ឋានអំពីលក្ខណៈកម្មសាធារណ៍

# Motivation - Feast or famine

ព័ត៌មានលើកទាំងអស់ មានចំណាំក្នុងការបង្ហាញក្នុងប្រព័ន្ធដែលមានការបង្ហាញ

- 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  
keyword กໍາລຳ ↗  
in apple
- Many queries are ambiguous
  - “Saturn”
    - Saturn the planet?
    - Saturn the car?

# Relevance Feedback

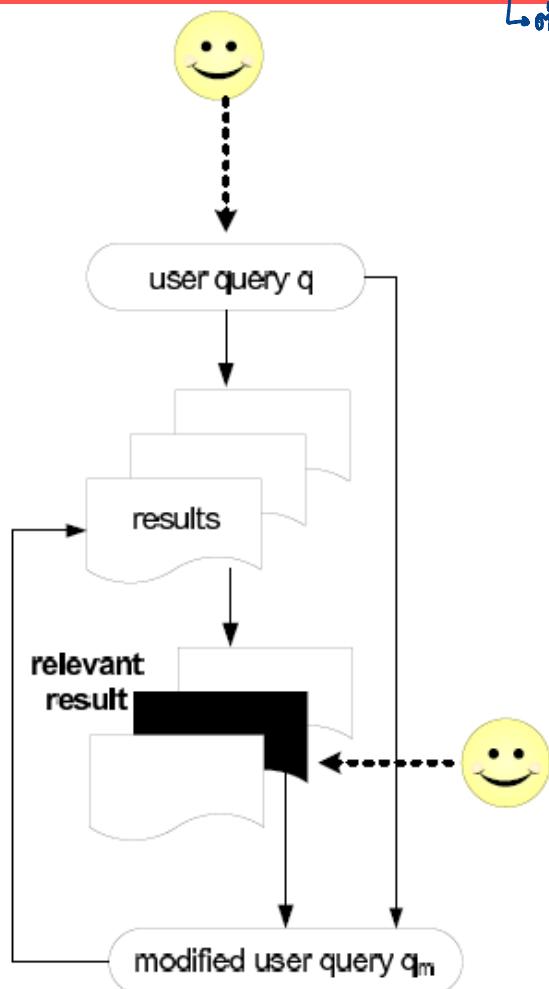
ការវិភាគ ដោយស្នើសុំជាមួយ

- Two general approaches:
  - Create new queries <sup>តាមរយៈ</sup> **with user feedback**  
**(explicit feedback)** <sup>ដ័ូរដោះសារពី user</sup>
  - Create new queries **automatically**  
**(implicit feedback)** <sup>ដើម្បីការងារបន្ថែម (អតិថិជន និងការកំណត់លក្ខណៈ)</sup>
- Re-compute document weights with new information  
<sup>ពិនិត្យ / គិត / ទិន្នន័យ keyword</sup>
- Expand or modify the query to more accurately reflect the user's desires

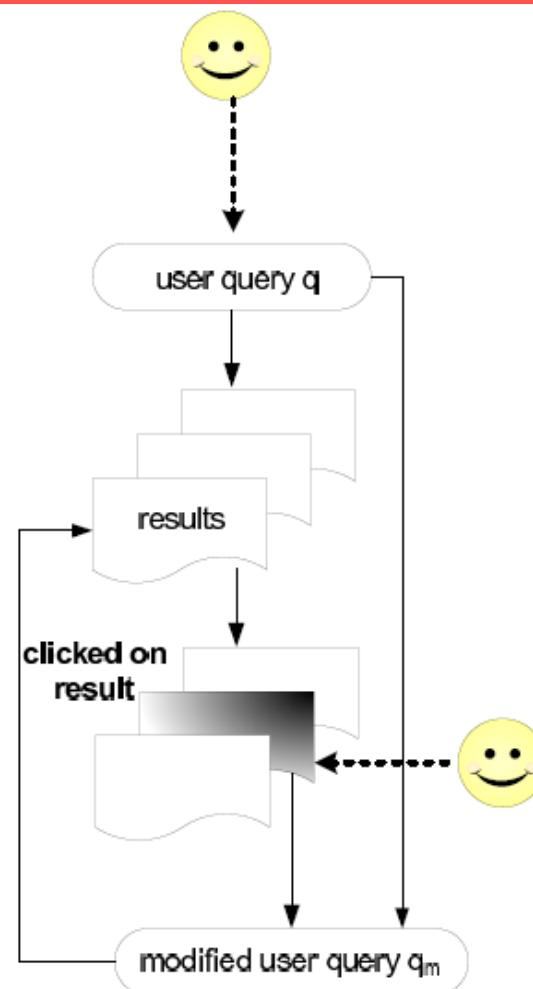
# Explicit Feedback

User Feedback

↳ explicit



(a) relevance feedback



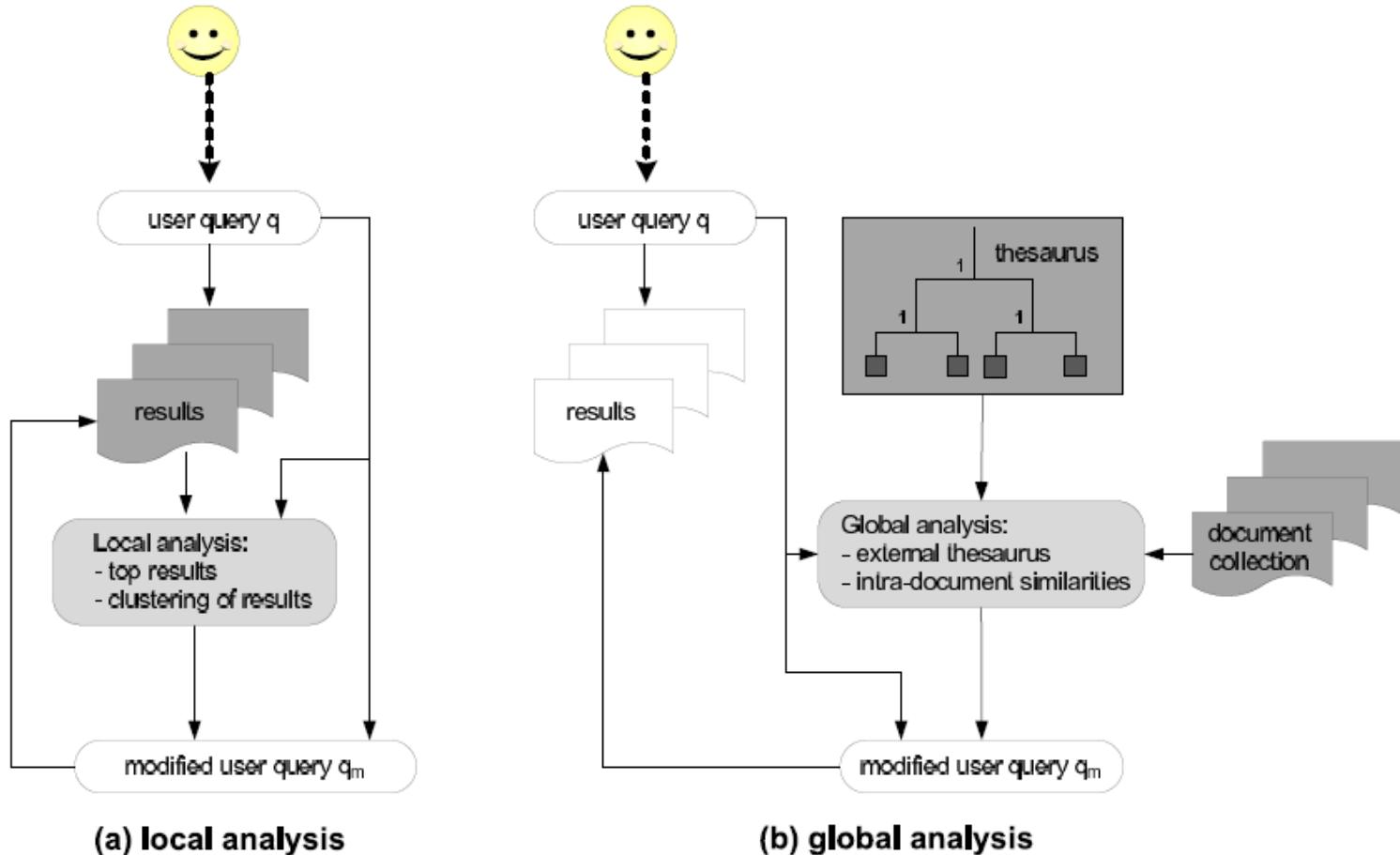
(b) click feedback

ກົດຕົວລາຍລະອຽດ  
(non-keyword)  $\rightarrow$  global

ກົດຕົວລາຍລະອຽດ  
(non-keyword)  $\rightarrow$  Local (minimum / nonreturn)

# Implicit Feedback

keyword ຂໍ້ມູນກົດຕົວລາຍລະອຽດ



# User Feedback

---

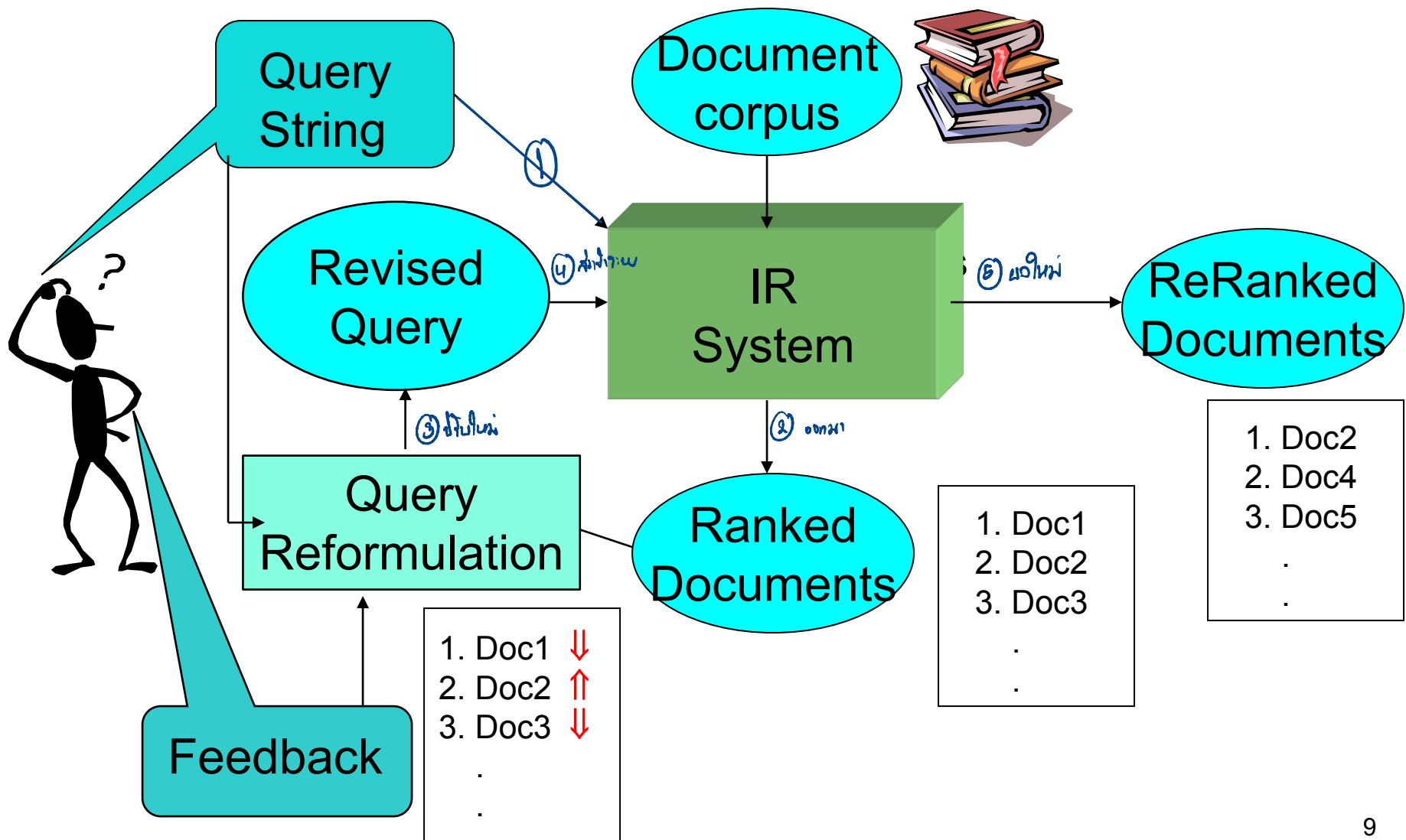
- After initial retrieval results are presented, allow the user to provide feedback on the relevance of one or more of the retrieved documents.
- Use this feedback information to reformulate the query.
- Produce new results based on reformulated query.
- Allows more interactive, ***multi-pass process***.

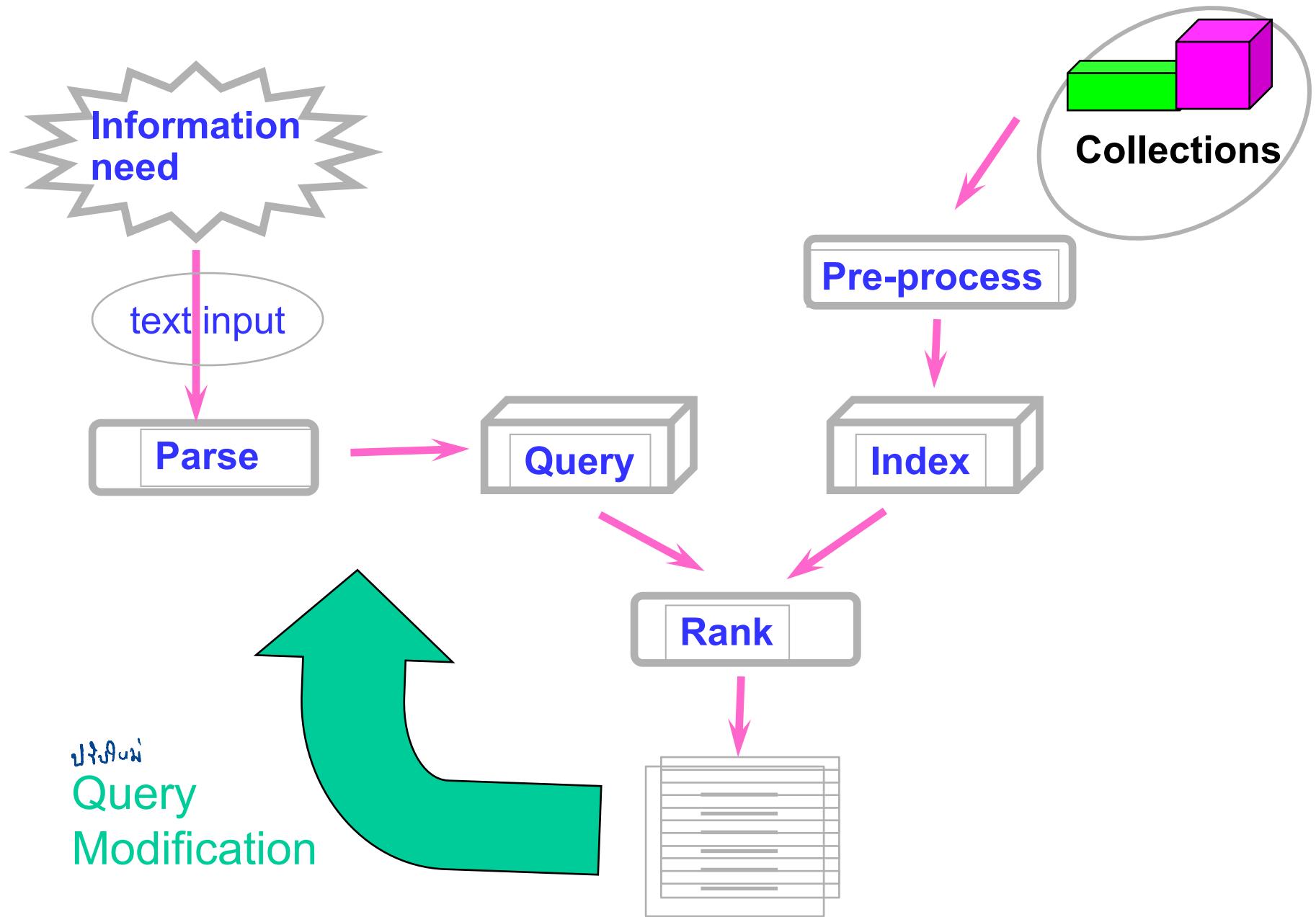
# User Feedback

---

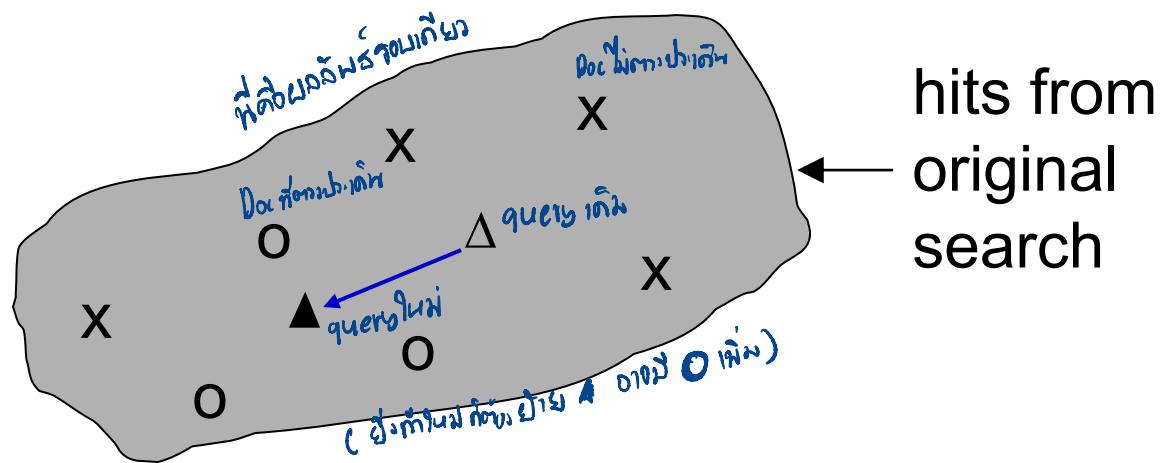
- ❑ The main idea consists of
  - selecting important terms from the documents that have been identified as relevant, and
  - enhancing the importance of these terms in a new query formulation

# User Feedback Architecture





# User Feedback (concept)



✗ documents identified as non-relevant

○ documents identified as relevant

▲ original query

reformulated query

# User Feedback (concept)

eResponder

DB list

New question

How do I license Mapuccino?

Go

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. Im 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 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

Refine Generate Draft

# User Feedback (concept)

The screenshot shows a Mozilla Firefox browser window with a red horizontal bar at the top. The title bar reads "regan - Google Search - Mozilla Firefox". The address bar shows the URL "http://www.google.com/search?q=regan&btnG=Search&hs=UnN&hl=en&lr=&client=firefox-a". The search term "regan" is entered in the search bar. The user's email "cmanning@gmail.com" and account links are visible in the top right. The search results are for "Web" and show 1 - 10 of about 13,200,000 results for "regan" (0.07 seconds). The results include:

- Brian Regan: The Official Site**  
Brian Regan is one of the best comedians performing today. His comedy, big enough for everyone, sharp enough for you, keeps audiences coming back time and ...  
[www.brianregan.com/](http://www.brianregan.com/) - 13k - Cached - Similar pages
- reganmusic.com**  
Color.  
[www.reganmusic.com/](http://www.reganmusic.com/) - 2k - Cached - Similar pages
- Regan Nursery Bare Root Roses**  
We offer over 1100 bareroot roses from one of the largest selections of Grade 1 bareroot roses in the US, including David Austin roses, Hybrid Tea roses, ...  
[www.regannursery.com/](http://www.regannursery.com/) - 14k - Cached - Similar pages

See results for: [ronald reagan](#)

- Biography of Ronald Reagan**  
Biography of Ronald Reagan, the fortieth President of the United States (1981-1989).  
[www.whitehouse.gov/history/presidents/rr40.html](http://www.whitehouse.gov/history/presidents/rr40.html)
- Ronald Reagan - Wikipedia, the free encyclopedia**  
Ronald Reagan visiting Nancy Reagan on the set of her movie Donovan's Brain, 1953.  
... Ronald Reagan on the cover of TIME as "Man of the Year," 1980 ...  
[en.wikipedia.org/wiki/Ronald\\_Reagan](http://en.wikipedia.org/wiki/Ronald_Reagan)
- RonaldReagan.com**  
Provides in-depth biographical information, message boards, video clips, and transcripts of historic speeches.  
[www.ronaldreagan.com/](http://www.ronaldreagan.com/)

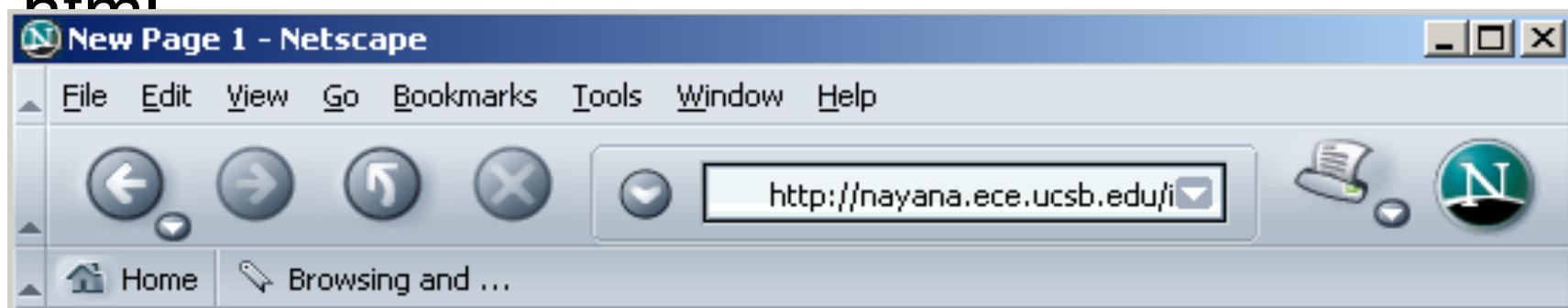
**Regan Family Genealogy Forum**  
Margaret Ann Regan 1878 Providence RI godmother - Barbara Glassel 3/05/06. James Regan Clarendon School Canton OH 1930s - Gregory Winters 3/05/06 ...  
[genforum.genealogy.com/regan/](http://genforum.genealogy.com/regan/) - 28k - Cached - Similar pages

Find: heinz   Find Next   Find Previous   Highlight all   Match case  
Done

# User Feedback: Example

- Image search engine

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

Designed by [Baris Sumengen](#) and [Shawn Newsam](#)

*Powered by JLAAMP2000 (Java, Linux, Apache, Mysql, Perl, Windows2000)*

# Results for Initial Query

Browse Search Prev Next Random



(144473, 16458)

0.0

0.0

0.0



(144457, 252140)

0.0

0.0

0.0



(144456, 262857)

0.0

0.0

0.0



(144456, 262863)

0.0

0.0

0.0



(144457, 252134)

0.0

0.0

0.0



(144483, 265154)

0.0

0.0

0.0



(144483, 264644)

0.0

0.0

0.0



(144483, 265153)

0.0

0.0

0.0



(144518, 257752)

0.0

0.0

0.0



(144538, 525937)

0.0

0.0

0.0



(144456, 249611)

0.0

0.0

0.0



(144456, 250064)

0.0

0.0

0.0

# User Feedback

Browse Search Prev Next Random



(144473, 16458)

0.0

0.0

0.0



(144457, 252140)

0.0

0.0

0.0



(144456, 262857)

0.0

0.0

0.0



(144456, 262863)

0.0

0.0

0.0



(144457, 252134)

0.0

0.0

0.0



(144483, 265154)

0.0

0.0

0.0



(144483, 264644)

0.0

0.0

0.0



(144483, 265153)

0.0

0.0

0.0



(144518, 257752)

0.0

0.0

0.0



(144538, 525937)

0.0

0.0

0.0



(144456, 249611)

0.0

0.0

0.0



(144456, 250064)

0.0

0.0

0.0

# Results after User Feedback

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

(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: នវត្ថុលេខ

(រៀងរាល់ key )  
ដូចជាមួយកំណែប្រភេទ

ប្រាក់បុគ្គលិក

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

Ide regular

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

Ide Dec\_hi

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

# 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

คือ ค่าของค่าคงที่

$Q_0$  = 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

ค่าคงที่ก็จะ เท่ากับ  $\alpha$  ค่าของค่าคงที่นี้ คือ  $\alpha = 1$  ค่าของค่าคงที่นี้ คือ  $\beta = 0.75$  และ  $\gamma = 0.25$

$\alpha, \beta$  and  $\gamma$  tune importance of relevant and nonrelevant terms

(in some studies best to set  $\alpha$  to 1  $\beta$  to 0.75 and  $\gamma$  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

$$\gamma = 0.25$$

จำนวน因子ค่าที่

จำนวน keyword หรือเรียกอีกอย่างหนึ่งว่า Vector มีค่า key = คำพิเศษ

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

Rocchio Calculation

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

Resulting feedback query

# 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:

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

$\alpha$ : Tunable weight for initial query.

$\beta$ : Tunable weight for relevant documents.

$\gamma$ : 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 :  $(5,0,3,0,1)$  *and "math." is main keyword*

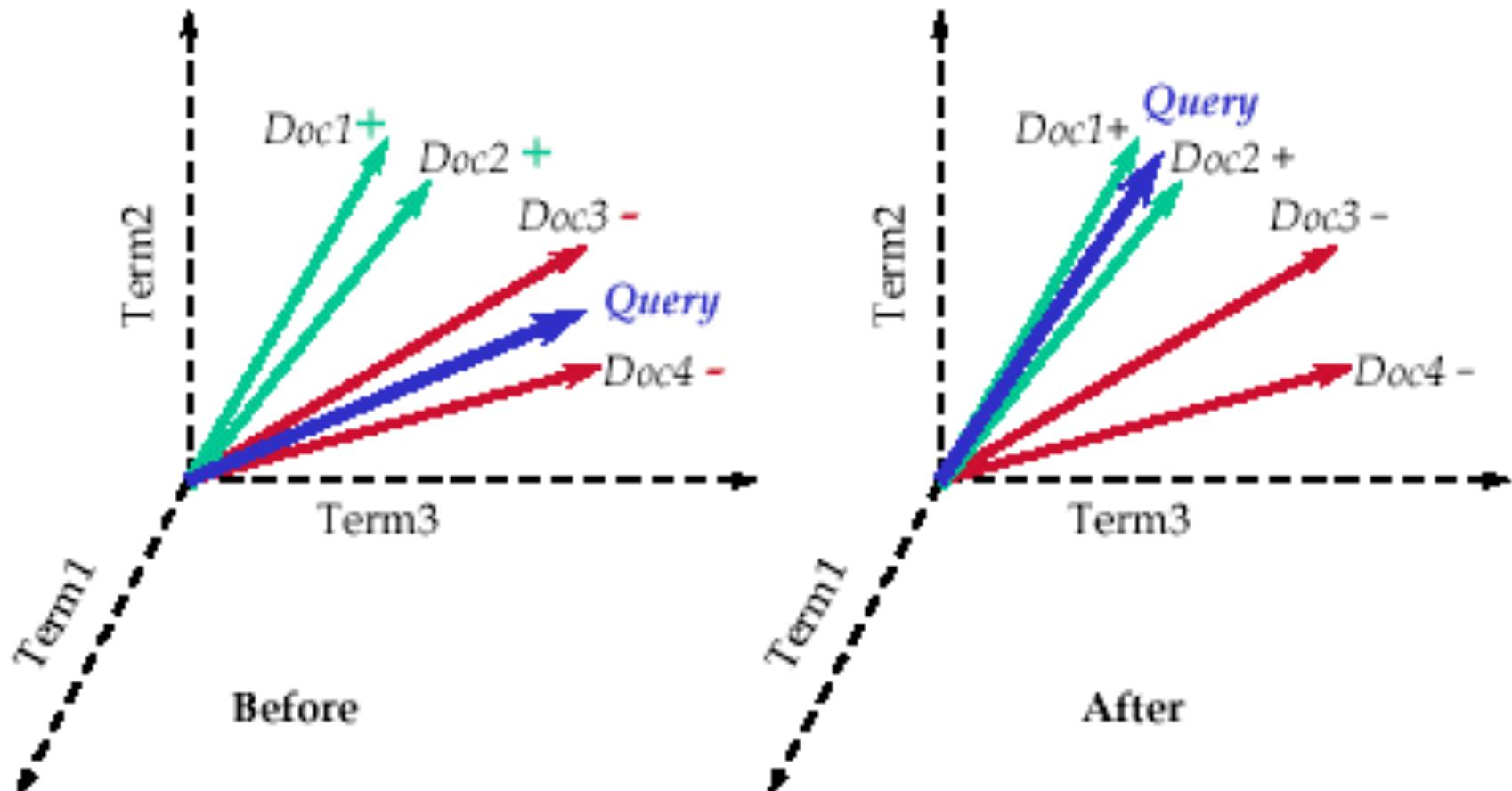
Document  $D_1$  Relevant :  $(2,1,2,0,0)$

Document  $D_2$  Nonrelevant :  $(1,0,0,0,2)$

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

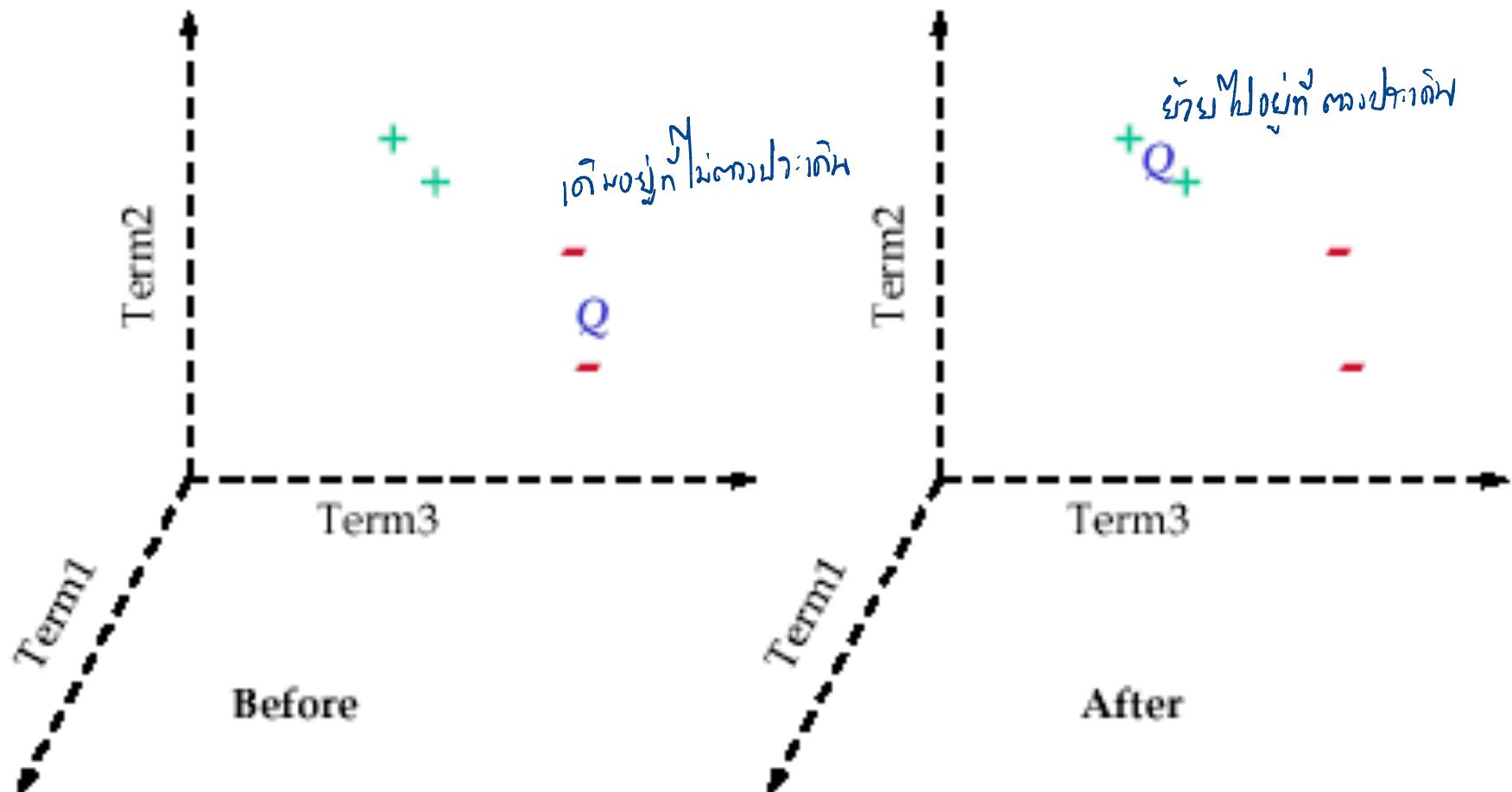
$$\begin{aligned} q' &= q + 0.5D_1 - 0.25D_2 \\ &= (5,0,3,0,1) + 0.5(2,1,2,0,0) - 0.25(1,0,0,0,2) \\ &= (5.75, 0.50, 4.0, 0.0, 0.5) \end{aligned}$$

# Relevance Feedback



# Relevance Feedback

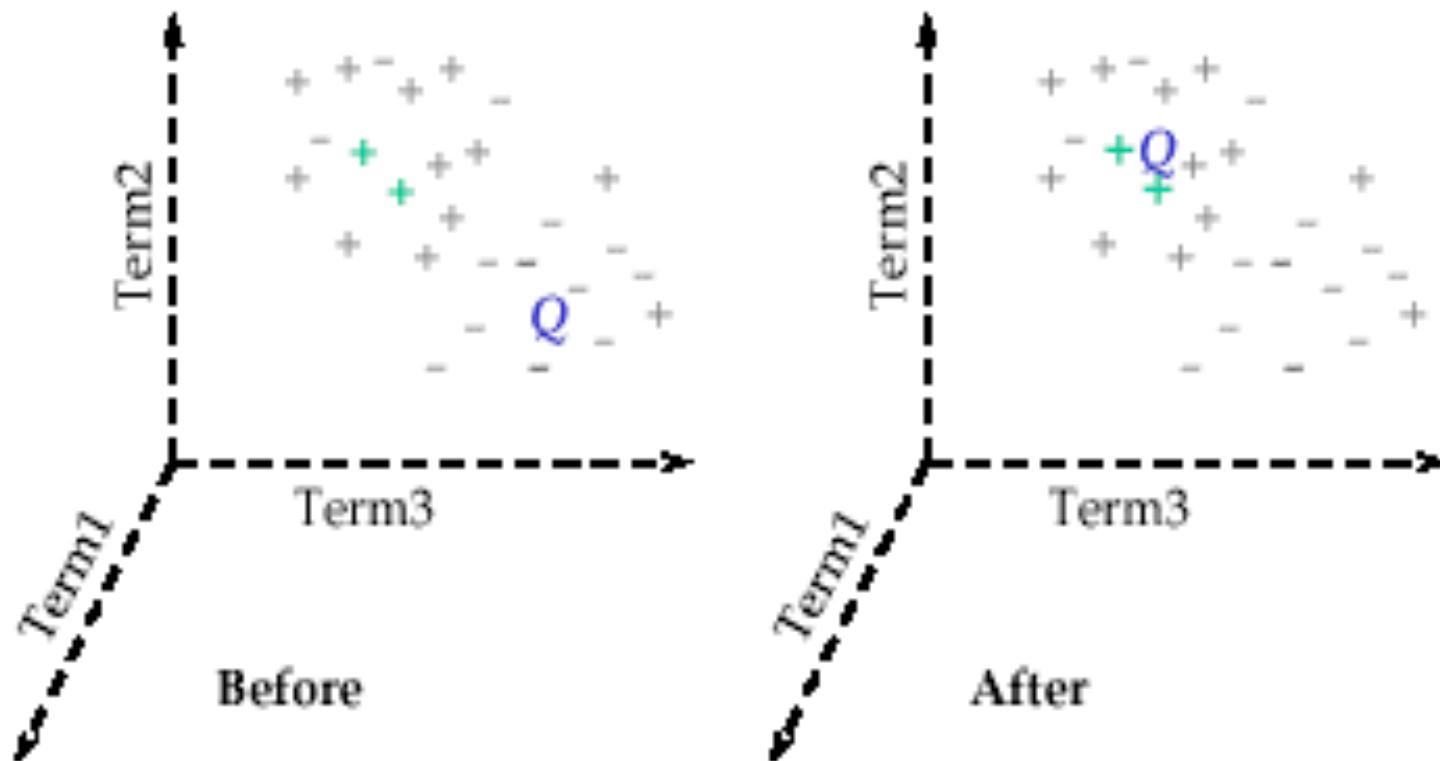
សម្រាប់ការសរសើរ query ក្នុងពាណិជ្ជការ



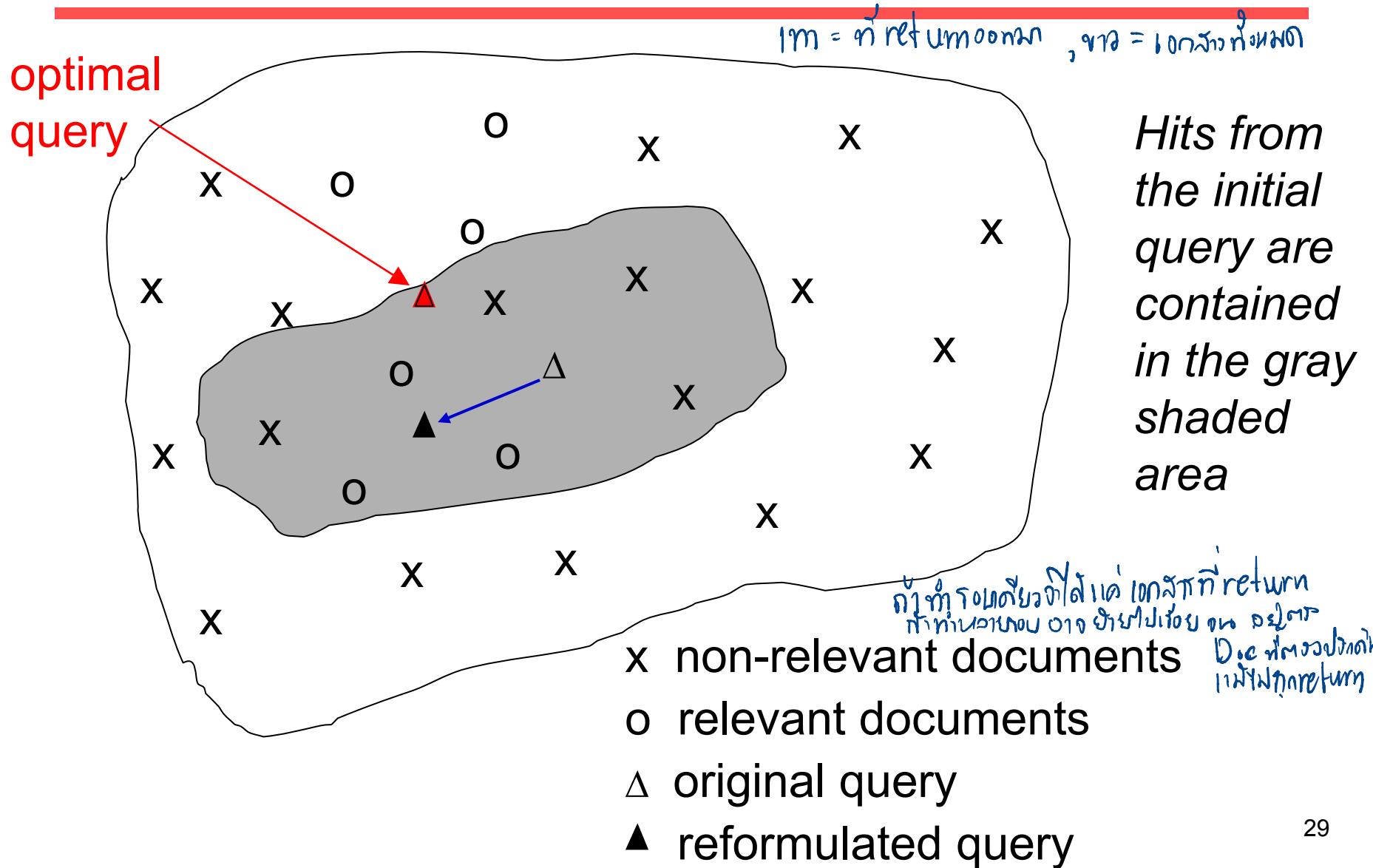
# Relevance Feedback

ຄ່າຍິນທຳກ່ອງອຸປະກອດ

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



# Difficulties with Relevance Feedback



# Vector Space Re-Weighting

---

- The initial query vector  $q_0$  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**

↳ Doc ក្នុង return (ស្ថាកិច្ចរដ្ឋមន្ត្រ)  
↳ Doc ក្នុងនៃ

# 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  $d_j$

$$itf_j = \log \frac{t_{\text{all term}}}{t_j}$$

term in document j

# Term weighting vs. Term concept space

มีหัวใจ

	Vector						$D_1$	$D_2$	....	$D_n$
$D_1$	$W_{11}$	$W_{21}$	...	$W_{t1}$		$\rightarrow$ จัดหัวใจ	$W_{11}$	$W_{12}$	...	$W_{1n}$
$D_2$	$W_{12}$	$W_{22}$	...	$W_{t2}$			$W_{21}$	$W_{22}$	...	$W_{2n}$
:	:	:		:			:	:		:
:	:	:		:			:	:		:
$D_n$	$W_{1n}$	$W_{2n}$	...	$W_{tn}$			$W_{t1}$	$W_{t2}$	...	$W_{tn}$

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

↓ หัวใจ กับ จัดหัวใจ

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

หัวใจ keyword

↓ หัวใจ กับ จัดหัวใจ

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

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

หัวใจ document

# Similarity Thesaurus

- Each term is associated with a vector

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

- ① — where  $w_{i,j}$  <sup>in row</sup> is a **weight** associated to the index-document pair

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

②

- The **relationship between two terms  $k_u$  and  $k_v$**  is <sup>between the two vectors</sup>

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

# Query Expansion Procedure with Similarity Thesaurus

1. និនីកសាយសាស្ត្រ  
2. គិតតម្លៃ

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

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

គិតតម្លៃ

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

(keyword)

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

# Query Expansion based on a Similarity Thesaurus

- A document  $d_j$  is represented term-concept space by  $\vec{d}_j = \sum_{k_v \in d_j} w_{v,j} \times \vec{k}_v$  ចូរការណ៍ទាយ នៅលើក្រុម
- 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 គាំងារក្នុង

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

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

- which is similar to the generalized vector space model

# Automatic Global Analysis Example

$$\begin{array}{c} K_1 \quad D_1 \quad D_2 \quad \dots \quad D_n \\ K_2 \quad w_{11} \quad w_{12} \quad \dots \quad w_{1n} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ K_t \quad w_{t1} \quad w_{t2} \quad \dots \quad w_{tn} \end{array}$$

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

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

# 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.

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

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

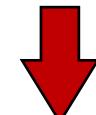
$$\text{sim}(q, k_3) = 3.98$$

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

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

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

$$\text{sim}(q, k_1)$$
  
$$\text{sim}(q, k_2)$$

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

$$\text{sim}(q, k_4)$$
  
$$\text{sim}(q, k_3)$$
  
$$\text{sim}(q, k_1)$$

Original Query

$$q = K_1 + K_4$$

New Query

$$q = K_1 + K_3 + K_4$$

**New Query**

$$q = \mathbf{K_1 + K_2 + K_3 + K_4}$$

# Automatic Global Analysis Example

---

Compute new weight terms for query

## Original Query

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

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

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

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

$w_{3,q'} = 5.4$

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

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

# Automatic Global Analysis Example

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

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

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

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

# 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 term  
Doc j st key term min

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
<small>term keyword frequency</small> $t_j$	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

ສໍາනະລັກ ສິນທີ່ Doc j

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

ເມນວ ( key i ວິຊາໂຄງລອນສົນ)

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

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

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

# Automatic Global Analysis Example

## Term Weight

W	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>
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



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

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

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

$$= 7.987$$

# 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 <sub>i</sub> )
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	

ပေးအနေဖြင့်  
နိုင်ပေးပို့ပါ

ပေးအနေဖြင့်  
လုပ်ချက်များ

**ADD K<sub>2</sub> to Query**

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

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

# Automatic Global Analysis Example

## Recompute term similarity

C	A	B	C	D	E	Sim(q,K <sub>i</sub> )
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	

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

ตั้งค่าเริ่มต้น Sim(q, k<sub>3</sub>) = 32.838

# Automatic Global Analysis Example

原始查询 → 全局化查询

## Compute new weight terms for query

### Original Query

$$q = 2.3K_1 + K_2 + K_3 \quad \text{Sum query weight} = 2.3 + 1 + 1 = 4.3$$

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

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

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

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

---

全局化查询

	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

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

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

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

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

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

# Automatic Global Analysis Example

## Arrange Relevance

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

	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>
Sim(q,d <sub>j</sub> )	1,291.282	1,531.123	1,538.429	1,079.324

Answer = D<sub>3</sub>, D<sub>2</sub>, D<sub>1</sub>, D<sub>4</sub>

# Automatically (Implicit)

---

ລະົງສູ່ຫຸ້ນວກັ້ງເໝາດ

- **Automatic Global Analysis**  
ເບີຍພາກເຕັກ \* ອົບເຄືອກວ່າຫຸ້ນວກັ້ງກ່ຽວ return ຕອນວາ / ດັບກໍ່ສອນໄດ້ຕອນວາ
- **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

- (ဝက်သွေ)
- 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 စံဆောင်ရွက် keyword ရှိသူ့ Doc
- metric clusters ဘုတ္ထော် keyword Ex. This is a book မျက် ၃ ခု
- scalar clusters တော်လောက်လောက် လောက်လောက်

ကြိမ်များ =  $\infty$   
စံဆောင် Doc ပါ၏ 1Doc စံဆောင်  
= စံဆောင် (min Doc ကြံများ)

# Association Clusters

---

- idea
  - Based on the co-occurrence of stems (or terms) **inside documents**
- association matrix
  - $\vec{f}_{s_i,j}$ : the frequency of a **stem**  $s_i$  in a document  $d_j$  ( $\in D_l$ )
  - $m = (\vec{f}_{s_i,j})$ : an association matrix with  $|S_l|$  rows and  $|D_l|$  columns
  - $s = m \vec{m}^T$  : a local **stem-stem** association matrix

# Association Clusters

- Idea
  - co-occurrence of stems (or terms) inside documents (frequency of stems in doc)
$$c(k_u, k_v) = \sum_{j=1}^{|D|} f_{u,j} \times f_{v,j}$$

(1) ໜ້າຕາມສູນໄຟ້າ 1:nin keyword
  - 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  $s(k_u, k_v) = \frac{c(k_u, k_v)}{c(k_u, k_u) + c(k_v, k_v) - c(k_u, k_v)}$ 

(2)  $\uparrow$  ມາແກ່ຍາມເນື້ອ query  
      (b)

# 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  $k_j$

① una-zenho

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

- ② – normalized form

③  $\overbrace{\text{with terms as query}}$   $s(k_u, k_v) = \frac{c(k_u, k_v)}{|V(k_u)| \times |V(k_v)|}$

# Scalar Clusters

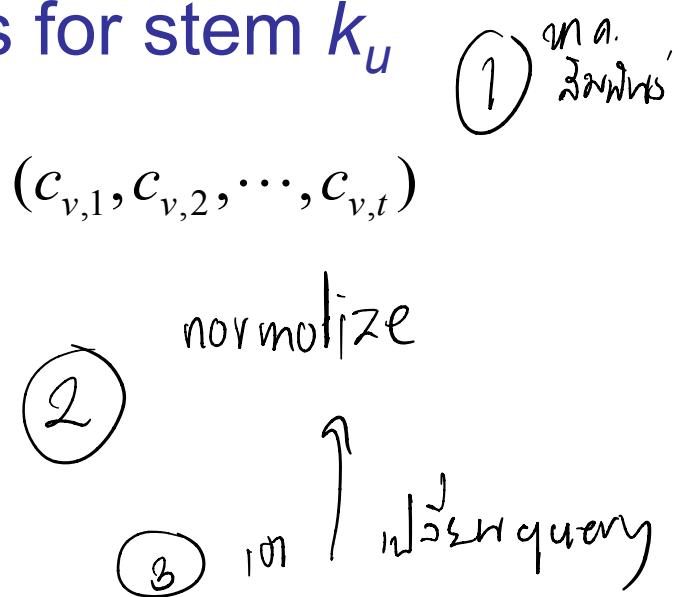
- Idea
  - two stems with similar neighborhoods have some **synonymity** relationships

- *Definition*

- $c_{u,v} = c(k_u, k_v)$
  - vectors of correlation values for stem  $k_u$  and  $k_v$   
 $\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t})$        $\vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$

- 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)

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

- $f_{u,j}$ : 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

②  $s(k_u, k_v) = \frac{c(k_u, k_v)}{c(k_u, k_u) + c(k_v, k_v) - c(k_u, k_v)}$

# Association Clusters

$c_{u,v} = \sum_{dj \in Dl} f_{su,j} \times f_{sv,j}$  : a correlation between the stems  $s_u$  and  $s_v$   
an element in  $\overrightarrow{m}\overrightarrow{m}^t$

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

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

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

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

3  
ज्ञानकेंद्रीय

# Association Clusters Example

$$q = A+B$$

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

$$d_1 = A, A, B, D$$

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

$$d_3 = A, B$$

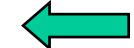
$$d_4 = B, C, D$$

$$d_5 = D$$

$$d_6 = A, B, D$$

$$d_7 = B, B, A$$

	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$
$A$	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



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

$C_{1,1}$

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

๑. សមារិភ័ណនរវាង A នូវ D នៅលើ

# Association Clusters Example

---

*Correlation Matrix ( $C$ )*

	$A$	$B$	$C$	$D$
$A$	8	7	2	4
$B$	7	9	3	4
$C$	2	3	5	3
$D$	4	4	3	5

# Association Clusters Example

*Other way to compute the Correlation Matrix*

$$c = \overrightarrow{m} \overleftarrow{m}^t$$

	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$
$A$	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

$m^t$

$$C_{1,4} = (m_{1,1} * m^t_{1,4}) + (m_{1,2} * m^t_{2,4}) + (m_{1,3} * m^t_{3,4}) + (m_{1,4} * m^t_{4,4}) + (m_{1,5} * m^t_{5,4}) + (m_{1,6} * m^t_{6,4}) + (m_{1,7} * m^t_{7,4})$$

$$= 2 * 1 + 1 * 1 + 1 * 0 + 0 * 1 + 0 * 1 + 1 * 1 + 1 * 0$$

$$= 4$$

# 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

ພົດຕະນາມໄມ້ເກີດ!

## Normalized Correlation Matrix ( $S$ )

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

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

	$A$	$B$	$C$	$D$
$A$	8	7	2	4
$B$	7	9	3	4
$C$	2	3	5	3
$D$	4	4	3	5

# Association Clusters Example

## Normalized Correlation Matrix

ຕົວນິທາຕາ ສິນພັກຊັກບັນຫາຕົ້ນ 1

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>A</b>	1	0.70	0.18	0.44
<b>B</b>	0.70	1	0.27	0.40
<b>C</b>	0.18	0.27	1	0.43
<b>D</b>	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**  
ຖານຂອງ A ໂດຍສະບັບ B ໄນກີບ B ສະບັບ (ຍືນດູຈົກລົງ)

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

	A	B	C	D
A	1	0.70	0.18	0.44
B	0.70	1	0.85	0.63
C	0.18	0.85	1	0.63
D	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

$$q = A+B$$



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

### 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

អំពីរបាយការណ៍ (ក្នុង)

- 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

ສາກສິພກ່າງການເກີມວິຈາ

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

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

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

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

# Metric Clusters Example

ภาษาต้นที่ คิดตามใจ

$$q = A + 2D$$

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

$S_1$  ภาษาต้นที่ 1

A, B, C base on  $S_1$  stem

D, E base on  $S_2$  stem

F base on  $S_3$  stem

Then

ภาษาต้นที่ 1 มี root 2 keyword

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

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

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

ภาษาต้นที่ return no word .... ห้ามกับ .... keyword keyword  
ห้าม ( ไม่ต้อง )

ระยะห่าง	A	B	C	D	E	F
ห่าง						
A	0	5	<small>ภาษาต้นที่ 1 มี root 2 keyword</small> $\infty$	$\infty$	1	2
B	5	0	3	2	1	1
C	$\infty$	3	0	3	4	$\infty$
D	$\infty$	2	3	0	$\infty$	5
E	1	1	4	$\infty$	0	1
F	2	1	$\infty$	5	1	0

# Metric Clusters Example

លទ្ធផល និង អារគុណ  
→ រូបភាពនេះ

ន័យោះ អំពី	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>
<b>A</b>	0	5	$\infty$	$\infty$	1	2
<b>B</b>	5	0	3	2	1	1
<b>C</b>	$\infty$	3	0	3	4	$\infty$
<b>D</b>	$\infty$	2	3	0	$\infty$	5
<b>E</b>	1	1	4	$\infty$	0	1
<b>F</b>	2	1	$\infty$	5	1	0



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

# Metric Clusters Example

๗. จงพิจารณาค่าต่อไปนี้ ว่าเป็น  $S_1$  หรือ  $S_2$   
 $A \times D, A \times E, B \times D, BE, CD, C \times E$

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

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

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

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

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

ค่าคงที่  $c_{1,2} = c(A,D) + c(A,E) + c(B,D) + c(B,E) + c(C,D) + c(C,E)$

$$= 0 + 1 + 0.50 + 1 + 0.33 + 0.25$$

$$= 3.08$$

# 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

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

↑Normalized

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

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

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

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

# 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

$$q = A + 2D \rightarrow q' = (S_1 + 0.51S_2) + 2*(S_2 + 0.60S_3)$$

$\downarrow$   
obj. matrix

$$= S_1 + 2.51S_2 + 1.2S_3$$

### New Query

# Scalar Clusters

keyword ពាណិជ្ជកម្ម

- 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  $k_v$   
 $\vec{s}_u = (c_{u,1}, c_{u,2}, \dots, c_{u,t})$        $\vec{s}_v = (c_{v,1}, c_{v,2}, \dots, c_{v,t})$   
  minh keyword keyword 2
  - scalar association matrix
  - scalar clusters
    - the set of  $k$  **largest values** of scalar association

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

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

# Scalar Clusters

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

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

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

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

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

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

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

$C_{\text{Database, Algebra}}$ ,  $C_{\text{Database, Math}}$ ,  $C_{\text{Database, Set}}$

$C_{\text{AI, Algebra}}$ ,  $C_{\text{AI, Math}}$ ,  $C_{\text{AI, Set}}$

$C_{\text{Network, Algebra}}$ ,  $C_{\text{Network, Math}}$ ,  $C_{\text{Network, Set}}$

କେବଳ - କେବଳ

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

# Scalar Clusters

Normalize

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

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

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

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

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

1 norm vector  $\sqrt{5^2 + 6^2 + 1^2}$

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

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

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

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

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

2

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

$$S_{1,3} = \frac{\vec{s}_1 \bullet \vec{s}_3}{|\vec{s}_1| \times |\vec{s}_3|} = \frac{(5 \times 1) + (6 \times 0) + (1 \times 2)}{7.874 \times 2.236}$$

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

# Scalar Clusters

*Normalized Correlation Matrix (S)*

$S$	$S_1$	$S_2$	$S_3$
$S_1$	1	0.986	0.398
$S_2$	0.986	1	0.248
$S_3$	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*

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

গুরুত্বপূর্ণ কাজ

# 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