Information Retrieval & Text Mining

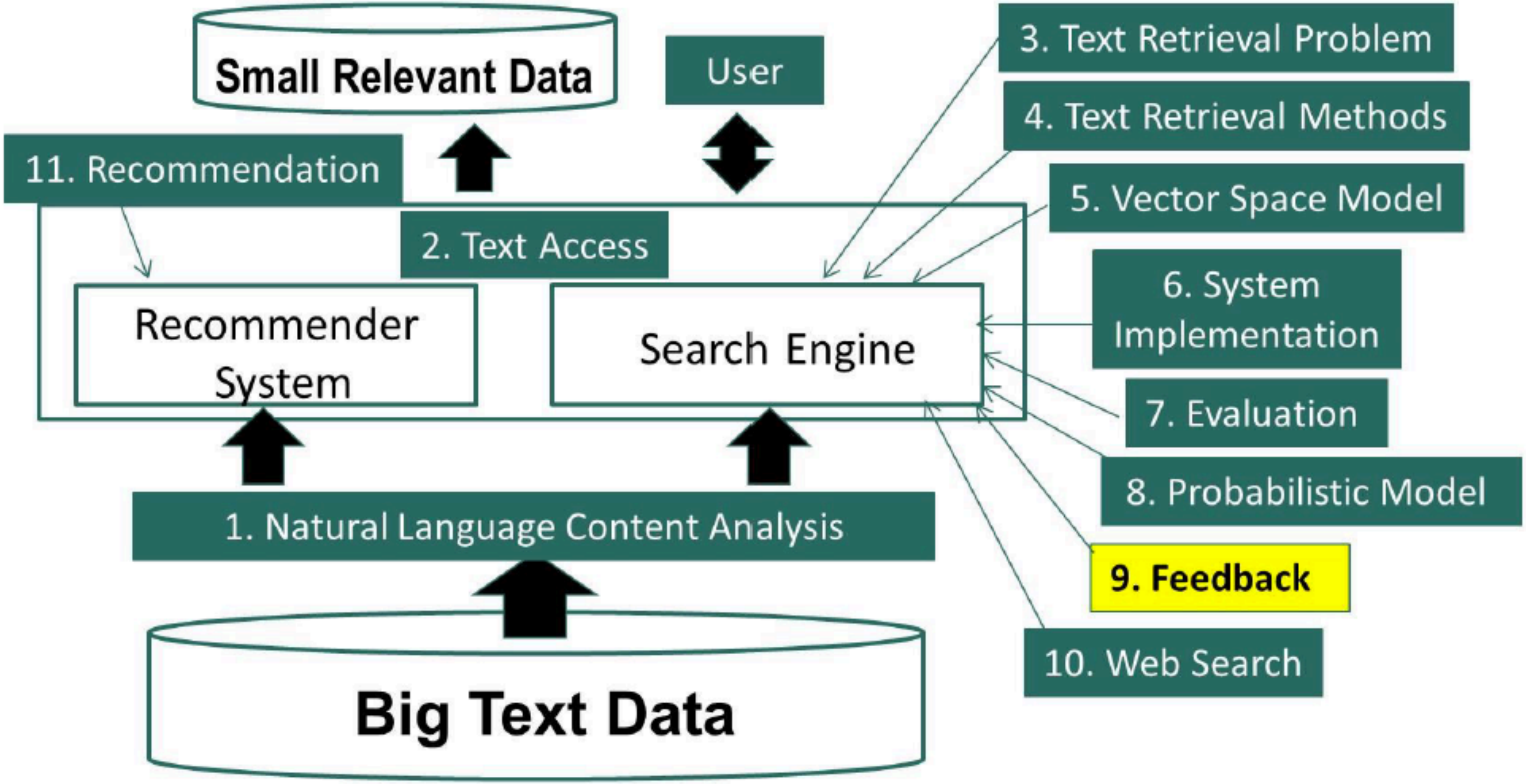
Retrieval Method:

Feedback in VSM

Dr. Saeed Ul Hassan

Information Technology University

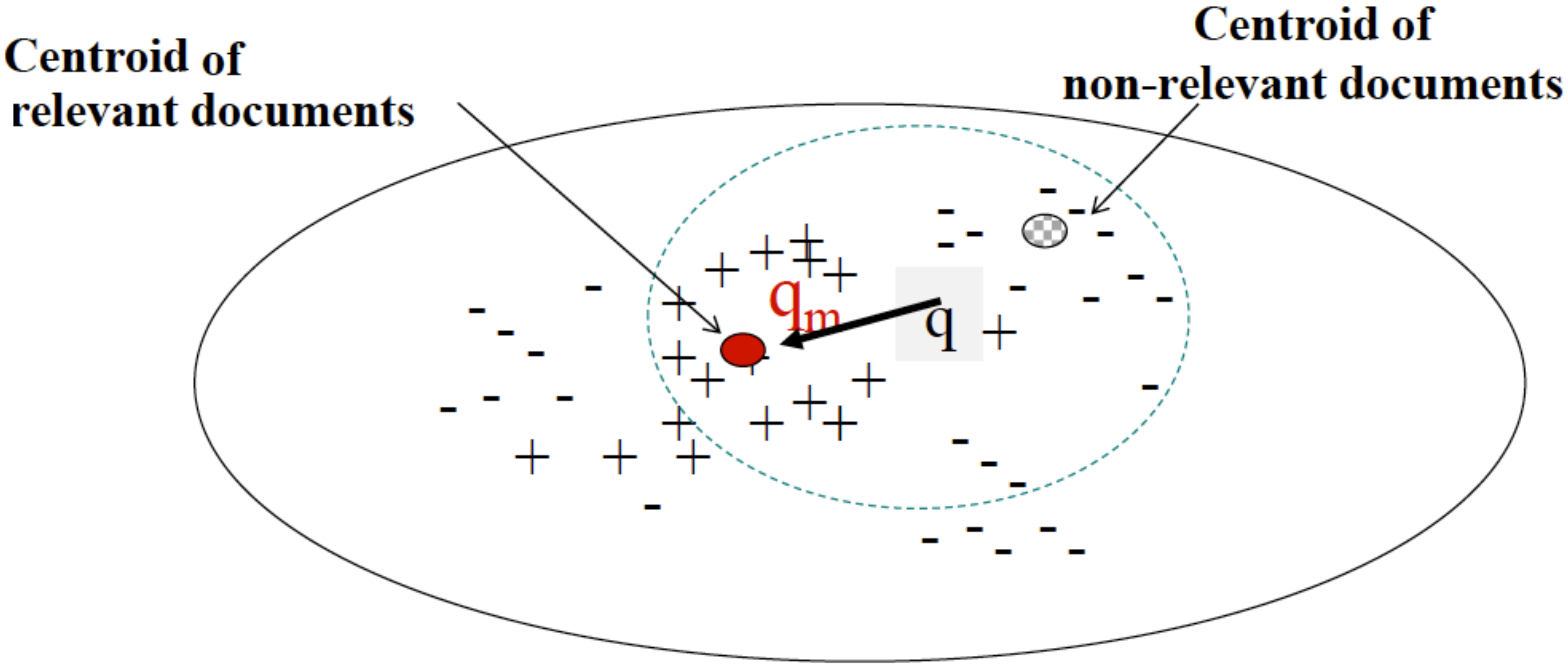
Text Retrieval Methods: Feedback in TR



Feedback in Vector Space Model

- How can a TR system learn from examples to improve retrieval accuracy?
  - Positive examples: docs known to be relevant
  - Negative examples: docs known to be non-relevant
- General method: query modification
  - Adding new (weighted) terms (query expansion)
  - Adjusting weights of old terms

Rocchio Feedback: Illustration



Rocchio Feedback: Formula

Parameters

New query

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

Rocchio Feedback: Formula

Parameters

New query

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

Origial query

Rel docs

Non-rel docs

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

D1 … news about …

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2 … news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3 … news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

D4 … news of presidential campaign …
… presidential candidate …

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

D5 … news of organic food campaign… campaign…campaign…campaign…

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

D1 … news about …

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2 … news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3 … news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

D5 … news of organic food campaign… campaign…campaign…campaign…

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = “news about presidential campaign”

Q= (1, 1, 1, 1, 0, 0, …)

- **D1= (1.5, 0.1, 0, 0, 0, 0, …)**

D2

… news about organic food campaign…

- **D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)**

D3

… news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

D4

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

- Centroid Vector= ((1.5+1.5+1.5)/3, (0.1+0.1+0)/3, 0, (0+2.0+6.0)/3, (0+2.0+2.0)/3, 0, …)
=(1.5 , 0.067, 0, 2.6, 1.3, 0,…)

- **D5= (1.5, 0, 0, 6.0, 2.0, 0, …)**

Example of Rocchio Feedback

V= {news about presidential camp. food …. }

Query = “news about presidential campaign”

$Q= (1, 1, 1, 1, 0, 0, \ldots)$

New Query $Q' = (\alpha*1+\beta*1.5-\gamma*1.5,\ \alpha*1-\gamma*0.067,\ \alpha*1+\beta*3.5,\ \alpha*1+\beta*2.0-\gamma*2.6,\ -\gamma*1.3,\ 0,\ 0,\ \ldots)$

- D1= (1.5, 0.1, 0, 0, 0, 0, …)

D2

… news about organic food campaign…

- D2= (1.5, 0.1, 0, 2.0, 2.0, 0, …)

D3

… news of presidential campaign …

+ D3= (1.5, 0, 3.0, 2.0, 0, 0, …)

D4

+ Centroid Vector= ((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, …)
=(1.5 , 0, 3.5, 2.0, 0, 0,…)

+ D4= (1.5, 0, 4.0, 2.0, 0, 0, …)

- Centroid Vector= ((1.5+1.5+1.5)/3, (0.1+0.1+0)/3, 0, (0+2.0+6.0)/3, (0+2.0+2.0)/3, 0, …)
=(1.5 , 0.067, 0, 2.6, 1.3, 0,…)

- D5= (1.5, 0, 0, 6.0, 2.0, 0, …)

Rocchio in Practice

- Negative (non-relevant) examples are not very important (why?)
- Often truncate the vector (i.e., consider only a small number of words that have highest weights in the centroid vector) (efficiency concern)
- Avoid "over-fitting" (keep relatively high weight on the original query weights) (why?)
- Can be used for relevance feedback and pseudo feedback (β should be set to a larger value for relevance feedback than for pseudo feedback)
- Usually robust and effective