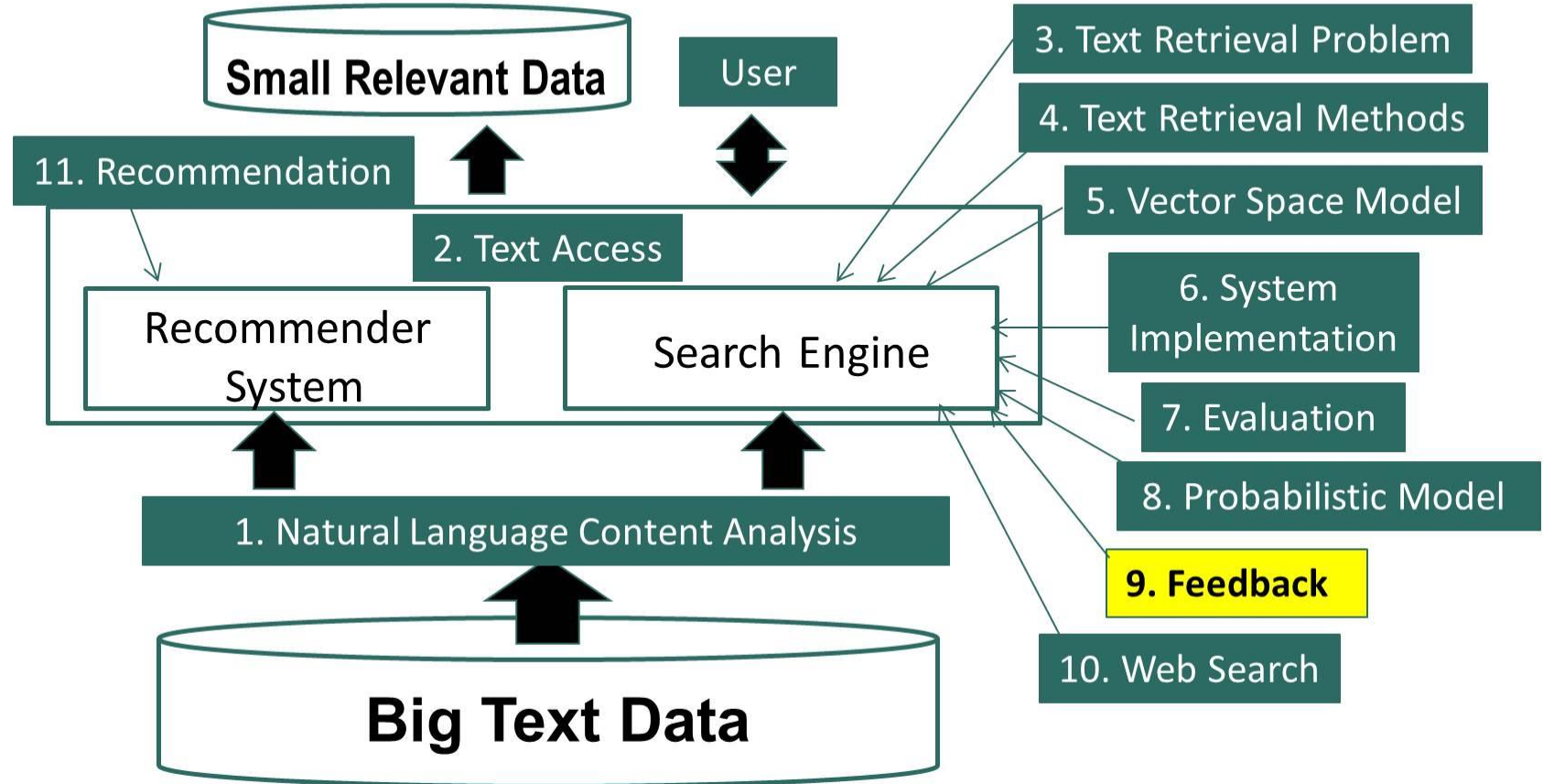


# Text Retrieval and Search Engines

Feedback in Text Retrieval: Feedback in VSM

ChengXiang “Cheng” Zhai  
Department of Computer Science  
University of Illinois at Urbana-Champaign

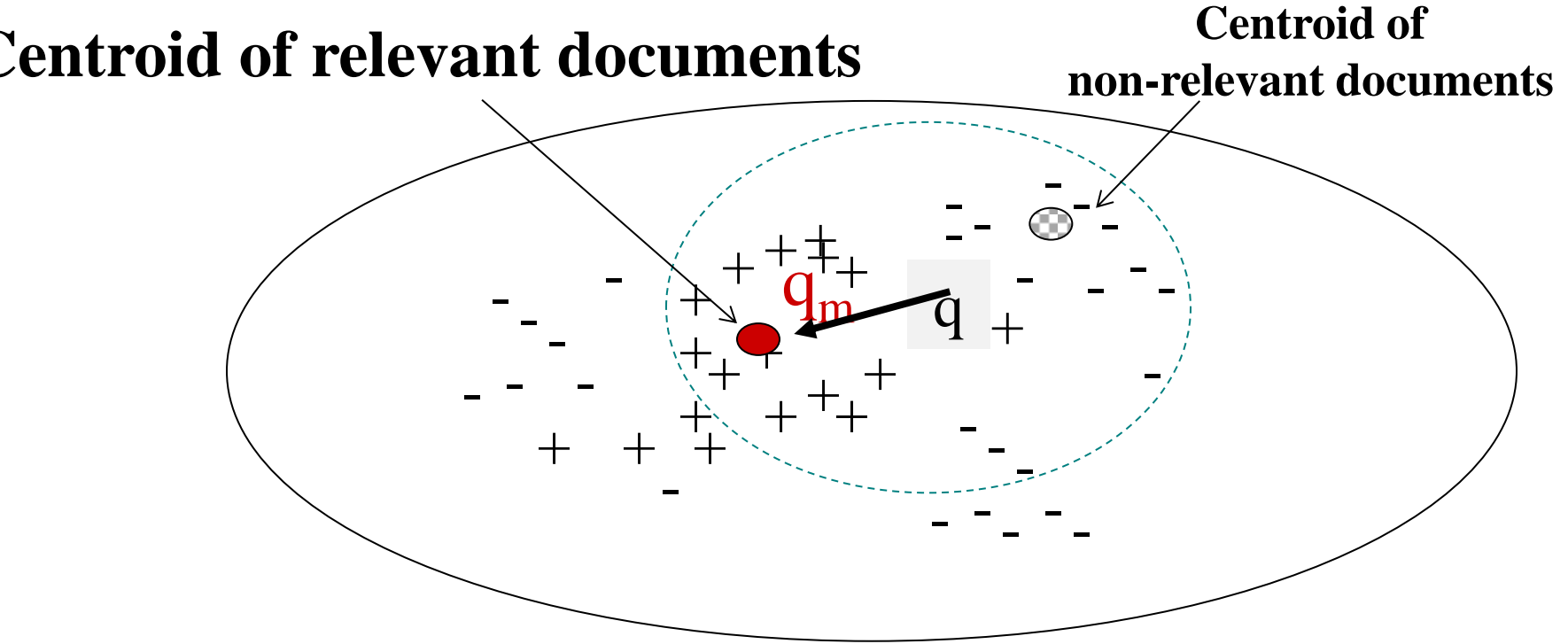
# Feedback in Text Retrieval: Feedback in VSM



# 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

New query

Parameters

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

Original query

Rel docs

Non-rel docs

The diagram illustrates the Rocchio Feedback formula. At the top, the word 'Parameters' has three arrows pointing to the coefficients  $\alpha$ ,  $\beta$ , and  $\gamma$  in the formula. On the left, 'New query' has an arrow pointing to  $\vec{q}_m$ . Below the formula, 'Original query' has an arrow pointing to  $\vec{q}$ . 'Rel docs' has an arrow pointing to the summation term  $\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j$ . 'Non-rel docs' has an arrow pointing to the summation term  $\sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$ .

# Example of Rocchio Feedback

$V = \{\text{news about presidential camp. food ....}\}$

Query = "news about presidential campaign"

$O = (1, 1, 1, 1, 0, 0, \dots)$

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

-  $D1 = (1.5, 0.1, 0, 0, 0, 0, \dots)$

D2

... news about organic food campaign...

-  $D2 = (1.5, 0.1, 0, 2.0, 2.0, 0, \dots)$

D3

... news of presidential campaign ...

+  $D3 = (1.5, 0, 3.0, 2.0, 0, 0, \dots)$

D4

+ Centroid Vector =  $((1.5+1.5)/2, 0, (3.0+4.0)/2, (2.0+2.0)/2, 0, 0, \dots)$   
 $= (1.5, 0, 3.5, 2.0, 0, 0, \dots)$

+  $D4 = (1.5, 0, 4.0, 2.0, 0, 0, \dots)$

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

-  $D5 = (1.5, 0, 0, 6.0, 2.0, 0, \dots)$

# 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 ( $\beta$  should be set to a larger value for relevance feedback than for pseudo feedback)
- Usually robust and effective