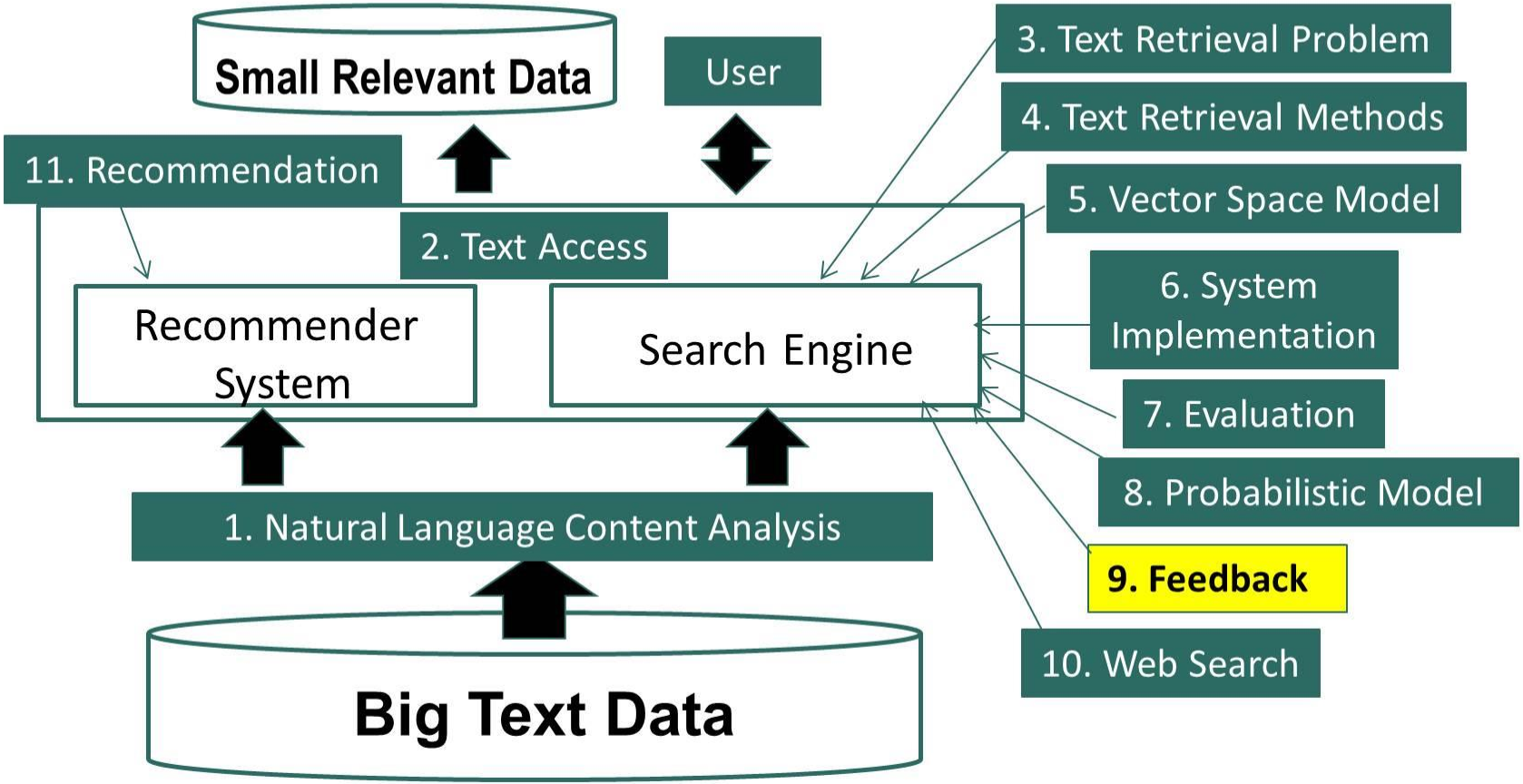
Text Retrieval and Search Engines

Feedback in Text Retrieval: Feedback in VSM

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

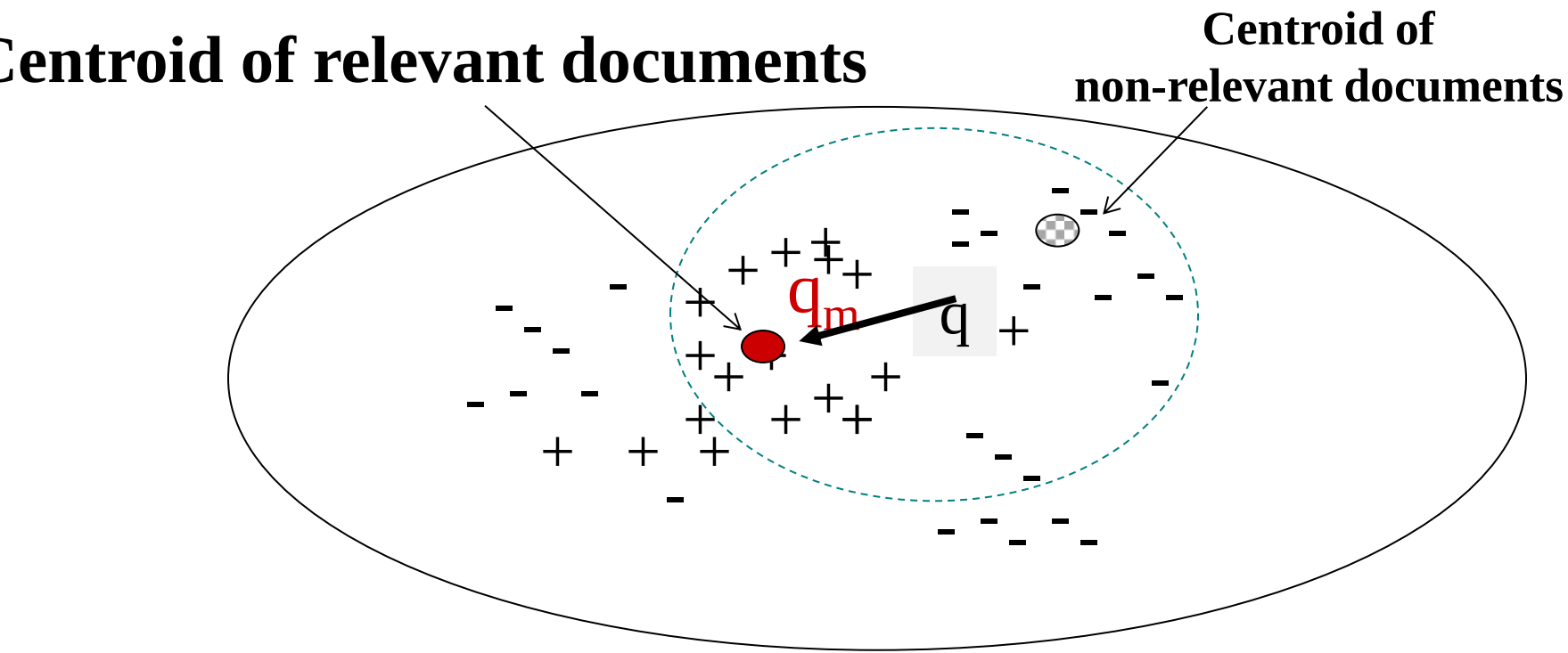
Feedback in Text Retrieval: Feedback in VSM



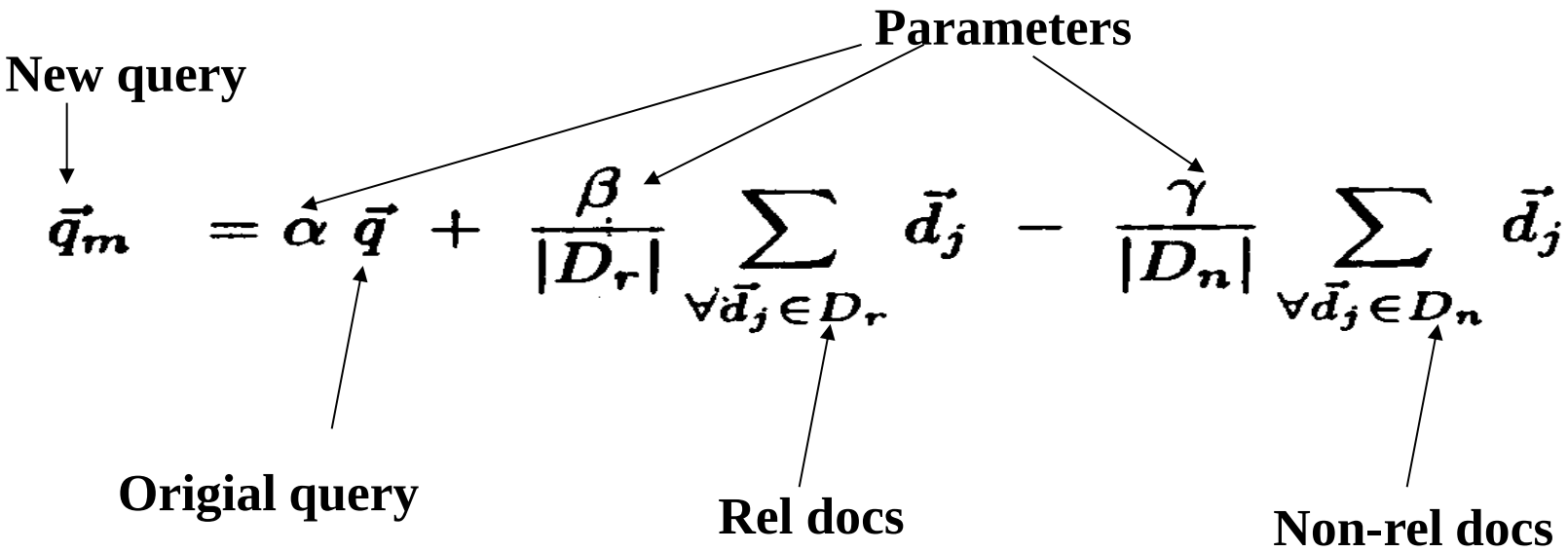
Feedback in Vector Space Model

- 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



Example of Rocchio Feedback

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

Query = "news about presidential campaign"

Q= (1, 1, 1, 1, 0, 0, …)

New Query Q'= (α*1+β*1.5-γ*1.5, α*1-γ*0.067, α*1+β*3.5, α*1+β*2.0-γ*2.6, -γ*1.3, 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, …)

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