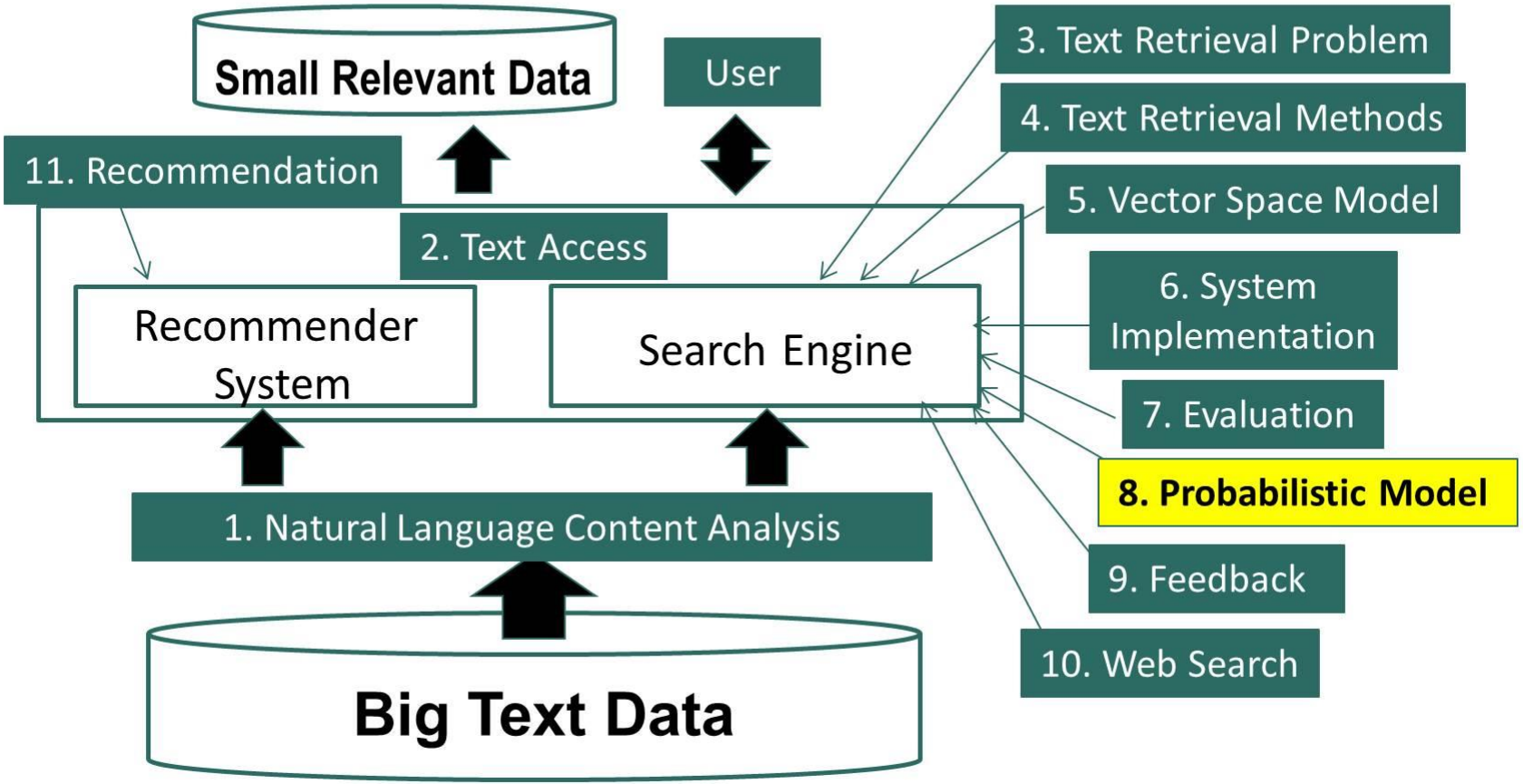
Text Retrieval and Search Engines

Probabilistic Retrieval Model: Smoothing

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

Probabilistic Retrieval Model: Smoothing



Benefit of Rewriting

- Better understanding of the ranking function
  - Smoothing with p(w|C) ➔ TF-IDF weighting + length norm.

TF weighting

Doc length normalization

$$\log p(q \mid d) = \sum_{\substack{w_i \in d \\ w_i \in q}} [\log \frac{p_{Seen}(w_i \mid d)}{\alpha_d \, p(w_i \mid C)}] + n \log \alpha_d + \sum_{i=1}^{n} \log p(w_i \mid C)$$

matched query terms

IDF weighting

Ignore for ranking

- Enable efficient computation

Summary

- Smoothing of p(w|d) is necessary for query likelihood
- General idea: smoothing with p(w|C)
  - The probability of an unseen word in d is assumed to be proportional to p(w|C)
  - Leads to a general ranking formula for query likelihood with TF-IDF weighting and document length normalization
  - Scoring is primarily based on sum of weights on matched query terms
- However, how exactly should we smooth?