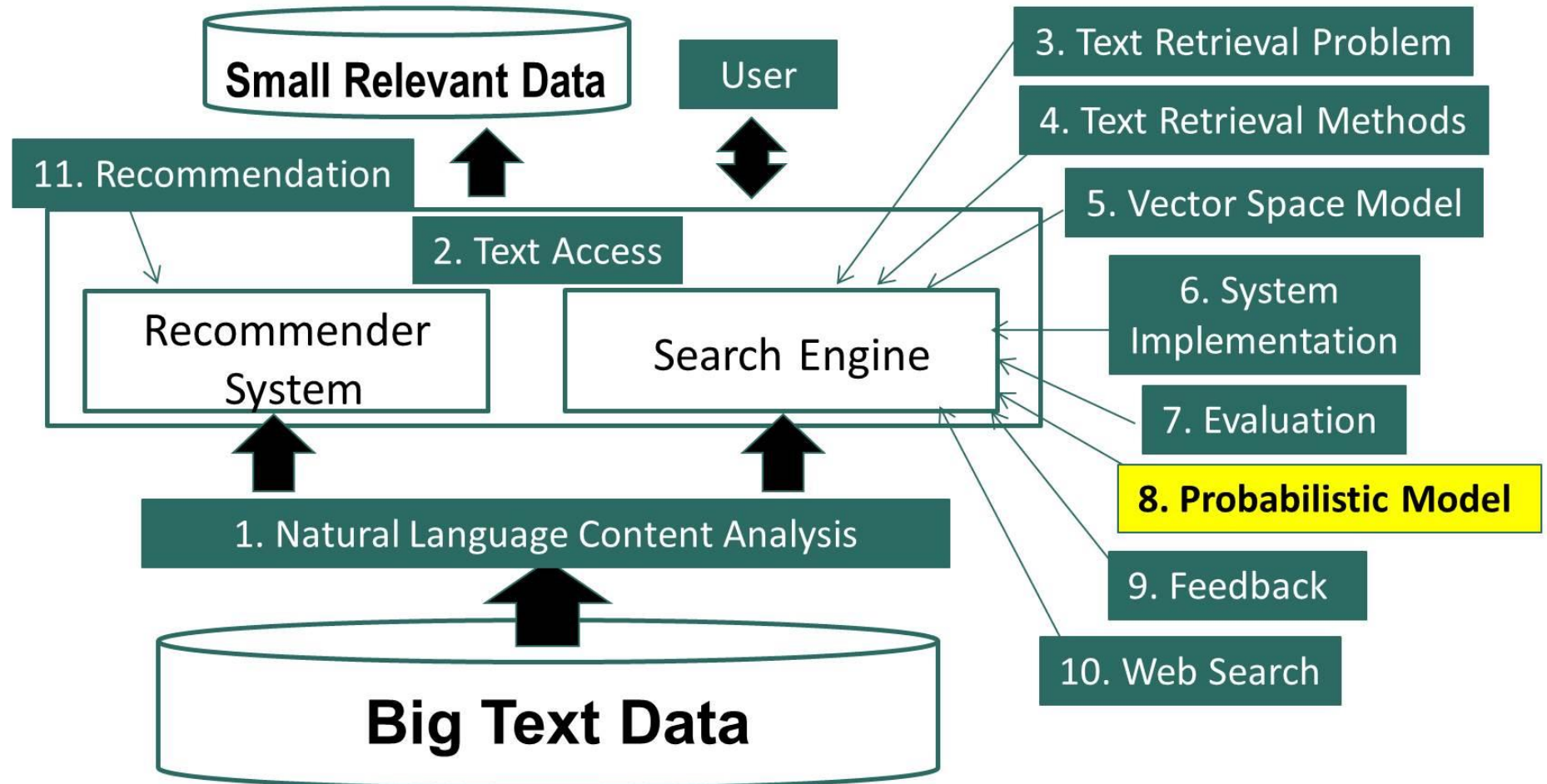# Text Retrieval and Search Engines
Probabilistic Retrieval Model: Smoothing Methods

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

Probabilistic Retrieval Model: Smoothing Methods

Ranking Function for JM Smoothing

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

$$p(w \mid d) = (1 - \lambda) \frac{c(w,d)}{\mid d \mid} + \lambda \, p(w \mid C) \qquad \lambda \in [0,1]$$

$$\frac{p_{seen}(w \mid d)}{\alpha_d p(w \mid C)} = \frac{(1-\lambda)p_{ML}(w \mid d) + \lambda p(w \mid C)}{\lambda p(w \mid C)} = 1 + \frac{1-\lambda}{\lambda} \frac{c(w,d)}{\mid d \mid p(w \mid C)}$$

$$f_{JM}(q,d) = \sum_{\substack{w \in d \\ w \in q}} c(w,q) \log[1 + \frac{1-\lambda}{\lambda} \frac{c(w,d)}{\mid d \mid p(w \mid C)}]$$

Ranking Function for Dirichlet Prior Smoothing

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

$$p(w \mid d) = \frac{c(w;d) + \mu p(w \mid C)}{d \mid + \mu} = \frac{\mid d \mid}{\mid d \mid + \mu} \frac{c(w,d)}{\mid d \mid} + \frac{\mu}{\mid d \mid + \mu} p(w \mid C)$$

$$\mu \in [0, +\infty)$$

$$\frac{p_{seen}(w \mid d)}{\alpha_d p(w \mid C)} = \frac{\dfrac{c(w,d) + \mu p(w \mid C)}{\mid d \mid + \mu}}{\dfrac{\mu p(w \mid C)}{\mid d \mid + \mu}} = 1 + \frac{c(w,d)}{\mu p(w \mid C)} \qquad \alpha_d = \frac{\mu}{\mid d \mid + \mu}$$

$$f_{DIR}(q,d) = [\sum_{\substack{w \in d \\ w \in q}} c(w,q) \log[1 + \frac{c(w,d)}{\mu p(w \mid C)}]] + n \log \frac{\mu}{\mu + \mid d \mid}$$

Summary

- Two smoothing methods
  - Jelinek-Mercer: Fixed coefficient linear interpolation
  - Dirichlet Prior: Adding pseudo counts; adaptive interpolation
- Both lead to state of the art retrieval functions with assumptions clearly articulated (less heuristic)
  - Also implementing TF-IDF weighting and doc length normalization
  - Has precisely one (smoothing) parameter

Summary of Query Likelihood Probabilistic Model

- Effective ranking functions obtained using pure probabilistic modeling
  - Assumption 1: Relevance(q,d) = $p(R=1|q,d) \approx p(q|d,R=1) \approx \mathbf{p(q|d)}$
  - Assumption 2: Query words are generated independently
  - Assumption 3: Smoothing with $p(w|C)$
  - Assumption 4: JM **or** Dirichlet prior smoothing
- Less heuristic compared with VSM
- Many extensions have been made [Zhai 08]

Additional Readings

- ChengXiang Zhai, *Statistical Language Models for Information Retrieval* (Synthesis Lectures Series on Human Language Technologies), Morgan & Claypool Publishers, 2008.

http://www.morganclaypool.com/doi/abs/10.2200/S00158ED1V01Y200811HLT001