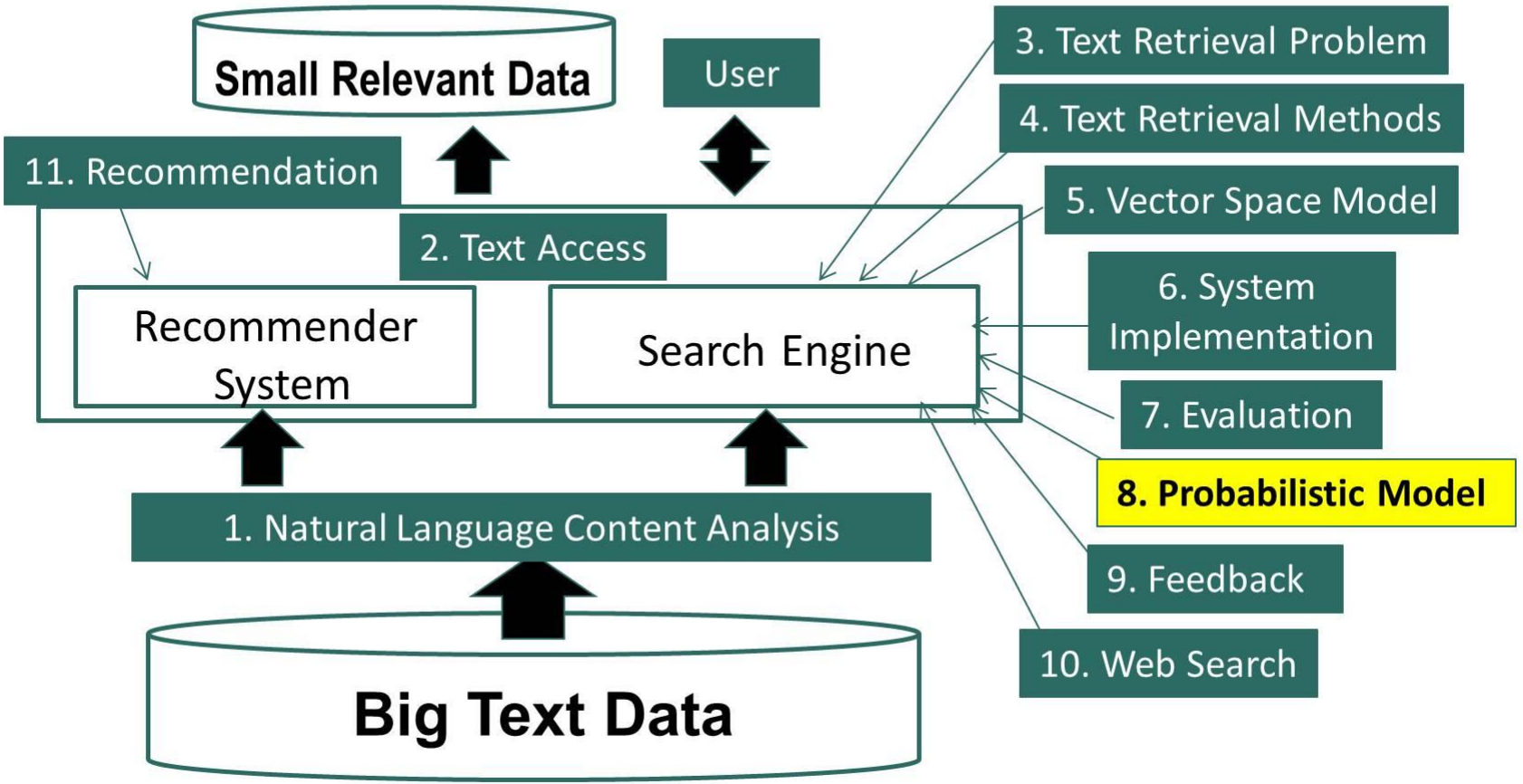
Text Retrieval and Search Engines

Probabilistic Retrieval Model: Smoothing Methods

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

Probabilistic Retrieval Model: Smoothing Methods



Query Likelihood + Smoothing with p(w|C)

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

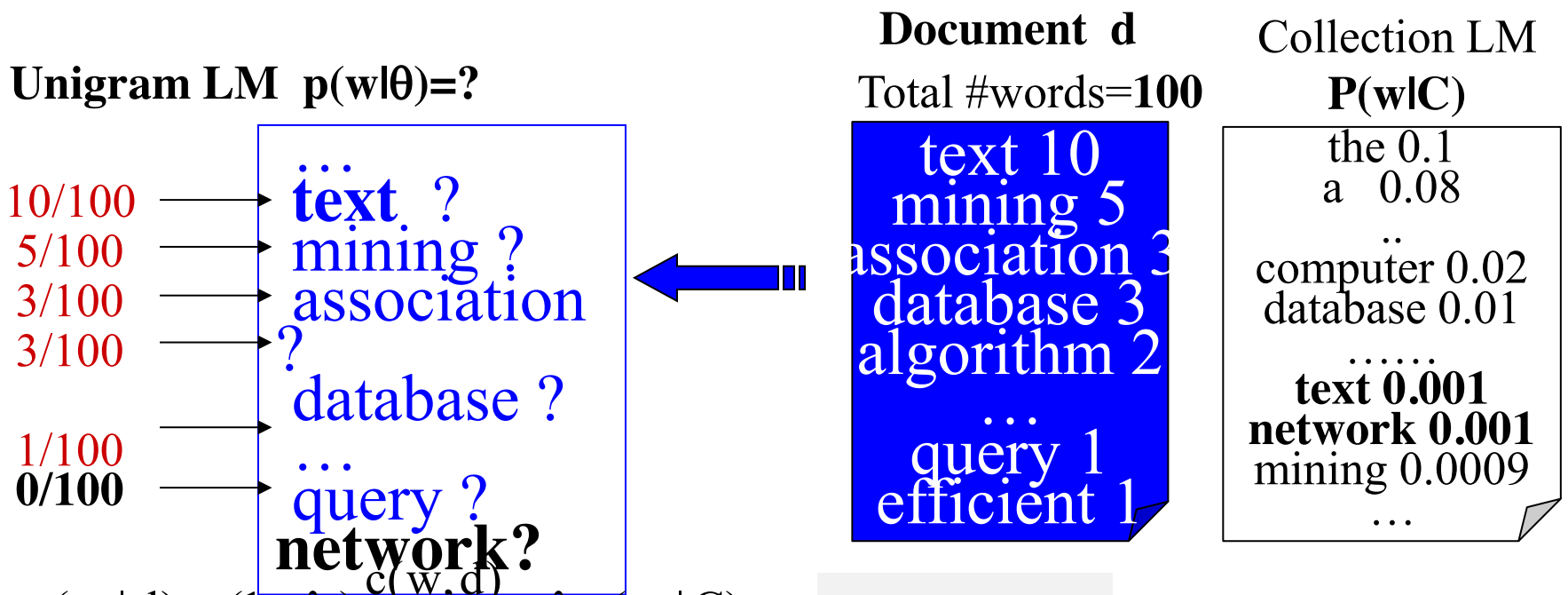
$$f(q,d) = \sum_{\substack{w_i \in d \\ w_i \in q}} c(w,q)[\log \frac{p_{Seen}(w_i \mid d)}{\alpha_d \, p(w_i \mid C)}] + n \log \alpha_d$$

$$p_{Seen}(w_i \mid d) = ?$$

$$\alpha_d = ?$$

How to smooth p(w|d)?

Linear Interpolation (Jelinek-Mercer) Smoothing



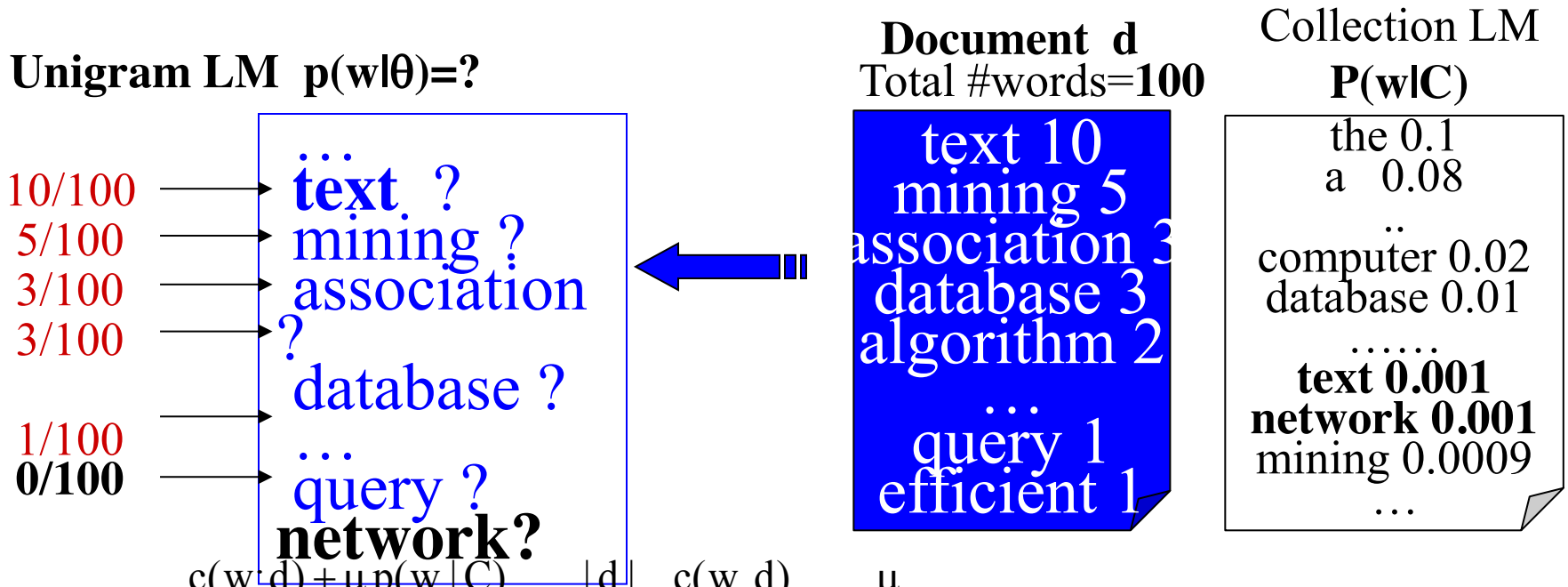
$$p(w \mid d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda\, p(w \mid C)$$

$$p(\text{"text"} \mid d) = (1-\lambda)\frac{10}{100} + \lambda * 0.001$$

$$\lambda \in [0,1]$$

$$p(\text{"network"} \mid d) = \lambda * 0.001$$

Dirichlet Prior (Bayesian) Smoothing



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

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

$$p("\text{text}"|d) = \frac{10 + \mu * 0.001}{100 + \mu}$$

$$p("\text{network}"|d) = \frac{\mu}{100 + \mu} * 0.001$$