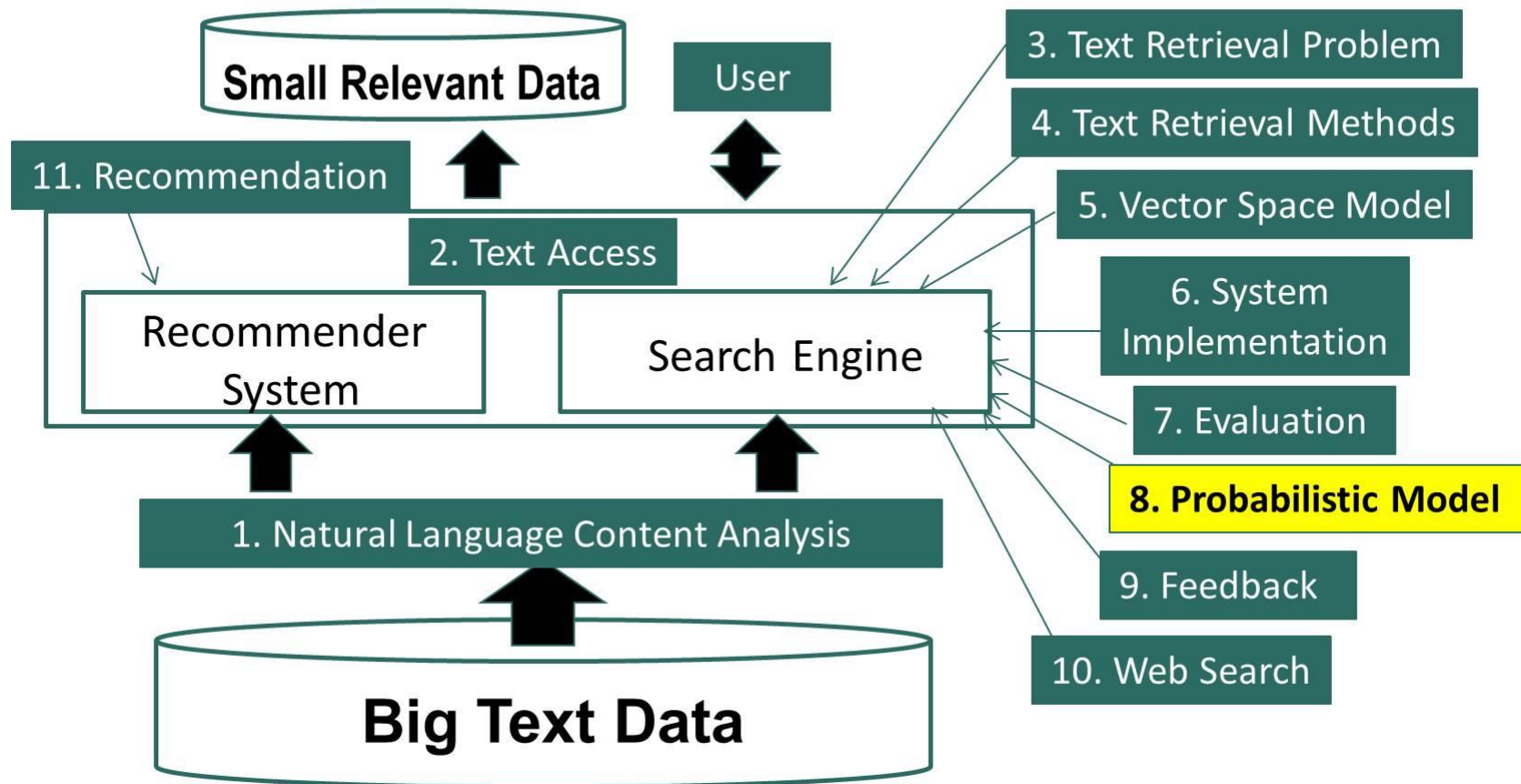


Text Retrieval and Search Engines

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 $p(w | C)$

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

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

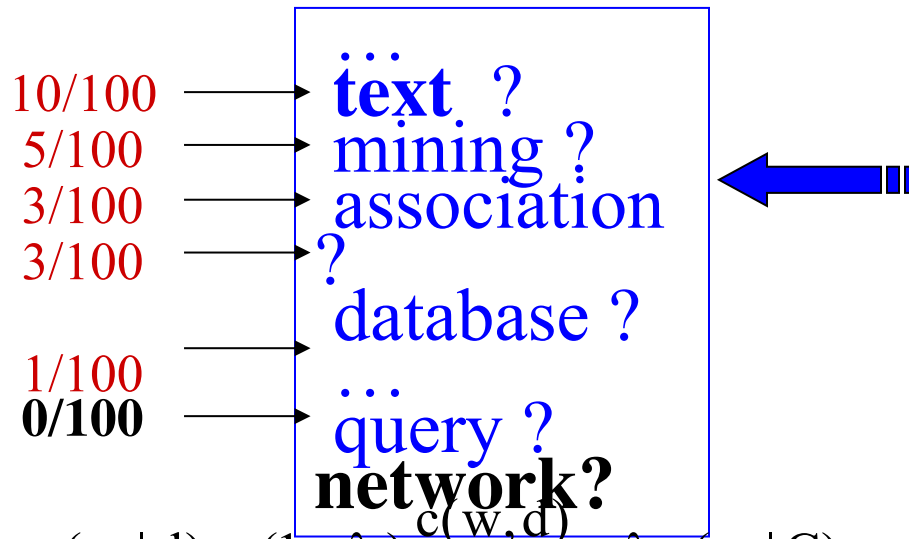
$$\boxed{p_{\text{Seen}}(w_i | d) = ?}$$

$$\boxed{\alpha_d = ?}$$

How to smooth $p(w | d)$?

Linear Interpolation (Jelinek-Mercer) Smoothing

Unigram LM $p(w|\theta)=?$



$$p(w | d) = (1 - \lambda) \frac{c(w, d)}{|d|_0} + \lambda p(w | C)$$

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

Document d
 Total #words=100

text 10
 mining 5
 association 3
 database 3
 algorithm 2
 query 1
 efficient 1

Collection LM
 $P(w|C)$

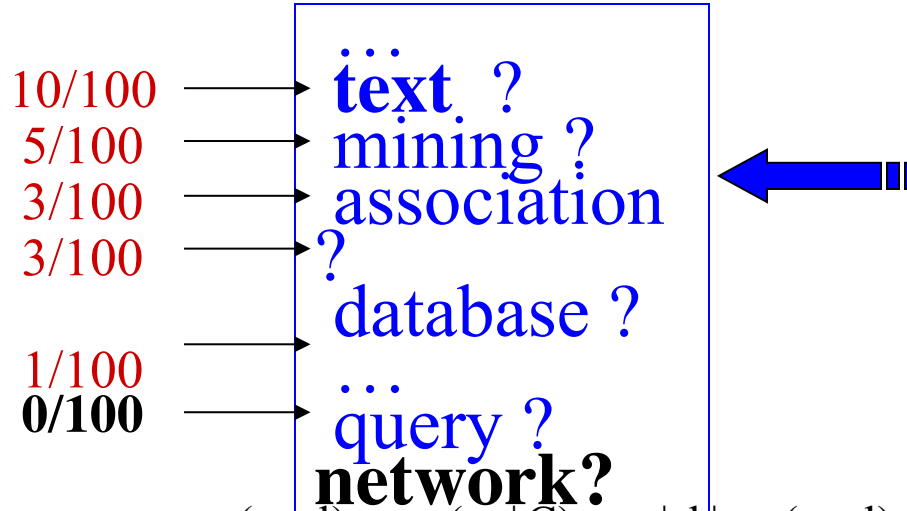
the 0.1
 a 0.08
 computer 0.02
 database 0.01
 text 0.001
 network 0.001
 mining 0.0009
 ...

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

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

Dirichlet Prior (Bayesian) Smoothing

Unigram LM $p(w|\theta)=?$



Document d
 Total #words=100

text 10
 mining 5
 association 3
 database 3
 algorithm 2
 query 1
 efficient 1

Collection LM
 $P(w|C)$

the 0.1
 a 0.08
 computer 0.02
 database 0.01
 text 0.001
 network 0.001
 mining 0.0009
 ...

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

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

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

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