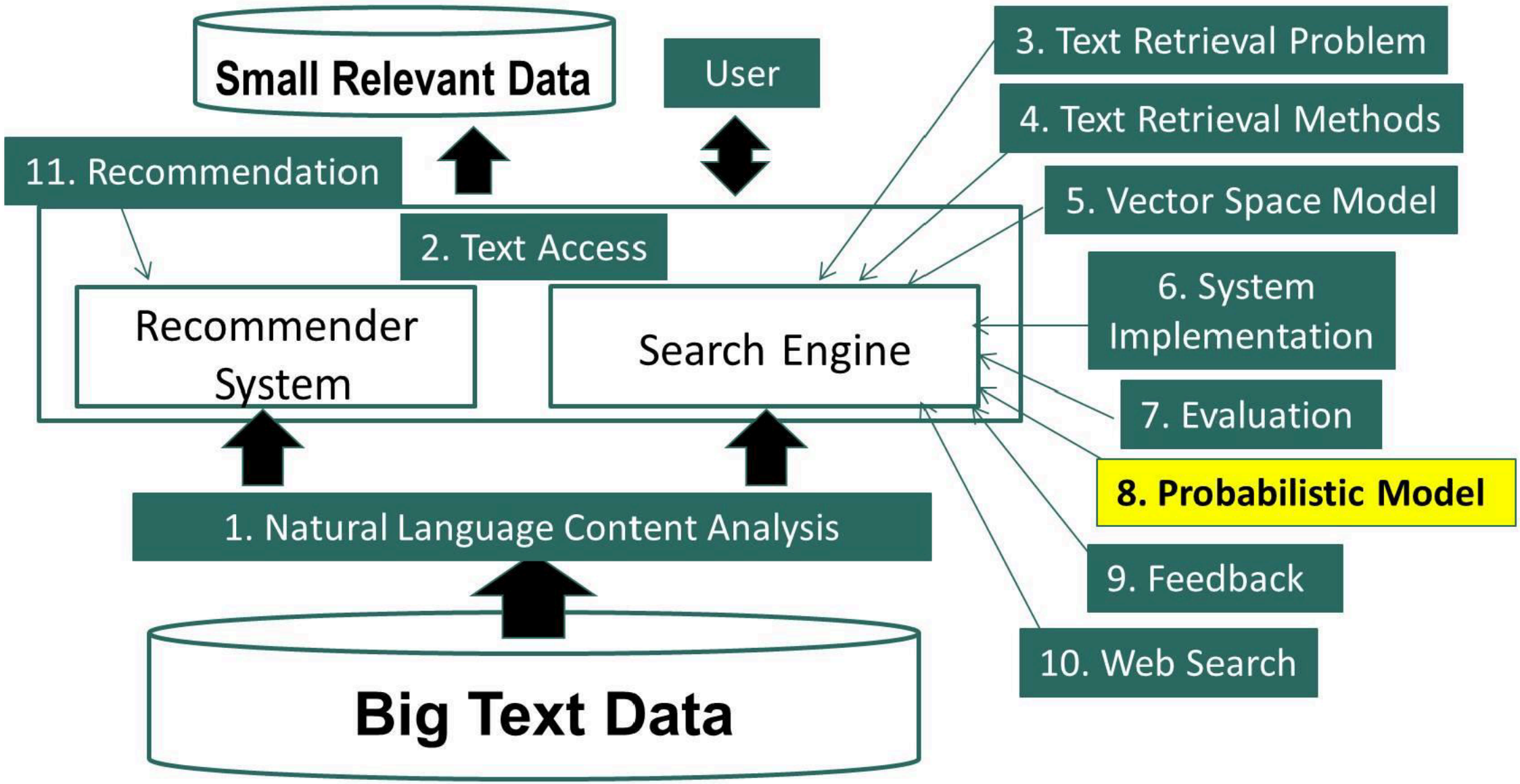
# Information Retrieval & Text Mining

## Probabilistic Retrieval Model:
## Smoothing Methods

Dr. Iqra Safder

# Probabilistic Retrieval Model: Smoothing Methods



# Query Likelihood + Smoothing with p(w|C)

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

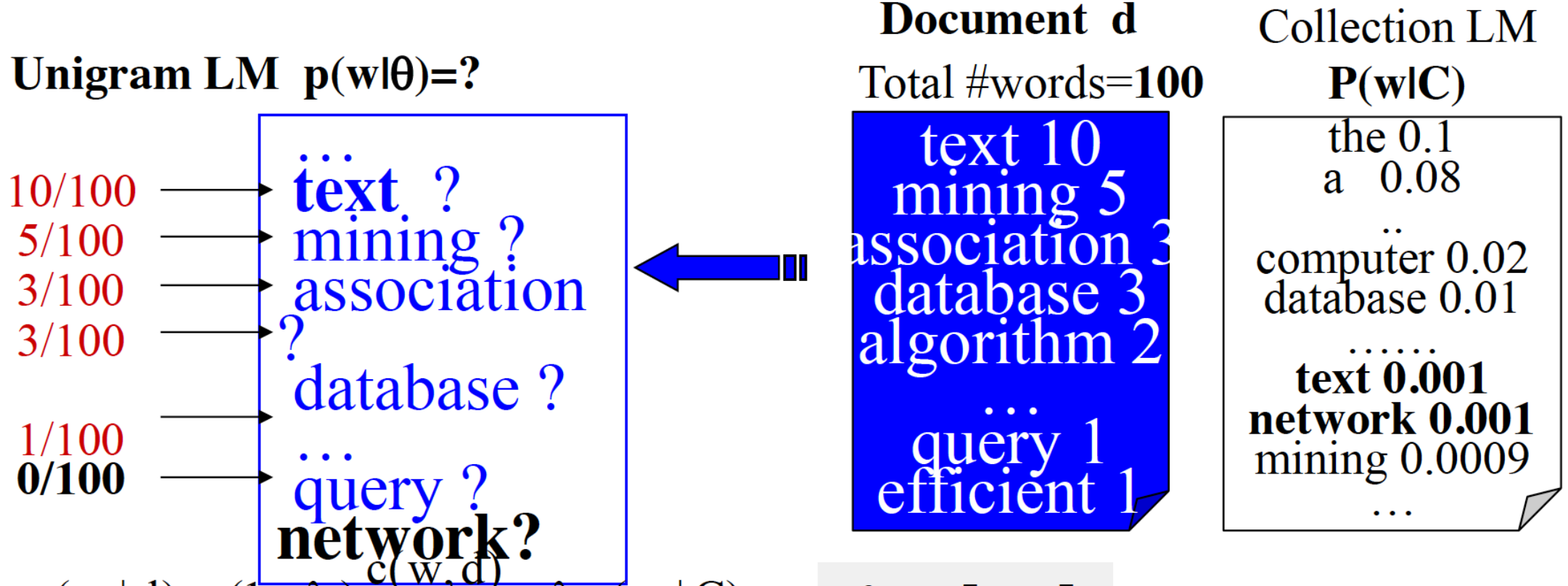
$$f(q,d) = \sum_{\substack{w_i \in d \\ w_i \in q}} c(w,q)[\log \frac{p_{Seen}(w_i \mid d)}{\alpha_d \, p(w_i \mid C)}] + n \log \alpha_d$$

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

$$\alpha_d = ?$$

How to smooth p(w|d)?

# Linear Interpolation (Jelinek-Mercer) Smoothing



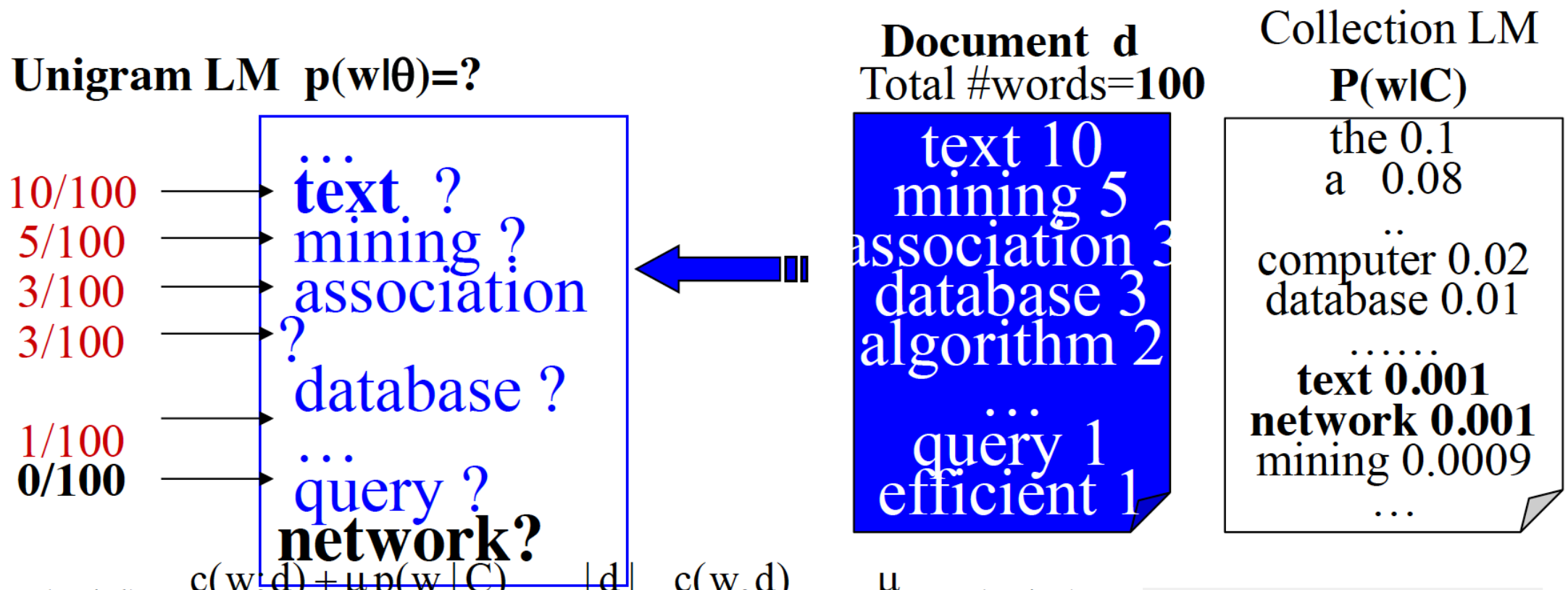

$$p(w \mid d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda\, p(w \mid C)$$

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

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

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

# Dirichlet Prior (Bayesian) Smoothing




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

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

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

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

# 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