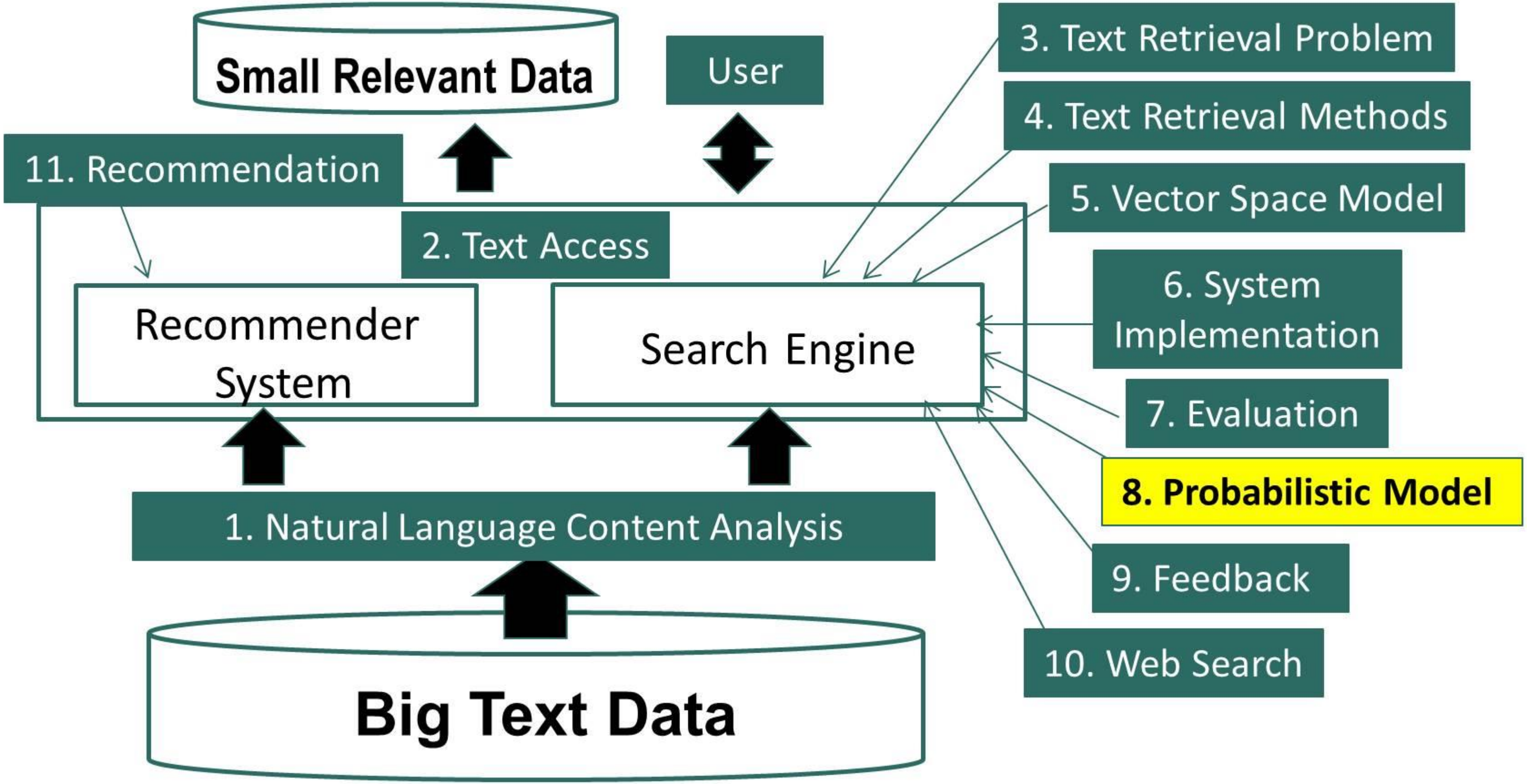
Text Retrieval and Search Engines

Probabilistic Retrieval Model: Smoothing Methods Part 1 & 2

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)

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

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

$$\boxed{\begin{array}{l} p_{Seen}(w_i|d) = ? \\ \alpha_d = ? \end{array}}$$

How to smooth p(w|d)?

Linear Interpolation (Jelinek-Mercer) Smoothing

Unigram LM p(w|θ)=?

10/100 → **text** ?
5/100 → mining ?
3/100 → association ?
3/100 → database ?
1/100 → …
0/100 → query ?
network?

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
…

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

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

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

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

Dirichlet Prior (Bayesian) Smoothing

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

…
10/100 → **text** ?
5/100 → mining ?
3/100 → association ?
3/100 → database ?
…
1/100 → query ?
0/100 → **network?**

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
…

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

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

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

Ranking Function for JM Smoothing

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

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

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

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

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

Ranking Function for Dirichlet Prior Smoothing

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

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

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

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

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

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) ≈ p(q|d,R=1) ≈**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