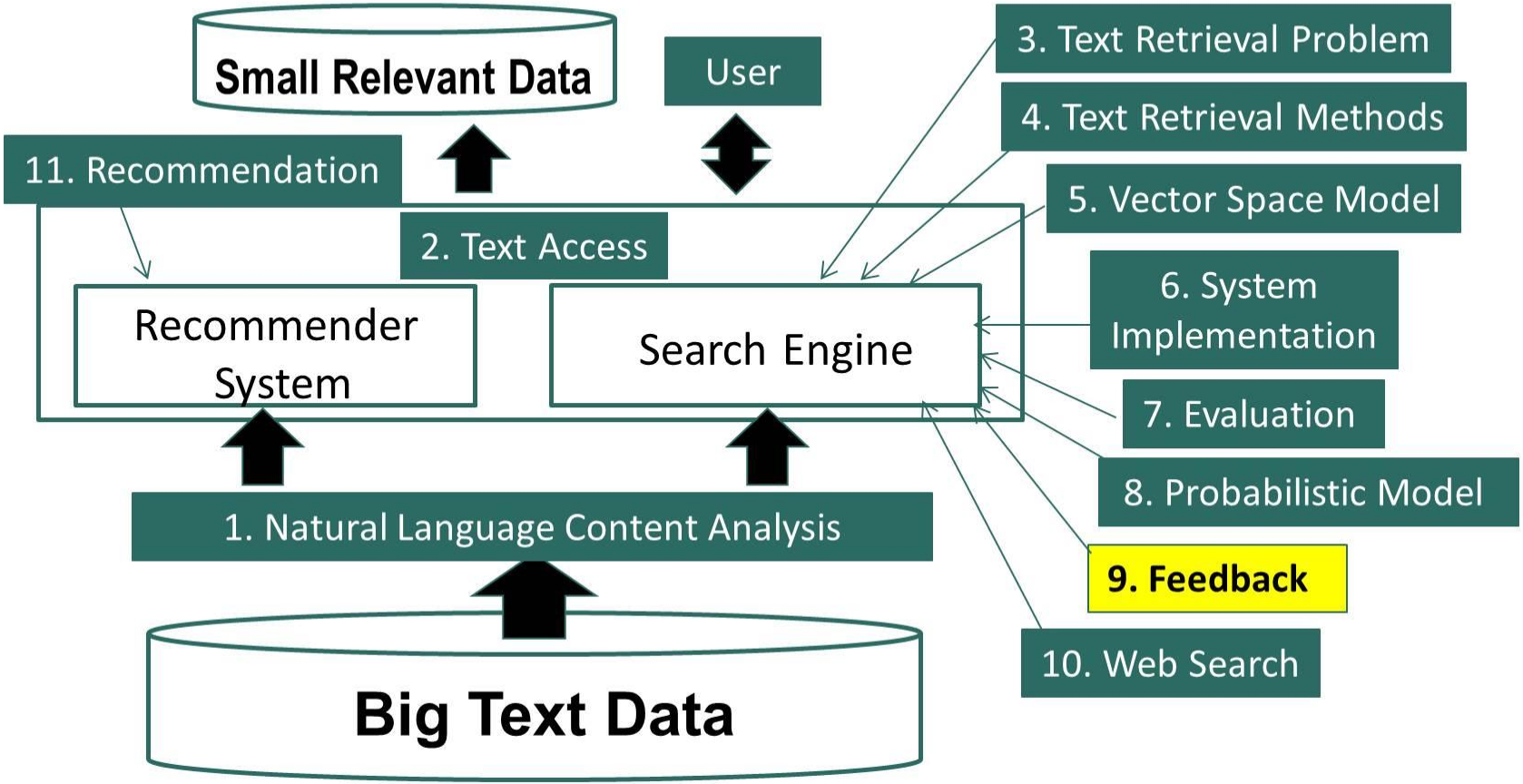
Text Retrieval and Search Engines

Feedback in Text Retrieval: Feedback in LM

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

Feedback in Text Retrieval: Feedback in LM



Feedback with Language Models

- Query likelihood method can't naturally support relevance feedback
- Solution:
  - Kullback-Leibler (KL) divergence retrieval model as a generalization of query likelihood
  - Feedback is achieved through query model estimation/updating

Kullback-Leibler (KL) Divergence Retrieval Model

Query Likelihood

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

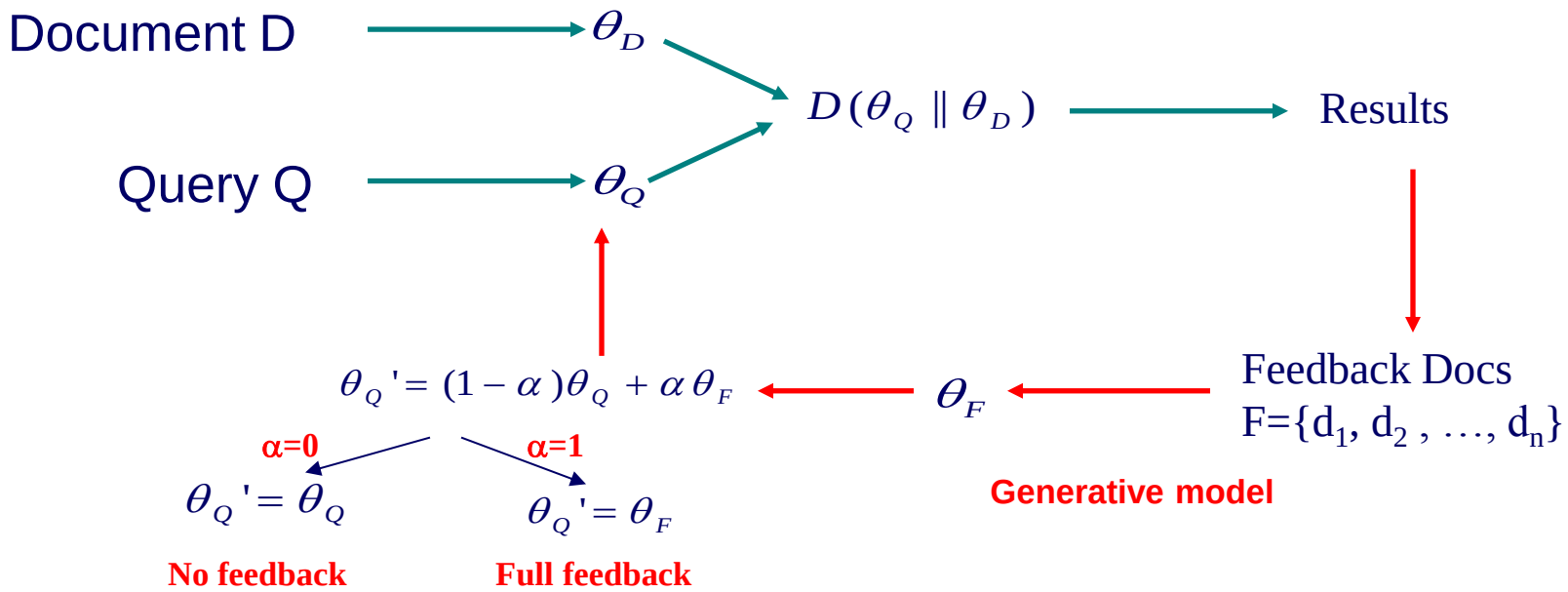
KL-divergence (cross entropy)

$$f(q,d) = \sum_{w \in d, p(w|\theta_Q)>0} [p(w \mid \hat{\theta}_Q)] \log \frac{p_{seen}(w \mid d)}{\alpha_d p(w \mid C)}] + \log \alpha_d$$

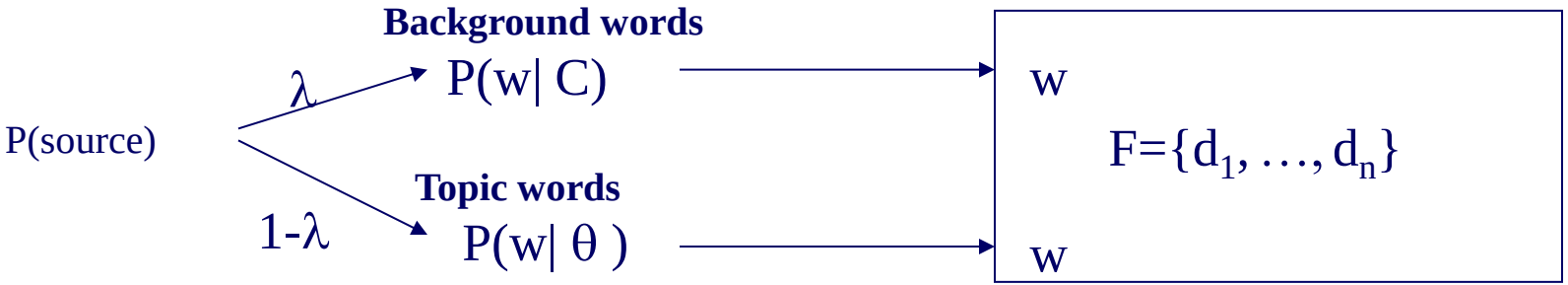
Query LM

$$p(w \mid \hat{\theta}_Q) = \frac{c(w,Q)}{|Q|}$$

Feedback as Model Interpolation



Generative Mixture Model



$$\log p(F \mid \theta) = \sum_i \sum_w c(w; d_i) \log[(1-\lambda)\, p(w \mid \theta) + \lambda\, p(w \mid C)]$$

Maximum Likelihood $\theta_F = \arg\max_{\theta} \log p(F \mid \theta)$

λ = Noise in feedback documents

Example of Pseudo-Feedback Query Model

Query: "airport security"

Mixture model approach

Web database

Top 10 docs

λ=0.9

W	$p(W|\theta_F)$
security	0.0558
airport	0.0546
beverage	0.0488
alcohol	0.0474
bomb	0.0236
terrorist	0.0217
author	0.0206
license	0.0188
bond	0.0186
counter-terror	0.0173
terror	0.0142
newsnet	0.0129
attack	0.0124
operation	0.0121
headline	0.0121

λ=0.7

W	$p(W|\theta_F)$
the	0.0405
security	0.0377
airport	0.0342
beverage	0.0305
alcohol	0.0304
to	0.0268
of	0.0241
and	0.0214
author	0.0156
bomb	0.0150
terrorist	0.0137
in	0.0135
license	0.0127
state	0.0127
by	0.0125

Summary of Feedback in Text Retrieval

- Feedback = learn from examples
- Three major feedback scenarios
  - Relevance, pseudo, and implicit feedback
- Rocchio for VSM
- Query model estimation for LM
  - Mixture model
  - Many other methods (e.g., relevance model) have been proposed [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

- Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of ACM SIGIR 2011.*