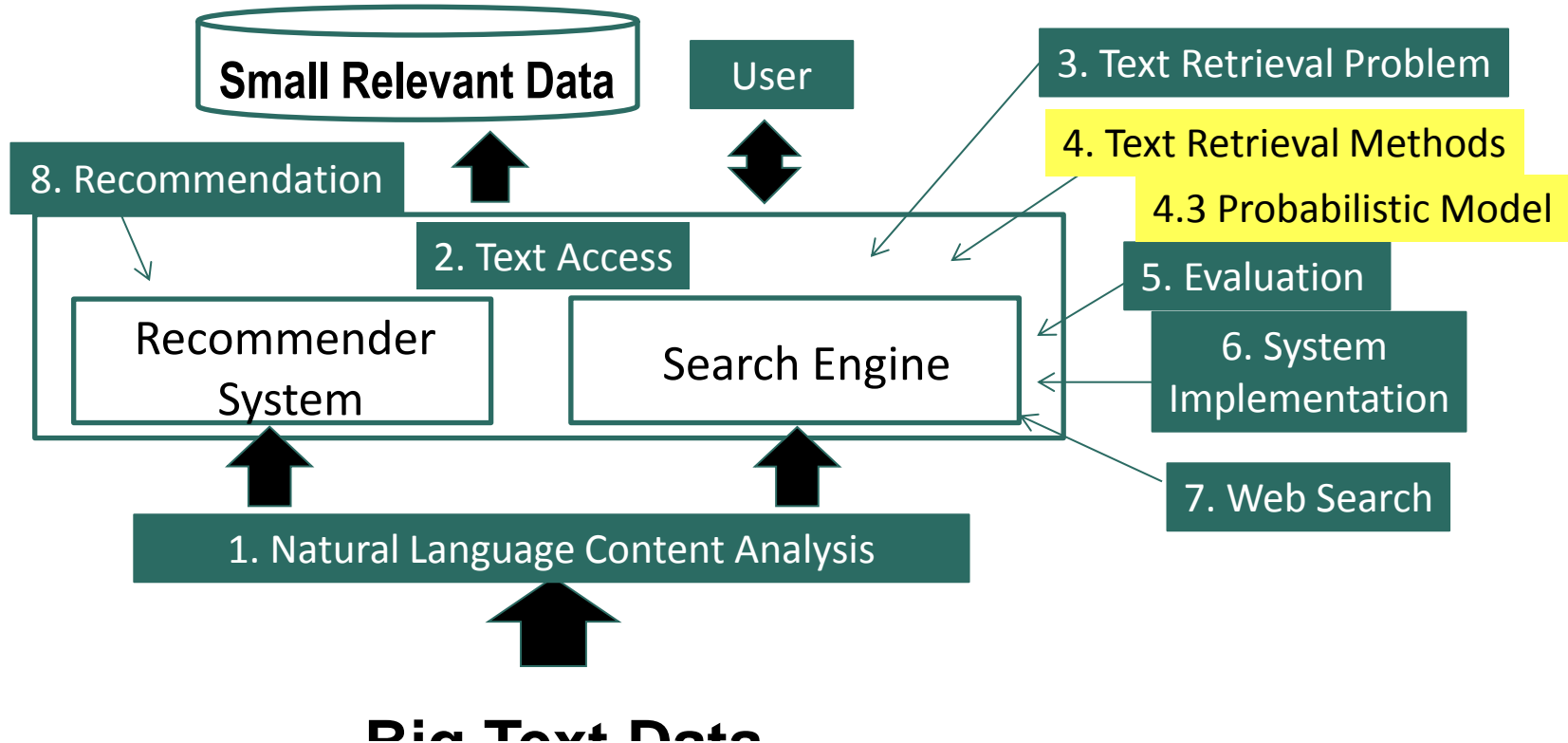


Text Retrieval and Search Engines

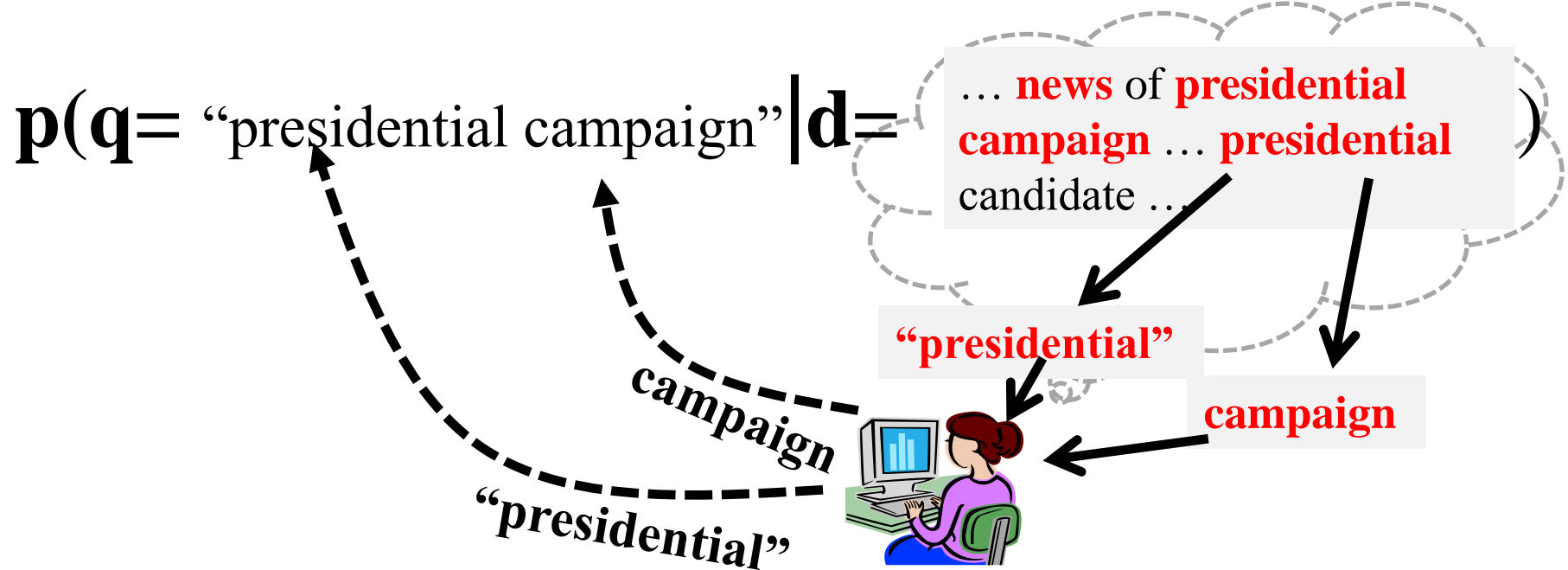
Probabilistic Retrieval Model: Query Likelihood

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

Probabilistic Retrieval Model: Query Likelihood



Query Generation by Sampling Words from Doc



If the user is **thinking of this doc** ,
how likely would she **pose this query**?

Unigram Query Likelihood

$$\begin{aligned} p(q = \text{"presidential campaign"} | d = \text{... news of presidential campaign ... presidential candidate ...}) \\ = p(\text{"presidential"} | d) * p(\text{"campaign"} | d) \\ = \frac{c(\text{"presidential"}, d)}{|d|} * \frac{c(\text{"campaign"}, d)}{|d|} \end{aligned}$$

Assumption:

Each query word is generated independently

Does Query Likelihood Make Sense?

$$p(q = \text{"presidential campaign"} | d) = \frac{c(\text{"presidential"}, d)}{|d|} * \frac{c(\text{"campaign"}, d)}{|d|}$$

$$p(q|d4 = \text{... news of presidential campaign ... presidential candidate ...}) = \frac{2}{|d4|} * \frac{1}{|d4|}$$

$$p(q|d3 = \text{... news of presidential campaign ...}) = \frac{1}{|d3|} * \frac{1}{|d3|}$$

$$p(q|d2 = \text{... news about organic food campaign...}) = \frac{0}{|d2|} * \frac{1}{|d2|} = 0$$

d4 > d3 > d2 as we expected

Try a Different Query?

q = “**presidential campaign** **update**”

$$p(q|d4 = \text{... news of } \textbf{presidential campaign} \text{ ... } \textbf{presidential} \text{ candidate ...}) = \frac{2}{|d4|} * \frac{1}{|d4|} * \frac{0}{|d4|} = 0!$$

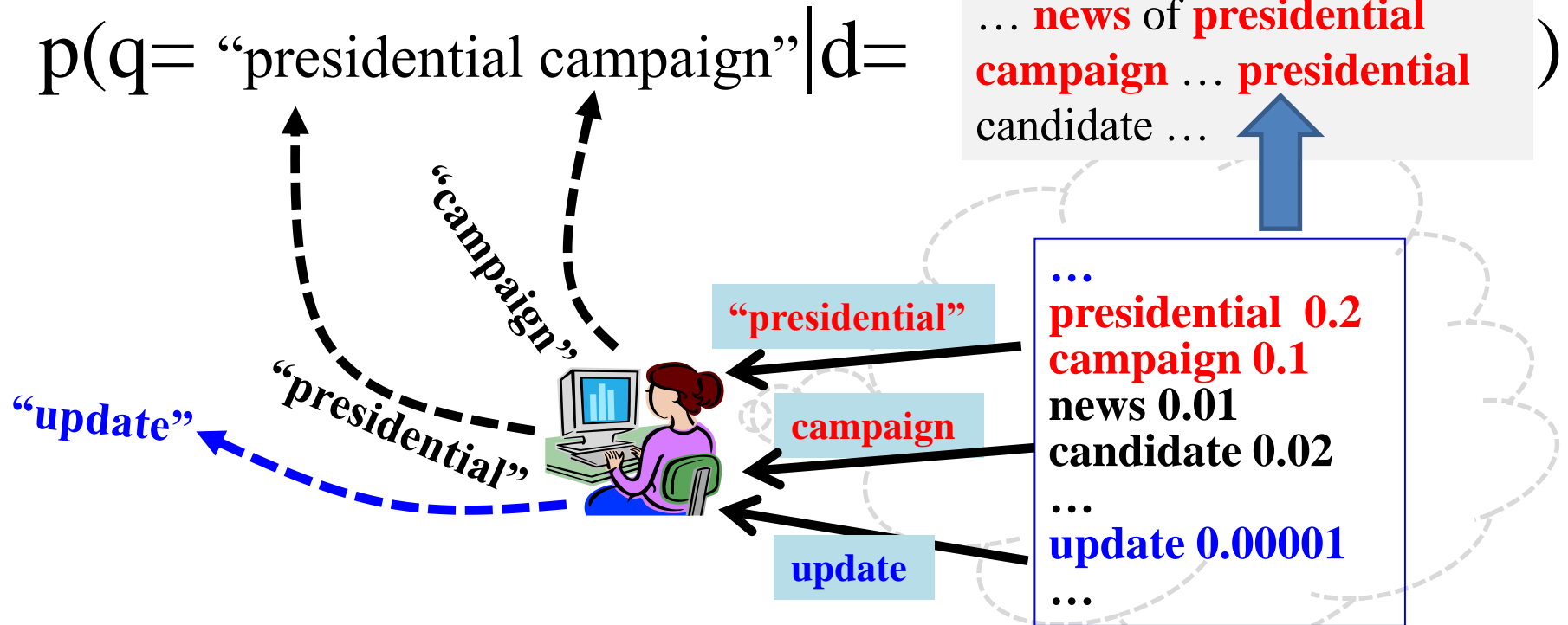
$$p(q|d3 = \text{... news of } \textbf{presidential campaign} \text{ ...}) = \frac{1}{|d3|} * \frac{1}{|d3|} * \frac{0}{|d3|} = 0!$$

$$p(q|d2 = \text{... news about organic food } \textbf{campaign} \text{ ...}) = \frac{0}{|d2|} * \frac{1}{|d2|} * \frac{0}{|d2|} = 0$$

What assumption has caused this problem? How do we fix it?

Improved Model: Sampling Words from a Doc Model

How likely would we observe **this query** from **this doc model**?



Computation of Query Likelihood

Document

d1

**Text mining
paper**



Document LM

$p(w|d1)$

...
text 0.2
mining 0.1
association 0.01
clustering 0.02
...
food 0.00001
...

d2

**Food nutrition
paper**



$p(w|d2)$

...
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
...

Query q =

“data mining algorithms”

$$\begin{aligned} p(\text{“data mining alg”}|d1) \\ &= p(\text{“data”}|d1) \\ &\quad \times p(\text{“mining”}|d1) \\ &\quad \times p(\text{“alg”}|d1) \end{aligned}$$

$$\begin{aligned} p(\text{“data mining alg”}|d2) \\ &= p(\text{“data”}|d2) \\ &\quad \times p(\text{“mining”}|d2) \\ &\quad \times p(\text{“alg”}|d2) \end{aligned}$$

Summary: Ranking based on Query Likelihood

$$q = w_1 w_2 \dots w_n \quad p(q | d) = p(w_1 | d) \times \dots \times p(w_n | d)$$

$$f(q, d) = \log p(q | d) = \sum_{i=1}^n \log p(w_i | d) = \sum_{w \in V} c(w, q) \log p(w | d)$$

Document language model

Retrieval problem → Estimation of $p(w_i | d)$

Different estimation methods → different ranking functions