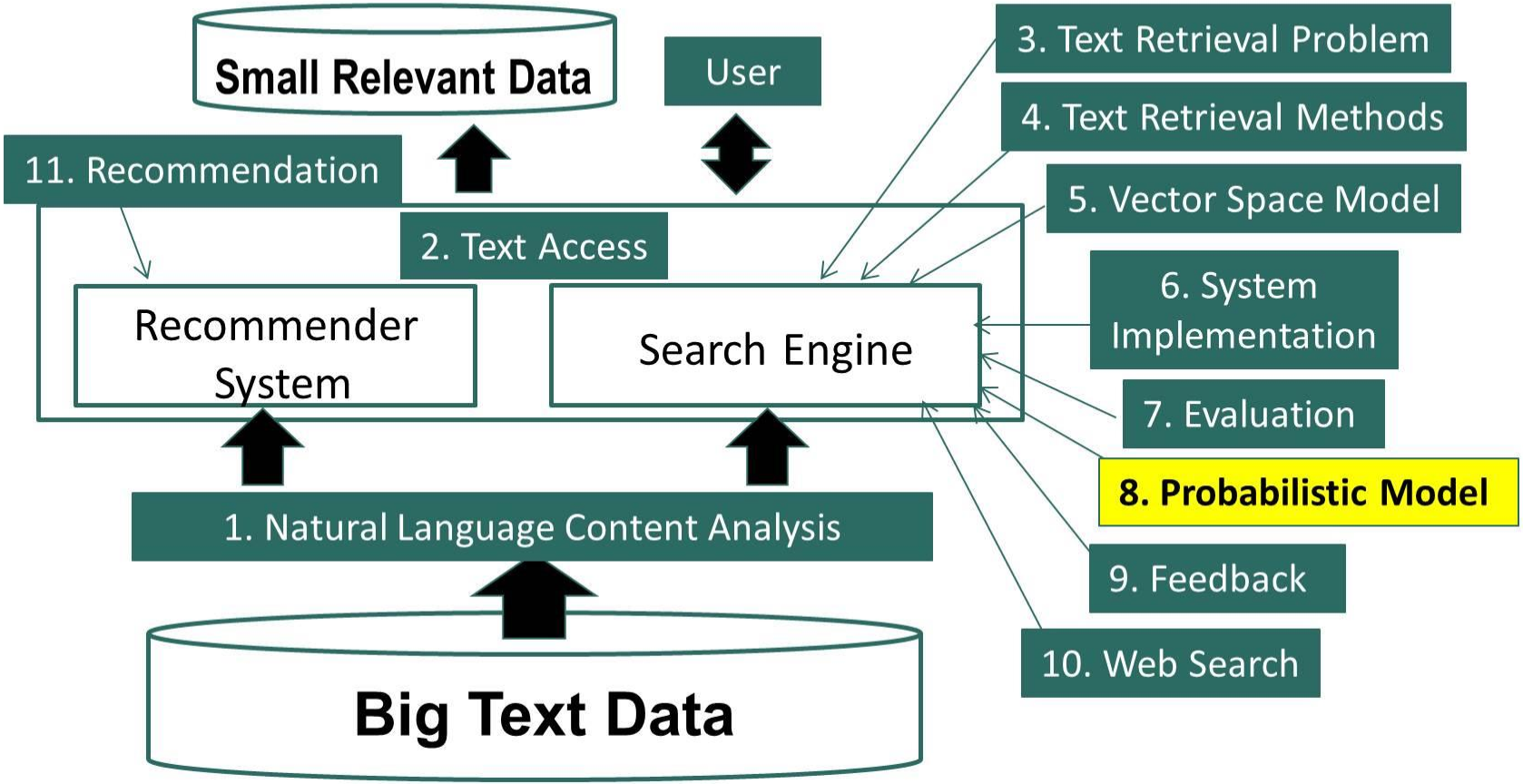
Text Retrieval and Search Engines

Probabilistic Retrieval Model: Basic Idea

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

Probabilistic Retrieval Model: Basic Idea



Many Different Retrieval Models

- **Probabilistic models**: $f(d,q) = p(R=1|d,q), \quad R \in \{0,1\}$
  - Classic probabilistic model ➔ BM25
  - **Language model ➔ Query Likelihood**
  - Divergence-from-randomness model ➔ PL2

$$p(R=1|d,q) \approx p(q|d,R=1)$$

If a user likes document d, how likely would the user enter query q (in order to retrieve d)?

Probabilistic Retrieval Models: Basic Idea

Query	Doc	Rel
q	**d**	**R**
q1	d1	1
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
…		
q1	d1	0
q1	d2	1
q1	d3	0
q2	d3	1
q3	d1	1
q4	d2	1
q4	d3	0

f(q,d)=p(R=1|d,q)=? $\frac{count(q,d,R=1)}{count(q,d)}$

P(R=1|q1,d1) = ? 1/2

P(R=1|q1,d2) = ? 2/2

P(R=1|q1,d3) = ? 0/2

What about unseen documents?

Unseen queries?

Query Likelihood Retrieval Model

Query	Doc	Rel
q	**d**	**R**
q1	d1	1
q1	d2	1
q1	d3	0
q1	d4	0
q1	d5	1
…		
q1	d1	0
q1	d2	1
q1	d3	0
q2	d3	1
q3	d1	1
q4	d2	1

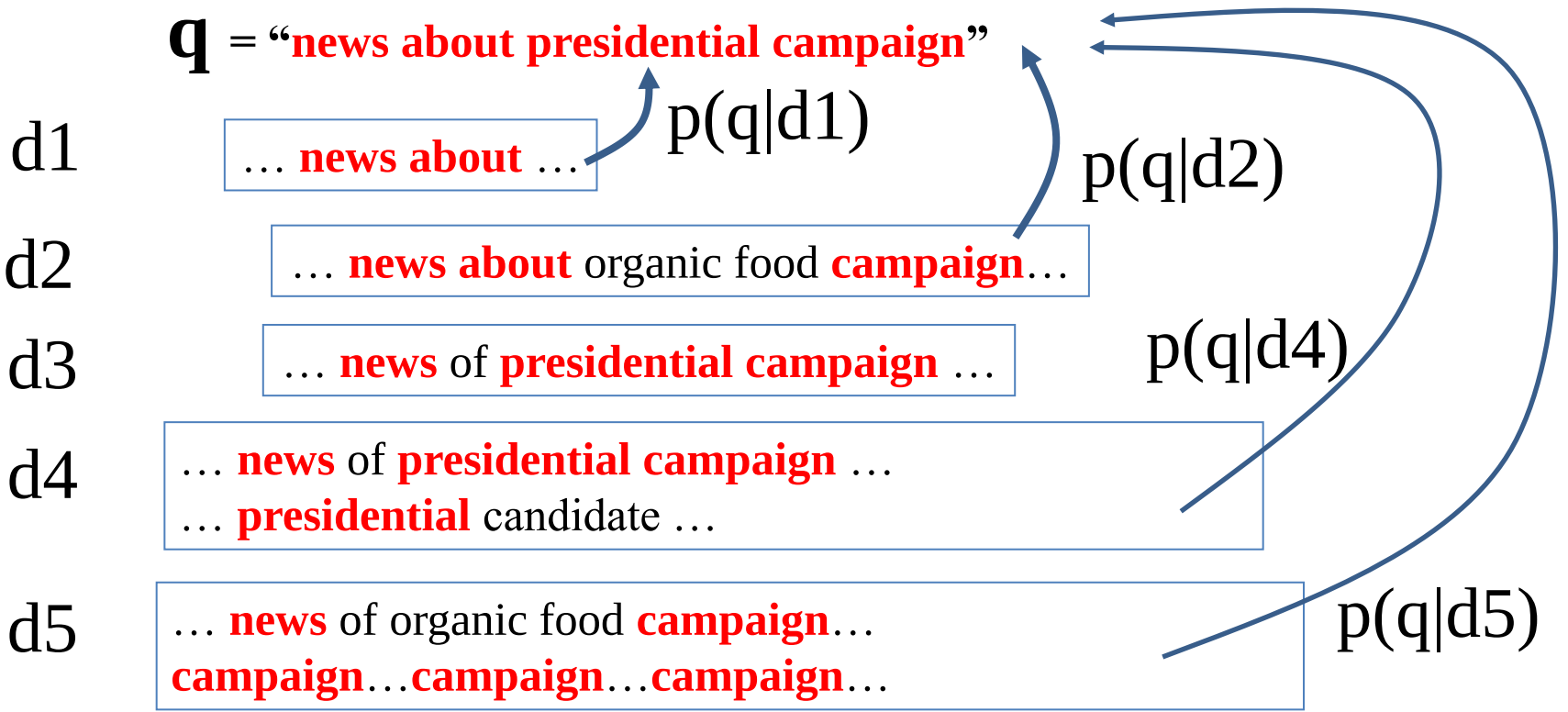
User likes d

$$f(q,d)=p(R=1|d,q)\approx \mathbf{p(q|d,R=1)}$$

How likely the user enters q

Assumption:
A user formulates a query based on an "**imaginary relevant document**"

Which doc is Most Likely the "Imaginary Relevant Doc"?



Summary

- Relevance(q,d) = p(R=1|q,d) ➔ p(q|d,R=1)
- **Query likelihood** ranking function: f(q,d)=p(q|d)
  - Probability that a user who likes d would pose query q
- How to compute p(q|d)? How to compute probability of text in general? ➔ Language Model

p(q= "presidential campaign"**|d=** … **news** of **presidential campaign** … **presidential** candidate … **)**