

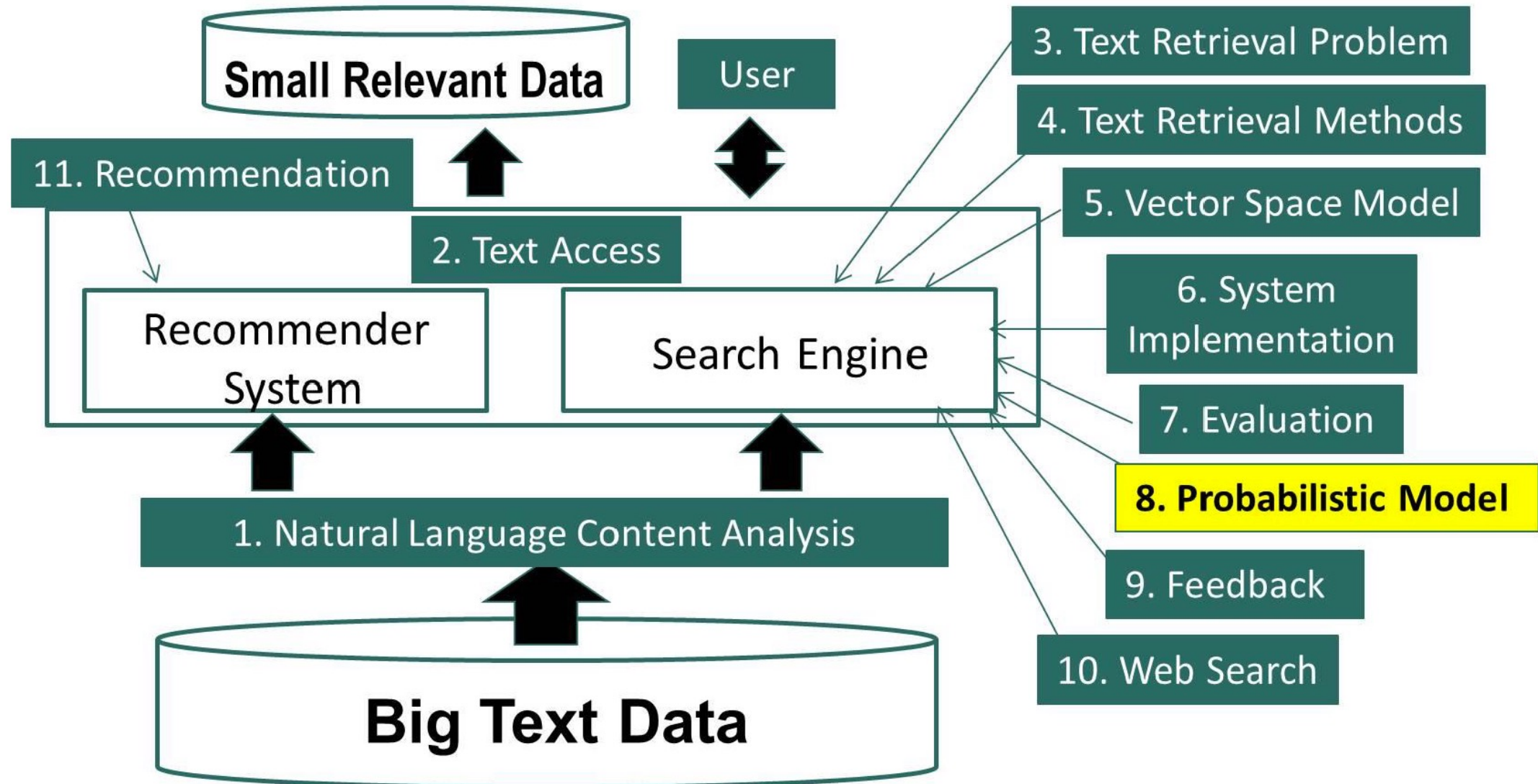


Information Retrieval

Probabilistic Retrieval Model: Basic Idea

Dr. Iqra Safder

Probabilistic Retrieval Model: Basic Idea



Many Different Retrieval Models

- **Similarity-based models:** $f(q,d) = \text{similarity}(q,d)$
 - Vector space model
- **Probabilistic models:** $f(d,q) = p(R=1 \mid d,q)$, where $R \in \{0,1\}$
 - Classic probabilistic model
 - Language model
 - Divergence-from-randomness model
- **Probabilistic inference model:** $f(q,d) = p(d \rightarrow q)$
- **Axiomatic model:** $f(q,d)$ must satisfy a set of constraints
- These different models tend to result in similar ranking functions involving similar variables

Probabilistic Model

- We define ranking function that a given document **D** is relevant to a given query **Q**.
- We introduce binary random variable $R \in \{0, 1\}$
- We assume that **Q** and **D** are observations from random variable, in vector space model we assume they are vectors
- Problem of retrieval now becomes the problem to estimate the probability of relevance.

$$f(d,q) = p(R=1 | d,q), \quad R \in \{0,1\}$$

Many Different Retrieval Models

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

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

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
----	----	---

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
----	----	---

$$f(q,d)=p(R=1 \mid d,q)=?$$

How can we estimate the probability of relevance?

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
----	----	---

$f(q,d)=p(R=1 | d,q)=?$

$$\frac{\text{count}(q, d, R = 1)}{\text{count}(q, d)}$$

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
----	----	---

$f(q,d)=p(R=1 \mid d,q)=?$

$$\frac{\text{count}(q, d, R = 1)}{\text{count}(q, d)}$$

$$P(R=1 \mid q1, d1) = ?$$

$$P(R=1 \mid q1, d2) = ?$$

$$P(R=1 \mid q1, d3) = ?$$

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
----	----	---

$$f(q,d)=p(R=1 \mid d,q)=?$$

$$\frac{\text{count}(q, d, R = 1)}{\text{count}(q, d)}$$

$$P(R=1 \mid q1, d1) = ? \quad 1/2$$

$$P(R=1 \mid q1, d2) = ? \quad 2/2$$

$$P(R=1 \mid q1, d3) = ? \quad 0/2$$

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

$$f(q,d)=p(R=1 \mid d,q)=?$$

$$\frac{\text{count}(q, d, R = 1)}{\text{count}(q, d)}$$

$$P(R=1 \mid q1, d1) = ? \quad 1/2$$

$$P(R=1 \mid q1, d2) = ? \quad 2/2$$

$$P(R=1 \mid q1, d3) = ? \quad 0/2$$

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
----	----	---

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

Approximations

In query likelihood, our assumption is that this probability of relevance can be approximated by the probability of a query given a document and relevance, $p(q \mid d, R = 1)$. Intuitively, this probability just captures the following probability: if a user likes document d , how likely would the user enter query q in order to retrieve document d ? The condition part contains document d and $R = 1$, which can be interpreted as the condition that the user likes document d .

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

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

User likes d

$$p(q \mid d, R=1)$$

In query likelihood, our assumption is that this probability of relevance can be approximated by the probability of a query given a document and relevance, $p(q \mid d, R = 1)$. Intuitively, this probability just captures the following probability: if a user likes document d , how likely would the user enter query q in order to retrieve document d ? The condition part contains document d and $R = 1$, which can be interpreted as the condition that the user likes document d .

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

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

User likes d

$$p(q \mid d, R=1)$$

How likely the user enters q

In query likelihood, our assumption is that this probability of relevance can be approximated by the probability of a query given a document and relevance, $p(q \mid d, R = 1)$. Intuitively, this probability just captures the following probability: if a user likes document d , how likely would the user enter query q in order to retrieve document d ? The condition part contains document d and $R = 1$, which can be interpreted as the condition that the user likes document d .

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

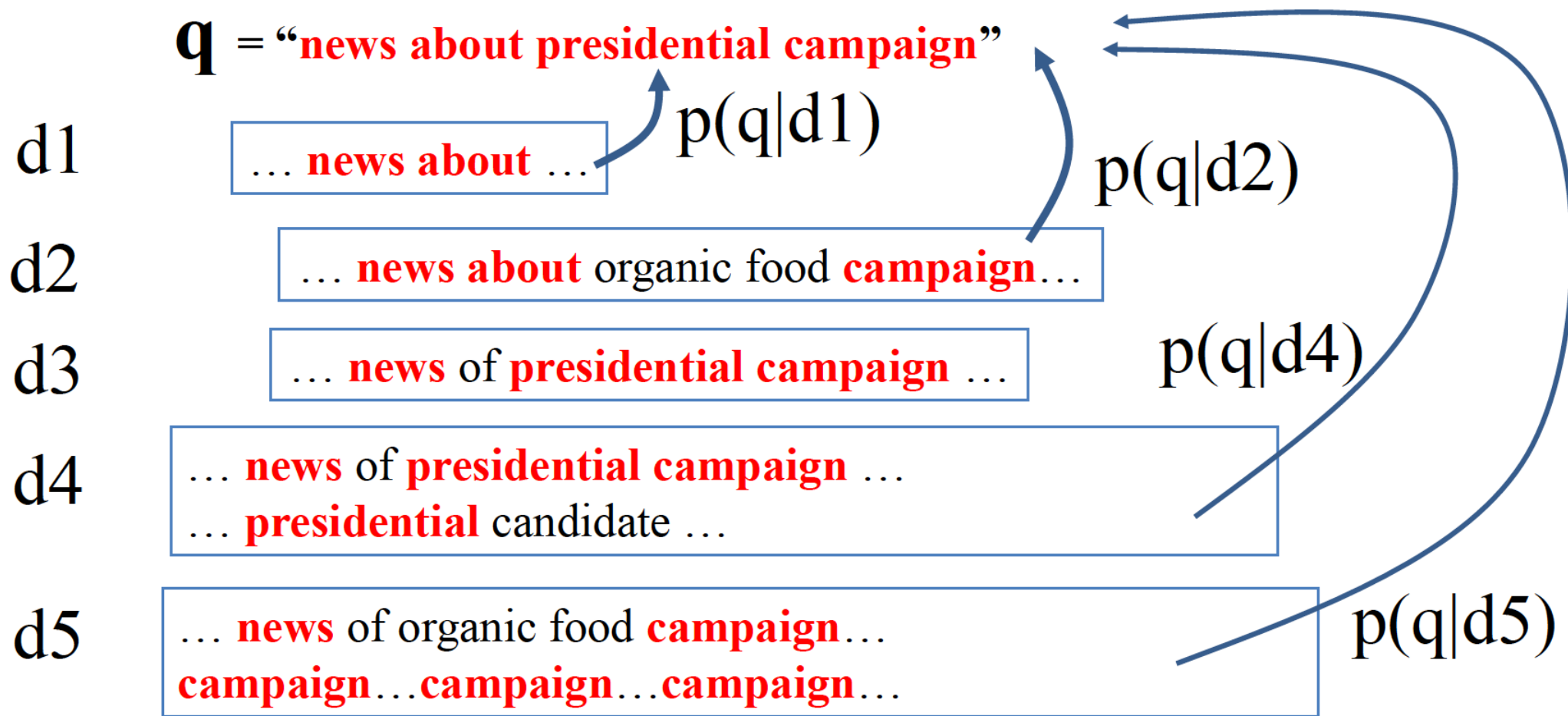
$$f(q,d)=p(R=1 \mid d,q) \approx p(q \mid d, R=1)$$

How likely the user enters q

User likes d

Assumption:
A user formulates a query based on an
“imaginary relevant document”

Which doc is Most Likely the “Imaginary Relevant Doc”?



Summary

- $\text{Relevance}(q,d) = p(R=1 | q,d) \rightarrow p(q | d, R=1)$
- **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? \rightarrow Language Model

$p(q = \text{"presidential campaign"} | d =$

... news of presidential
campaign ... presidential
candidate ...)