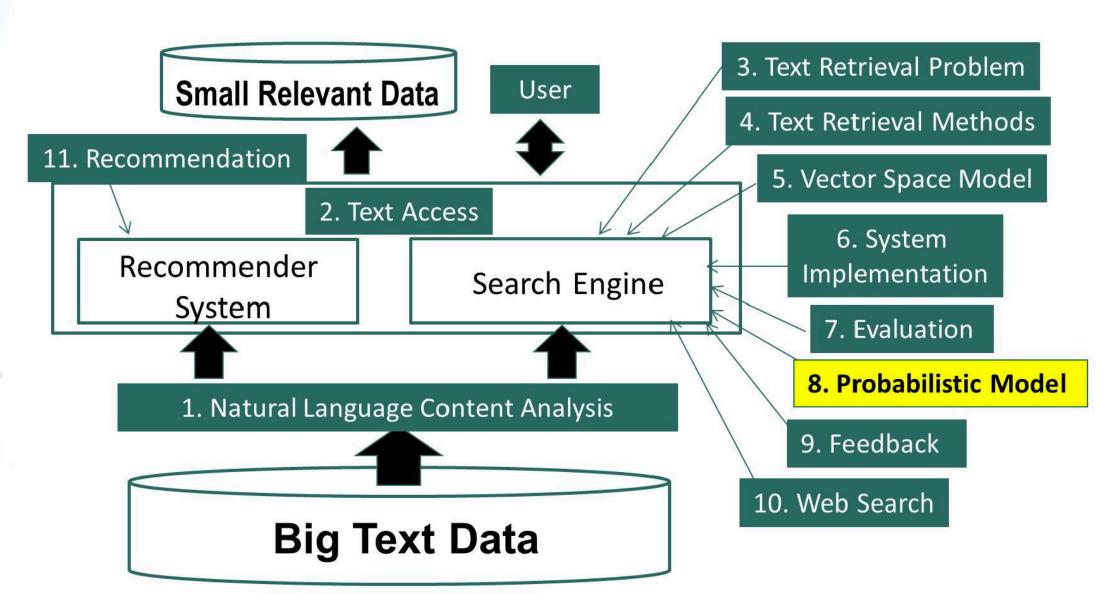
Information Retrieval & Text Mining

Probabilistic Retrieval Model:

Basic Idea

Dr. Saeed UI Hassan
Information Technology University



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.

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

Many Different Retrieval Models

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

Query	Do	C	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	

Query	Do	2	Rel	
q	d		R	
q1	d1	1		f(q,d)=p(R=1 d,q)=?
q1	d2	1		(-1//-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		

Query	Do	С	Rel		
q	d		R		count(q,d,R)
q1	d1	1		f(q,d)=p(R=1 d,q)=?	
ql	d2	1			count(q, d
q1	d3	0			
ql	d4	0			
q1	d5	1			
ql	d1	0			
ql	d2	1			
q1	d3	0			
q2	d3	1			

d1

d2

Query	Do	c Re	<u> </u>	
q	d	R		count(q, d, R)
q1	d1	1	f(q,d)=p(R=1 d,q)=?	
ql	d2	1		count(q, a
ql	d3	0		
ql	d4	0	P(R=1 q1,d1) = ?	
q1	d5	1		
			P(R=1 q1,d2) = ?	
q1	d1	0	P(R=1 q1,d3) = ?	
q1	d2	1	. (= 9=,00,	
ql	d3	0		
q2	d3	1		
q3	d1	1		
q4	d2	1		

Query	Do	C	Rel		
q	d		R		count(q, d, R = 1)
q1	d1	1		f(q,d)=p(R=1 d,q)=?	
ql	d2	1		(1) / 1 (1 / 1)	count(q,d)
ql	d3	0			
q1	d4	0		D/D = 1 a1 d1 = 2	1/2
q1	d5	1		P(R=1 q1,d1) = ?	•
				P(R=1 q1,d2) = ?	2/2
ql	d1	0		P(R=1 q1,d3) = ?	0/2
ql	d2	1		. (= 9=,00,	•
q1	d3	0			
q2	d3	1			

d1

d2

q3

Query	Do	С	Rel		
q	d		R		count(q, d, R = 1)
q1	d1	1		f(q,d)=p(R=1 d,q)=?	
ql	d2	1		(1/ / 1 (count(q,d)
ql	d3	0			
ql	d4	0		P(R=1 q1,d1) = ?	1/2
q1	d5	1			•
				P(R=1 q1,d2) = ?	2/2
q1	d1	0		P(R=1 q1,d3) = ?	0/2
q1	d2	1		. (= 9 =) 0.0 / .	
q1	d3	0		What about unseen	documents?
q2	d3	1		Uncoon quarios2	
q3	d1	1		Unseen queries?	
q4	d2	1			4

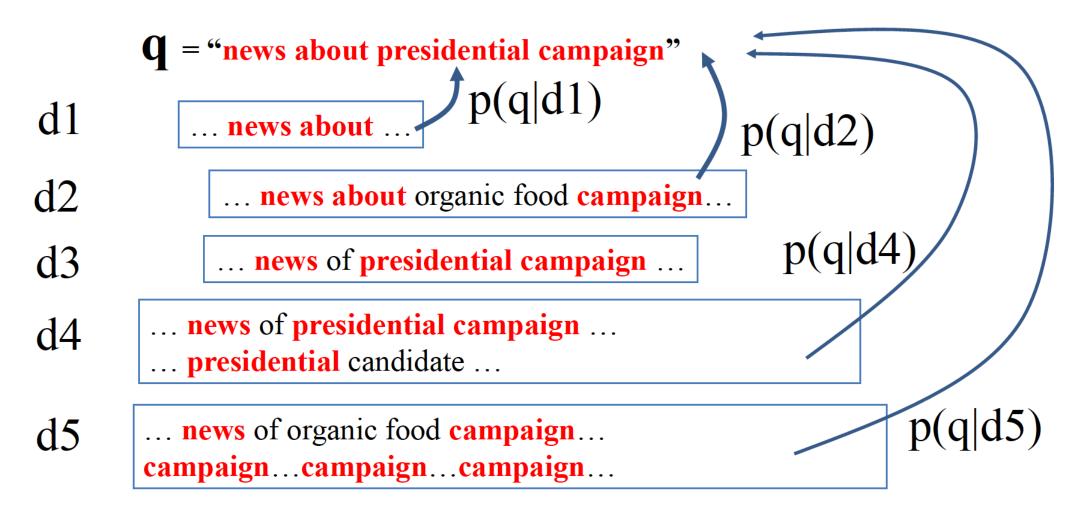
Query	Doc	Rel		
q	d	R		
q1	d1	1	$f(q,d)=p(R=1 d,q)\approx$	p(q d,R=1)
q1	d2	1	(3),37 [2(3 - 1373])	
q1	d3	0		
q1	d4	0		
q1	d5	1		
			Approximati	one
q1	d1	0	Approximati	Ulis
q1	d2	1		
q1	d3	0		
q2	d3	1		
q3	d1	1		
q4	d2	1		

Query	Do	С	Rel		User likes d
q	d		R		1 1 1
ql	d1	1		$f(q,d)=p(R=1 d,q)\approx$	p(q d,R=1)
ql	d2	1			
ql	d3	0			
q1	d4	0			
q1	d5	1			
ql	d1	0			
ql	d2	1			
ql	d3	0			
q2	d3	1			
q3	d1	1			
q4	d2	1			

Query	Do	С	Rel		User likes d
q	d		R		\wedge
ql	d1	1		$f(q,d)=p(R=1 d,q)\approx$	p(q d,R=1)
ql	d2	1			
ql	d3	0		سممين مطاح بالمباثل بينمانا	
ql	d4	0		How likely the user	enters q
q1	d5	1			
ql	d1	0			
ql	d2	1			
ql	d3	0			
q2	d3	1			
q3	d1	1			
q4	d2	1			

Query	Do	С	Rel	User likes d
q	d		R	1 1 1 1 1 1 1 1 1 1
q1	d1	1		$f(q,d)=p(R=1 d,q)\approx p(q d,R=1)$
ql	d2	1		
ql	d3	0		Have likely the green enters of
ql	d4	0		How likely the user enters q
ql	d5	1		
				Accumption
ql	d1	0		Assumption:
ql	d2	1		A user formulates a query based on an
ql	d3	0		"imaginary relevant document"
q2	d3	1		imaginary relevant document
q3	d1	1		
q4	d2	1		

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



Summary

- 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= "presidential campaign" | d= | ... news of presidential campaign ... presidential candidate ...
```