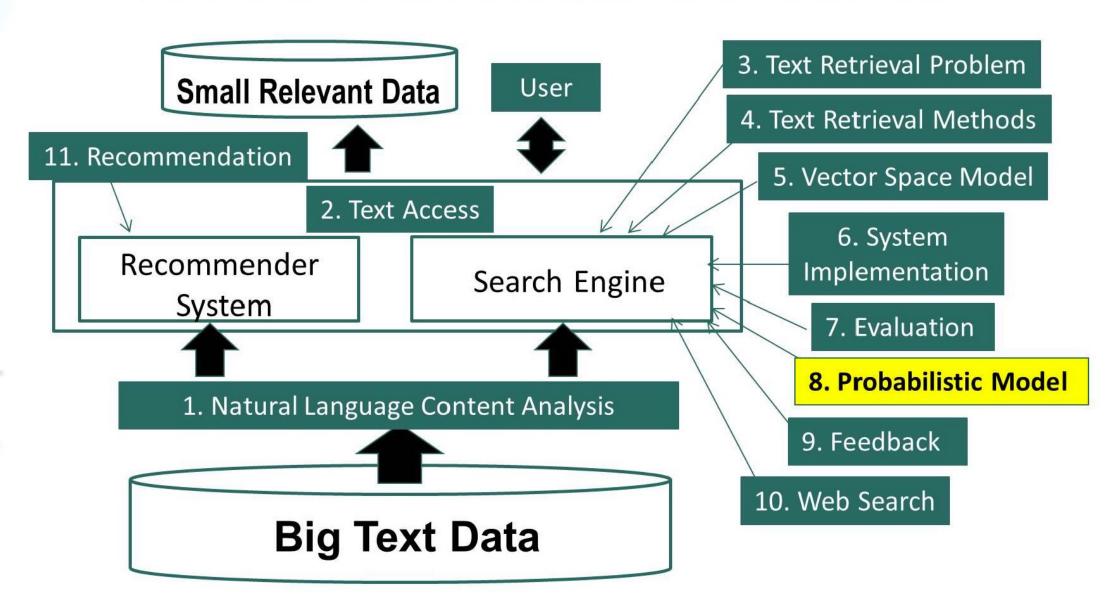
Information Retrieval

Probabilistic Retrieval Model:

Basic Idea



Many Different Retrieval Models

- **Similarity-based models**: f(q,d) = similarity(q,d)
 - Vector space model
- Probabilistic models: f(d,q) = p(R=1|d,q), where $R \in \{0,1\}$
 - Classic probabilistic model
 - Language model
 - Divergence-from-randomness model
- Probabilistic inference model: f(q,d) = p(d→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.

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

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	С	Rel
q	d		R
q1	d1	1	
q1	d2	1	
q1	d3	0	
ql	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	
ql	d1	1		f(q,d)=p(R=1 d,q)=?
q1	d2	1		
q1	d3	0		
q1	d4	0		How can we estimate the probability of relevance?
q1	d5	1		new can we commute the probability of relevance.
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 = 1)
ql	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			
ql	d5	1			
ql	d1	0			
ql	d2	1			
ql	d3	0			
q2	d3	1			
q3	d1	1			
q4	d2	1			

Query	Do	С	Rel		
q	d		R		count(q, d, R = 1)
ql	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	
q1	d5	1		P(R=1 q1,d1) = ?	
				P(R=1 q1,d2) = ?	
ql	d1	0		P(R=1 q1,d3) = ?	
q1	d2	1		. (= 9=)00)	
q1	d3	0			
q2	d3	1			
q3	d1	1			

d2

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			
ql	d4	0		D/D = 1 a1 d1 = 2	1/2
ql	d5	1		P(R=1 q1,d1) = ?	
				P(R=1 q1,d2) = ?	2/2
ql	d1	0		P(R=1 q1,d3) = ?	0/2
q1	d2	1		. (= 9 =/5.5/	
ql	d3	0			
q2	d3	1			

d1

d2

Query	Do	С	Rel		
q	d		R		count(q, d, R = 1)
ql	d1	1		f(q,d)=p(R=1 d,q)=?	
ql	d2	1		(1) / 1 (1 / 1)	count(q,d)
ql	d3	0			
ql	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
q1	d1	0		P(R=1 q1,d3) = ?	0/2
q1	d2	1		. (= 9=,00,	
q1	d3	0		What about unseen	documents?
q2	d3	1		Uncoon quarios2	
q3	d1	1		Unseen queries?	
q4	d2	1			4

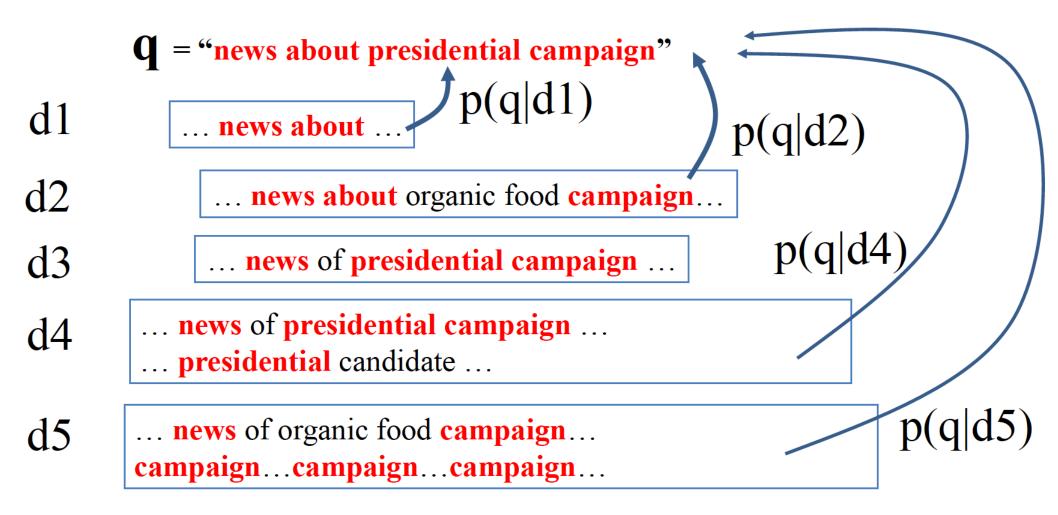
Query	Do	c Rel	
q	d	R	
q1	d1	1	$f(q,d)=p(R=1 d,q)\approx p(q d,R=1)$
q1	d2	1	
q1	d3	0	
q1	d4	0	
q1	d5	1	
			Approximations
q1	d1	0	Approximations
ql	d2	1	
q1	d3	0	In query likelihood, our assumption is that this probability of relevance can be
q2	d3	1	approximated by the probability of a query given a document and relevance, $p(q \mid d)$, $R = 1$). Intuitively, this probability just captures the following probability: if a user likes
q3	d1	1	document d, how likely would the user enter query q in order to retrieve document
q4	d2	1	d? The condition part contains document d and $R = 1$, which can be interpreted as the condition that the user likes document d.

Query	Do	c Rel	User likes d
q	d	R	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	
q1	d4	0	
q1	d5	1	
ql	d1	0	
q1	d2	1	
q1	d3	0	In query likelihood, our assumption is that this probability of relevance can be
q2	d3	1	approximated by the probability of a query given a document and relevance, $p(q \mid d)$
q3	d1	1	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
q4	d2	1	d? The condition part contains document d and R = 1, which can be interpreted as the condition that the user likes document d.

Query	Doc	Rel	User likes d
q l ql	d1 d2 d3	R 1 1 0	$f(q,d)=p(R=1 d,q)\approx p(q d,R=1)$
ql ql ql 	d4 d5	0	How likely the user enters q
q1 q1 q1 q2 q3 q4	d1 d2 d3 d3 d1 d2	0 1 0 1 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, $p(q \mid d \mid R = 1)$. Intuitively, this probability just captures the following probability: if a user like 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	Do	С	Rel	User likes d
q	d		R	^ ^
ql	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 consequent one of
ql	d4	0		How likely the user enters q
ql	d5	1		
				Accumpation
ql	d1	0		Assumption:
ql	d2	1		A user formulates a query based on an
ql	d3	0		"imaginary relevant document"
q2	d3	1		inaginary relevant accument
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? → Language Model

```
p(q= "presidential campaign" | d= | ... news of presidential | campaign ... presidential | candidate ...
```