Probabilistic Retrieval Models: Intuitions

Suppose we have a large number of relevance judgments (e.g., clickthroughs: "1"=clicked; "0"= skipped)

Query(Q)	Doc (D)	Rel (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
…		

We can score documents based on

$$P(R=1|Q,D)=\frac{count(Q,D,R=1)}{count(Q,D)}$$

P(R=1|Q1, D1)=1/2
P(R=1|Q1,D2)=2/2
P(R=1|Q1,D3)=0/2
…

What if we don't have (sufficient) search log?

We can approximate p(R=1|Q,D)!

Different assumptions lead to different models