TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 8

Information Retrieval

11. Probabilistic Information Retrieval

Warm up

Dresden Database Systems Group

Exercise 11.1

- Are the following statements true or false? Give reasons for your answer.

a) Probabilistic IR is about estimating the probability that a document is relevant to a query.

b) The odds of an event A is defined as $O(A) = P(A)/P(\bar{A})$.

c) The classical Binary Independence Retrieval (BIR) model takes term frequencies into account.

d) The BIR model assumes that terms appear independently from each other in the documents.

e) The cluster-hypothesis states, term distributions differ between relevant and irrelevant documents.

f) The system Okapi was a much simpler predecessor of the BIR Model.

Calculating the Retrieval Status Values

Dresden Database Systems Group

状态检索数值--每个term出现次数来决定权重，从而计算概率
所以首先统计出现次数
但是最后评价无关的term可以不计算 === query中-在最后评价时候不出现的（所以term不多）

-很好的题目，
和ppt上的类似
容易被忽略的薄弱环节
next

Exercise 11.2

a) A user submits the query

`tasty hot coffee sugar`

to an IR system using the BIR Model

The Relevance or Non-relevance of the documents D1-D6 is already given

Rank the documents D7 and D8 according to their Retrieval Status Values

Before you start, sketch how to approach this task: what do you need to do and what can be omitted in order to save time?

Doc	R/N	Contents
D1	N	coffee prices rose this year
D2	N	enjoy hot tasty black tea
D3	N	tasty hot tea for sale
D4	R	tasty hot coffee really tasty
D5	R	recipe hot coffee chocolate
D6	R	that great black drink
D7		coffee is more tasty than tea
D8		i need some tasty hot tea

Calculating the Retrieval Status Values

Dresden Database Systems Group

Exercise 11.2

b) Now the user submits another query.

Which of the calculations done so far have to be done anew for the new query?

nothing can be used again

c) In a) we used some initial relevance feedback. other technics

How would the system work without such initial feedback?

d) The classical BIR model does not consider term frequencies.
Is there a way to take these into account as well?

Rekursive Parameterschätzung ! 很可能问 NEXT

Doc	R/N	Contents
D1	N	coffee prices rose this year
D2	N	enjoy hot tasty black tea
D3	N	tasty hot tea for sale
D4	R	tasty hot coffee really tasty
D5	R	recipe hot coffee chocolate
D6	R	that great black drink
D7		coffee is more tasty than tea
D8		i need some tasty hot tea

Calculating the Retrieval Status Values

Dresden Database Systems Group

Exercise 11.2

ok

e) Use BM25 – Okapi to reevaluate the relevance of D7 and D8. The RSV is given as:

$$\sum_{i \in X | t_i \in q \cap d} \log \frac{N}{df(t_i)} \cdot \frac{(k_1 + 1) tf_d(t_i)}{k_1 \cdot ((1-b) + b \cdot \frac{l(d)}{l_{avg}}) + tf_d(t_i)}$$

Use $k_1 = 1.2, b = 0.5, l_{avg} = 6$ and $N = 10000$.
The global document frequencies are given as

t_i	$df(t_i)$
tasty	1000
coffee	100
hot	1000

Doc	R/N	Contents
D1	N	coffee prices rose this year
D2	N	enjoy hot tasty black tea
D3	N	tasty hot tea for sale
D4	R	tasty hot coffee really tasty
D5	R	recipe hot coffee chocolate
D6	R	that great black drink
D7		coffee is more tasty than tea
D8		i need some tasty hot tea

(In)dependent Occurrence of Terms

Dresden Database Systems Group

Exercise 11.3

- In the lecture, we formally derived the formulas for the Retrieval Status Value and the term weights
- We made the assumption that terms appear independent from each other in a document
- Thus, we could rewrite $\frac{P(\overrightarrow{d_m} \mid R \cap q_k)}{P(\overrightarrow{d_m} \mid \bar{R} \cap q_k)}$ as $\prod_{i=1}^{x} \frac{P(w_{i,m} \mid R, q_k)}{P(w_{i,m} \mid \bar{R}, q_k)}$ (see lecture slide 31)

- Provide an example of two terms for which this assumption is
  a) not likely to hold
  b) likely to hold `an example that two words are independent:: stop words`