TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 7

Information Retrieval

10. Relevance Feedback and Query Expansion

Warm up

Dresden Database Systems Group

Exercise 10.1

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

a) The main motivation for relevance feedback and query expansion is to decrease recall. ok

b) The basic idea of the Rocchio algorithm is to move the query vector towards the vectors of relevant documents and away from the vectors of irrelevant documents.

c) Relevance feedback is widely available in web search engines, because users can easily understand it and are willing to “tune” their query.

d) Pseudo-relevance feedback can lead to query drift.

e) In thesaurus-based query expansion, we add terms to a query which are semantically related to the query terms specified by the user.

f) Thesauri can only be built manually, there is no way to build them automatically.

Rocchio Relevance Feedback

Dresden Database Systems Group

Exercise 10.2

- Suppose that a user's initial query is `tasty hot chocolate`
- The user examines the following documents and marks some of them as Relevant or Non-relevant

DocId	Mark	Content
1	R	`milk chocolate hot milk chocolate`
2	R	`hot chocolate milk hot milk chocolate hot`
3	N	`tasty hot tea tasty hot tea milk`
4		`tasty milk chocolate milk`

- Assume that we use direct term frequency to build the vectors
  → no scaling, no document frequency and no need to length-normalize vectors
- Assume $\alpha = 1$, $\beta = 0.75$, and $\gamma = 0.25$
- Using Rocchio relevance feedback, what is the revised query vector be after relevance feedback?

用p27公式，小于零忽略

Some More Questions on Relevance Feedback

Dresden Database Systems Group

Exercise 10.3

NEXT 考公式

- Under what conditions would the modified query q_m in Rocchio relevance feedback be the same as the original query q_0?
- In the cases when q_m is different from q_0, is q_m always closer than q_0 to the centroid of the relevant documents?

Exercise 10.4

- In Rocchio's algorithm, what weight setting for α, β, γ does a "Find pages like this one" search correspond to?

Exercise 10.5

- Why is positive feedback likely to be more useful than negative feedback to an IR system?
- Why might only using one non-relevant document be more effective than using several?

? ? NEXT