workshop-04

2016年8月17日 7:38

Department of Computing and Information Systems
The University of Melbourne
COMP30018/COMP90049
Knowledge Technologies (Semester 2, 2016)
Workshop exercises: Week 4

1. What is the difference between "data retrieval" and "information retrieval"? Why is the latter a knowledge task, but the former is not?

2. [**EXTENSION**] How many books are there in an average library? How many words are there in an average library? How many documents are there on the World Wide Web? How many words?

3. Identify some different types of "informational needs." Often they are given in 3 categories — which categories might subsume other categories?

   (a) Give examples of queries which might indicate a particular type of informational need. Are some of them ambiguous for the type of need?

   (b) [**EXTENSION**] Input some queries of different types of informational needs into a web search engine like Google. Are search engines better at responding to some types of informational need than others? Is there any indication that the search engine is identifying the type of need and tailoring the results toward that?

4. Identify some differences between Boolean querying and ranked querying.

   (a) [**EXTENSION**] Do search engines like Google use Boolean querying or ranked querying? How can we tell?

   (b) [**EXTENSION**] With respect to the contents of a single web page, issues queries of successively greater length until only a single document is returned. Why does this occur? Pay attention to the time taken to resolve each query: what happens as you change the number of terms? Why?

5. Identify the two (sometimes three) components of "TF-IDF" models. Indicate the rationale behind them as in, why would they contribute to a "better" result set?

6. Given a document set made up of five documents, with the indicated term frequencies f_t:

DocID	`apple`	`ibm`	`lemon`	`sun`
Doc_1	4	0	0	1
Doc_2	5	0	5	0
Doc_3	2	5	0	0
Doc_4	1	2	1	7
Doc_5	1	1	3	0

   calculate the document ranking for the (conjunctive) queries: (a) `apple` and (b) `apple lemon`, based on the following TF-IDF term weighting model:

$$w_{d,t} = \begin{cases} 1 + \log_2 f_{d,t} & \text{if } f_{d,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$w_{q,t} = \begin{cases} \log(1 + \frac{N}{f_t}) & \text{if } f_{q,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

7. Revise "language model" approaches toward querying; in which ways is it similar to the "vector space model"? In which ways is it different?

1