

Information Retrieval and Text Mining: Assignment 4

Problem 1.

Download a new corpus.

1. The Corpus: <http://dg3rtljvitrle.cloudfront.net/cacm.corpus>
2. The relevance judgements: <http://www.search-engines-book.com/data/cacm.rel>
3. The Queries: <http://www.search-engines-book.com/data/cacm.query.xml>

CACM is a collection of abstracts of articles published in the Communications of the ACM journal between 1958 and 1979

Excercise

1. Use the above queries as batch queries to retrieve 100 results for each query and store them in “cacm.results”.

17 Q0 CACM-2572 23 -110.13338470 galago

Interpretation: For query number 17 the document CACM-2572 is ranked 23 and has a score -110.13338470.

2. For any four queries having more than 15 relevance judgements (e.g. query 26 has 30 relevance judgements.) Examine the documents for which relevance judgements are given to see why they might be relevant.
3. Calculate the Precision, Recall and F-Score for each query. Assume Galago returns only 100 results and only the documents in relevance judgements are relevant for the query.

4. Use Galago’s inbuilt evaluation on the above queries:

bin/galago eval -path to cacm results- -path to query judgements-

MAP: Mean Average Precision.

For a single query, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved. This value is then averaged over all queries. That is, if the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \dots, d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k then

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \quad (1)$$

Compare MAP scores with the above precision-recall and F-Scores for each query.

5. Why does average precision approximate area under precision-recall curve?
6. Can you give a good reason why it is more useful to use MAP rather than precision-recall measures in search engines today? Illustrate with any 1 observation from above collection or otherwise.

Exercise date: Friday, December 2, 2011, 14:00.