11741 Reading Summary, Ch 8.1-8.4 WWW2002

Xin Qian (xinq@cs.cmu.edu)

Ch 8.1

Ad hoc information retrieval effectiveness measurement needs three pieces: the document collection, the test suite and the relevance judgments set. Notice that relevance is relative to an information need. A document that contains all the words in the query might still be irrelevant. The usual procedure to tune system performance is to do so ona development test collections and then runs the tuned system on the test collection to get an unbiased estimate of performance.

Ch 8.2

Several well-known standard test collections including ad hoc information retrieval collection such as *Cranfield* collection, TREC collection, NTCIR collection and CLEF collection, and text classification collection such as Reuters and 20 Newsgroups.

Ch 8.3

Precision and Recall, both concentrates on the return of true position (relevant documents), are used to measure the effectiveness of unranked retrieval situations. The two quantities trade off against each other. Recall is a non-decreasing function but precision tends to decrease as the number of documents retrieved increases. Accuracy is the fraction of correct classification which can be used if we see an information retrieval system as a two-class classifier. F score is the weighted harmonic mean of precision and recall.

Ch 8.4

For ranked retrieval results, the set of retrieved documents can be potted to have a saw-tooth shaped *precision-recall curve*. *Interpolated precision* is every highest precision at each certain recall level. *11-point interpolated average precision* is used to examine the entire precision-recall curve. *Mean Average Precision* the the average over information needs on the individual average of precision for the top k (k from 1 to the total number of documents retrieved) documents retrieved. *R-precision* is a break-even point when the precision and the recall are identical. ROC curve plots sensitivity against (1-specificity). NDCG is for graded notions of relevance.

Haveliwala 2002

This paper discussed a topic-sensitive modification to the original PageRank algorithm. Instead of a signle importance score, they compute a set of scores w.r.t. various topics. The first step of topic-sensitive PageRank is to generate a set of "basis" topics. We replace the uniform damping vector with the nonuniform damping vector. At query time, each query has a set of class probability for each topic class, parameters are tuned by maximum-likelihood estimates using a unigram language model. The probabilistic interpretation of this query-sensitive PageRank follows the "random surfer" model. Experiment results are evaluated using two measures, absolute ranking overlap measures and relative ranking agreement rate. The paper conducted a user study to compare the query-sensitive approach to ordinary PageRank. A majority of users prefers rankings induced from topic-sensitive PageRanks socres.