Evaluation of System Measures for Incomplete Relevance Judgment in IR

Shengli Wu and Sally McClean

School of Computing and Mathematics, University of Ulster, UK
{s.wu1, si.mcclean}@ulster.ac.uk

Abstract. Incomplete relevance judgment has become a norm for the evaluation of some major information retrieval evaluation events such as TREC, but its effect on some system measures has not been well understood. In this paper, we evaluate four system measures, namely mean average precision, R-precision, normalized average precision over all documents, and normalized discount cumulative gain, under incomplete relevance judgment. Among them, the measure of normalized average precision over all documents is introduced, and both mean average precision and R-precision are generalized for graded relevance judgment. These four measures have a common characteristic: complete relevance judgment is required for the calculation of their accurate values. We empirically investigate these measures through extensive experimentation of TREC data and aim to find the effect of incomplete relevance judgment on them. From these experiments, we conclude that incomplete relevance judgment affects all these four measures' values significantly. When using the pooling method in TREC, the more incomplete the relevance judgment is, the higher the values of all these measures usually become. We also conclude that mean average precision is the most sensitive but least reliable measure, normalized discount cumulative gain and normalized average precision over all documents are the most reliable but least sensitive measures, while R-precision is in the middle.

1 Introduction

To evaluate the effectiveness of an information retrieval system, a test collection, which includes a set of documents, a set of topics, and a set of relevance judgments indicating which documents are relevant to which topics, is required. Among them, "relevance" is an equivocal concept [3, 11, 12] and relevance judgment is a task which demands huge human effort. In some situations such as to evaluate some searching services on the World Wide Web, complete relevance judgment is not possible. It is also not affordable when using some large document collections for the evaluation of information retrieval systems. For example, in the Text REtrieval Conferences (TREC) held by the National Institute of Standards and Technology of the USA, only partial relevance judgment is conducted due to the large number of documents (from 0.5 to several million) in the whole collection. A pooling method [8] has been used in TREC. For every query (topic) a document pool is formed from the

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 245 – 256, 2006.

top 100 documents of all or a subset of all the runs submitted. Only those documents in the pool are judged by human judges and those documents which are not in the pool are not judged and are assumed to be irrelevant to the topic. Therefore, many relevant documents can be missed out in such processing [17]. "Partial relevance judgment" or "incomplete relevance judgment" are the terms used to refer to such situations.

The TREC's pooling method does not affect some measures such as precision at a given cut-off document level. However, in the evaluation of information retrieval systems, both precision and recall are important aspects and many measures concern both of them at the same time. In order to calculate accurate values for such measures, complete relevance judgment is required. Probably mean average precision (MAP) and R-precision are two such measures that are most often used recently. There have been some papers [1, 2, 5, 13] which investigate the reliability and sensitivity of MAP and R-precision.

In the context of TREC, Zobel [17] investigated the reliability of the pooling method. He found that in general the pooling method was reliable, but that recall was overestimated since it was likely that 30% ~ 50% of the relevant documents had not been found.

Buckley and Voorhees [5] conducted an experiment to investigate the stability of different measures when using different query formats. Results submitted to the TREC 8 query track were used. In their experiment, recall at 1000 document level had the least error rate, which was followed by precision at 1000 document level, R-precision, and mean average precision, while precision at 1, 10, and 30 document levels had the biggest error rates.

Voorhees and Buckley [14] also investigated the effect of topic size on retrieval results by taking account of the consistency of rankings when using two different sets of topics for the same group of retrieval systems. They found that the error rates incurred were larger than anticipated, therefore, researchers needed to be careful when concluding one method was better than another, especially if few topics were used. Their investigation also suggested that using precision at 10 document level incurred higher error rate than using MAP.

Concerning that fact that some existing evaluation measures (such as MAP, R-precision and precision at 10 document level) are not reliable for substantially incomplete relevance judgment, Buckley and Voorhees [6] introduced a new measure, which was related to the number of irrelevant documents occurring before a given number of relevant documents in a resultant list, to cope with such a situation.

Sanderson and Zobel [13] reran Voorhees and Buckley's experiment (Voorhees & Buckley, 2002) and had similar observations. But they argued that precision at 10 document level was as good as MAP if considering both the error rate of ranking and the human judgmental effort.

Järvelin and Kekäläinen [7] introduced cumulated gain-based evaluation measures. Among them, normalized discount cumulated gain (NDCG) concerns both precision and recall, which can be used as an alternative for MAP. Using cumulated gain-based evaluation measures, Kekäläinen [9] compared the effect of binary and graded relevance judgment on the rankings of information retrieval systems. She found that these measures correlated strongly under binary relevance judgment, but the