Review of Paper 8: - Ramkumar Rajabaskaran (85241493)

This paper compares the sensitivity of Information retrieval metrics by comparing the effectiveness of retrieval systems based on a set of judged queries. Though there exist systems like MAP, NDCG that perform such comparisons, recently there is another system that performs the comparison through interleaving of two rankings and tracking user clicks. Through this method, it was found that large differences in effectiveness can be identified with much better reliability. As the effective measures mature, sensitivity of measurement becomes critical as we may tend to reject many small but critical improvements that may affect the overall effectiveness.

The primary form of evaluation is based on a set of test collections comprising of query topics and human judgements on the topic-documents, but sensitivity and fidelity of this method tends to be dependent on the number of topics and whether these reflect real world judgement. Another approach is based on user behavior and by measuring the clicks and general browsing patterns of the user. This tends to be cheaper for a system with real users. This paper considers the reliability, sensitivity and agreement of these evaluation approaches. In a few related studies done in this aspect, it was found that an interleaved evaluation allowed clicks to identify the better of two rankings quickly and reliably.

To test this system of measuring sensitivity, they perform 5 experiments, A, B, C are performed as Major Experiments with change of 0.5% between MAP and NDGC. The other two are considered Minor Experiments with change of less than 0.2%. Each ranker was evaluated using both standard information using approximately 12000 queries as well as user traffic. The relevance of the top ten results returned by each ranker was assessed by trained judges.

Interleaving evaluation combines the results of two retrieval functions and presents this combination to the user. The user's clicks indicate a relative performance comparing the quality of the two retrieval functions. The ranking that contained the most clicked results is considered. This Interleaving is achieved using Team-Draft Algorithm. The paper addresses the following questions. How many queries must be judged to obtain significant results, does interleaving produce correlated metrics, how many impressions are needed for comparable results and how do design choices affect the outcome of the analysis.

Firstly, from the sample of 12000 queries, subsample of n subqueries is taken and score of each input as per the three ranking algorithms are taken. It is seen that for small query sets, the preference is on the NDCG@5. For a higher 50 to 200 queries, there is preference for the better ranker 50 to 90% of the time. As expected larger changes in ranking quality is found for smaller query sizes. The sensitivity of the interleaving was found to be 95% reliable after about 50000 impressions for about 5000 judged queries. It was also seen that there was a high degree of binomial confidence and correlation. Apart from this Team-Draft method of interleaving, other methods like Impression Aggregation, to count how many rankers is preferred by a query and Credit assignment where each click is weighted and then decision is made among other methods were also presented.

My summary: - This paper compares the sensitivity of rankers to the small changes in evaluation metrics as it is said that rankers may ignore small critical changes. This can be avoided by using an interleaving system to perform the evaluation metrics by choosing the better ranker. Thus, we can improve the judgement and ranking performance of evaluation systems with the aid of user clicks. We can see that with the help of Interleaver, we can get a high correlation between the results. Thus, they have provided a strong agreement between the judgement based and click based evaluations and the volume of query needed to make realistic changes in retrieval quality. Thus, giving us better relevant search results