Ad Hoc IR - Not Much Room for Improvement

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

David Keeler
Department of Computer Science
University of Otago
Dunedin, New Zealand
dkeeler@cs.otago.ac.nz

ABSTRACT
Ranking function performance reached a plateau in 1994. The reason for this is investigated. First the performance of BM25 is measured as the proportion of queries satisfied on the first page of 10 results – it performs well. The performance is then compared to human performance. They perform comparably. The conclusion is there isn't much room for ranking function improvement.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search process*.

General Terms
Measurement, Performance, Experimentation, Human Factors.

Keywords
Relevance Ranking, Document Retrieval.

1. INTRODUCTION
Armstrong et al. [2, 3] investigate the magnitude of the improvements in ad hoc searching seen on the TREC collections since 1994. They find "no discernible upward trend in Ad-Hoc scores over time" and find the results "reported in 2008 generally indistinguishable from those reported in 1999". In 2009 the best ad hoc relevance-in-context run submitted to INEX did not use structure in ranking [5]. A better ranking function and new features that improve ranking have remained elusive.

As a rule of thumb, the best ranking algorithms all perform at about the same level. BM25 with two tuning parameters performs comparably to language models with one and Divergence From Randomness [1] with none. Ranking performance has plateaued.

This plateau could be crossed if we knew the cause. Buckley [4] investigates the reasons why search engines fail and concludes that significant improvement may be seen if we could identify the cases where we incorrectly emphasized some aspect of the query. Query performance prediction research has not been fruitful.

We ask: *Is there room for improvement in ad hoc ranking?*

We explore this question by comparing stemmed BM25 performance to human performance (language models might equally be used). To do this we take all the TREC and INEX test collections that were multiply assessed and generate synthetic human runs. Comparing the performance of these runs to a search engine suggests that *there is little room for improvement*. We validate this by comparing manual and automatic runs submitted to TREC and again see the same result.

SIGIR'11, July 24-28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

2. DOCUMENT COLLECTION
Two TREC and four INEX ad hoc collections were multiple assessed. The TREC 4 collection is 567,529 documents with 49 queries. The TREC 6 collection is 556,077 documents with 50 queries. After TREC 4 assessment, up-to 200 relevant and 200 non-relevant documents were randomly chosen for reassessment by two alternate assessors. At TREC 6 the University of Waterloo submitted a manual run of documents thought to be relevant for each topic. For TREC experiments we trained on TREC 3.

The INEX 2002-2004 IEEE collection contains 12,107 academic articles, extended to 16,819 in 2005. The INEX 2006-2008 collection contains 659,388 Wikipedia articles and the INEX 2009-2010 collection has 2,666,190 Wikipedia articles. Between 2002 and 2010 the topic sets contained 24, 32, 34, 29, 114, 107, 70, 68, and 51 topics respectively. INEX topic pools were double assessed in parallel by participants who did not know their topic was chosen for reassessment. 5 topics were double assessed in 2004, 4 in 2005, and 15 in 2006. For our INEX experiments assessments were converted to whole document binary assessments by considering a document relevant if any part of it was relevant. The <title> field of the Content-Only (CO) topics were used (those without structural hints). Training was on INEX 2008.

The TREC and INEX designation of original and alternative assessor is maintained throughout these experiments.

3. EXPERIMENT 1:HOW GOOD IS IR?
Mean un-interpolated Average Precision (MAP) has been criticized for being recall oriented [7] and so a precision metric is used; success at *n*, S@n, the proportion of queries for which a relevant document is found at or before position *n* in the results.

The performance of the search engine for all topics in the sets was measured at each position on the first page of 10 results. The first page was chosen because clicking to a subsequent page is infrequent [6]. The results are in Figure 1 which shows that a relevant document is returned at position 1 between 56% and 86% of the time with TREC 6 being the outlier. By the end of the first page satisfaction levels are in excess of 96%, 88%, and 84% respectively for the Wikipedia, IEEE and TREC sets. Search engines are very effective at putting a relevant result high in the results list.

4. EXPERIMENT 2:HOW GOOD ARE WE?
Despite MAP criticism, it is important to ground this work in well understood metrics. The MAP of the search engine is compared against possible scores for the alternate assessor by constructing a synthetic best, typical (expected) and worst run for the assessor.

If all relevant documents are equally relevant then all permutations of relevant documents are equally good (to the alternate assessor). The highest scoring, however, starts with those documents both assessors consider relevant. The lowest has those the alternate assessor considers relevant and the original assessor does

not, followed by those they both agree are relevant. These lists are the upper and lower bound on human performance, but it is important to compute the *expected* performance.

For each topic, 10,000 random permutations of the alternate assessor's relevant documents were generated and Average Precision (AP) computed for each. The average of these APs is the *expected* AP. The mean over all topics is the *expected* MAP.

The *expected* MAP scores were compared to those of the search engine. In Figure 2 the line represents the range of human scores (highest to lowest) with a tick at the *expected* score. The bar is the search engine. For TREC 4 the score of the 2 alternate assessors are designated alt1 and alt2. For TREC 6 the designation is alt; but also included is the original assessor's performance measured against the alternate assessor's ground truth (designated org).

Manual and automatic runs were submitted to TREC 4 and 6. The best, worst, and mean MAP of these runs is presented in the last two columns of Figure 2. The line is the range of manual runs (the tick is the mean). The bar is the best automatic run.

It can be seen that the search engine often performs inside the human range and sometimes better than *expected*. The notable exception is TREC 4. This, we believe, is because the alternate assessors were given a small and biased pool and were not asked to assess the original pool (they were unlikely to find new relevant documents). We also note that where there were more than 200 relevant documents for a topic those that were not reassessed were added to the alternate assessor's results (artificially inflating MAP). The runs submitted to TREC 4 are a better performance indicator – and the figure shows that by TREC 4 (1995) manual and automatic runs were performing comparably.

Although of questionable utility for the INEX collection (with few topics), a paired 2-tailed *t*-test was conducted. The results are presented in Table 1 where it can be seen that with TREC 6 and INEX IEEE, there is no statistically significant difference between the performance of the alternate assessor and the search engine.

Table 1: *t*-test scores, search engine against assessor

	2004	2005	2006	T4 alt 1	T4 alt 2	T6 alt	T6 org
Upper	0.19	0.54	0.00	0.00	0.00	0.00	0.00
Expected	0.06	0.32	0.03	0.00	0.00	0.08	0.01
Lower	0.01	0.13	0.51	0.00	0.00	0.63	0.58

5. DISCUSSION AND CONCLUSIONS

It is pertinent to ask whether this result is consequence of the methodology and search engine training. The high performance of the alternate assessors at TREC 4 is discussed in the previous section. In 1995 at TREC 4 and in 1997 at TREC 6 the manual and automatic runs performed at comparable levels. At INEX the number of multiple assessed topics is low which is likely reflected in the *t*-test results. None the less, it is reasonable to conclude that search engine performance is comparable to human performance.

To examine training, EXPERIMENT 1 was re-conducted on the INEX 2006 double-assessed topics. The nonparametric Divergence from Randomness (I(ne)B2), TF.IDF (inner product), and BM25 functions were used. The results are presented in Figure 3 from where it can be seen that when BM25 was introduced (circa 1994) there was room for improvement on TF.IDF. It can also be seen that the non-parametric I(ne)B2 function performs comparably to BM25 and the assessor.

Recent ranking functions are a substantial improvement on TF.IDF but there remains very little room for improvement in ad hoc search. In future work we will examine this result using graded relevance assessments and in known-entity searching where preliminary results suggests improvements can be made.

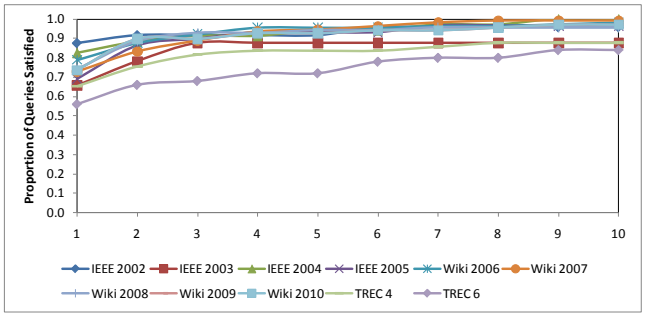

Figure 1: S@N of BM25 on the TREC and INEX collections

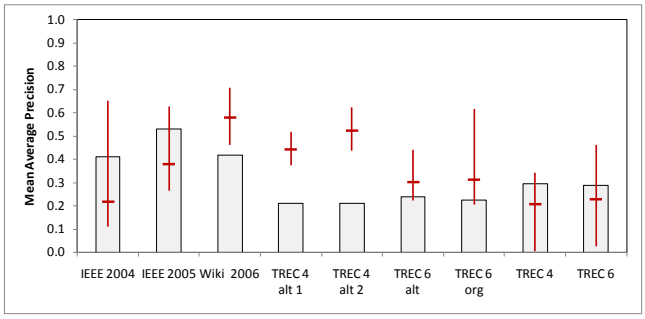

Figure 2: MAP of assessor (line) and search engine (box)

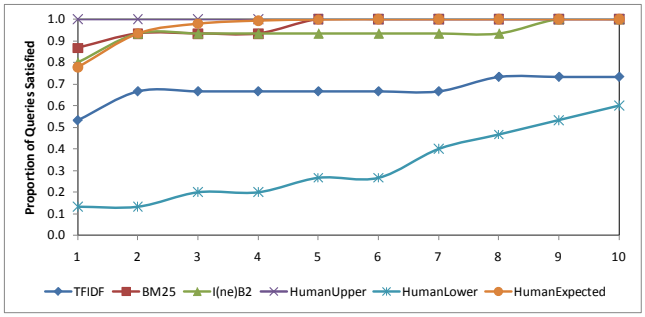

Figure 3:S@N of assessor and different ranking functions

REFERENCES

[1] Amati, G., van Rijsbergen, C.J. (2002) Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness, TOIS 20(4):357-389

[2] Armstrong, T., Moffat, A., Webber, W., Zobel, J. (2009) Has adhoc retrieval improved since 1994? SIGIR 2009:692-693

[3] Armstrong, T., Moffat, A., Webber, W., Zobel, J. (2009) Improvements that don't add up: ad-hoc retrieval results since 1998, CIKM 2009:601-610

[4] Buckley, C. (2009) Why current IR engines fail, Inf. Retr. 12(6):652-665

[5] Trotman, A., Jia, X.-F., Geva, S. (2009) Fast and Effective Focused Retrieval, Proceedings of INEX 2009:229-241

[6] Zhang, Y., Moffat, A. (2006) Some observations on user search behavior, ADCS 2006:1-8

[7] Zobel, J., Moffat, A., Park, L. (2009). Against recall: is it persistence, cardinality, density, coverage, or totality? SIGIR Forum 43:1:3-8