A Comparison of Evaluation Measures Given How Users Perform on Search Tasks

James A. Thom *Falk Scholer*

School of Computer Science and IT
RMIT University
Vic 3001 Australia

{*james.thom,falk.scholer*}*@rmit.edu.au*

Abstract *Information retrieval has a strong foundation of empirical investigation: based on the position of relevant resources in a ranked answer list, a variety of system performance metrics can be calculated. One of the most widely reported measures, mean average precision (MAP), provides a single numerical value that aims to capture the overall performance of a retrieval system. However, recent work has suggested that broad measures such as MAP do not relate to actual user performance on a number of search tasks. In this paper, we investigate the relationship between various retrieval metrics, and consider how these reflect user search performance. Our results suggest that there are two distinct categories of measures: those that focus on high precision in an answer list, and those that attempt to capture a broader summary, for example by including a recall component. Analysis of runs submitted to the TREC terabyte track in 2006 suggests that the relative performance of systems can differ significantly depending on which group of measures is being used.*

Keywords Information Retrieval, evaluation, metrics

1 Introduction

Information retrieval (IR) has a long history of experimental evaluation, following the “Cranfield” methodology: a set of queries (or topics) are run over a static collection of documents. For each returned query and document combination, a human judges whether the document is *relevant* to the query. This methodology is widely applied in IR research, and forms the basis of the on-going series of Text Retrieval Conferences (TREC).

Based on the relevance judgements, a variety of metrics can be calculated, aiming to reflect the performance of the retrieval system. Such metrics are generally based on two underlying concepts: the *precision* of a retrieval system, defined as the number of relevant documents retrieved as a proportion of the total number of documents that have been retrieved; and the *recall*, defined as the number of relevant documents that have been retrieved as a proportion of the total number of relevant documents in the collection. Precision therefore reflects the accuracy of an answer list, while recall measures the completeness.

Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007. Copyright for this article remains with the authors.

Since information retrieval experiments generally focus on the performance of a system across a set of 50 or more queries, various summary measures are widely used. However, recent work has suggested that some of the most widely reported IR metrics have no relationship with user-based evaluation measures on precision-based search tasks [7]. Motivated by these findings, we investigate the relationship between different retrieval measures. Our results indicate that there is a distinct difference between measures that focus only on high-precision and those that aim to provide a more inclusive summary of retrieval performance.

In Section 2 we survey related work on comparison of information retrieval measures. Five commonly used information retrieval measures—precision at 1 (P@1), precision at 10 (P@10), mean average precision (MAP), mean reciprocal rank (MRR), and R-precision (RP)—are presented in Section 3. These measures are compared with each other on retrieval system runs submitted to the TREC Terabyte track. Finally, in Section 4 we discuss the implications of these results on how to determine what measures should be used when evaluating IR systems.

2 Related Work

A variety of information retrieval metrics have been proposed in the literature. While the relationship between some metrics has been considered previously [2, 8], such studies have not focused on precision at 1 and mean reciprocal rank. Recent studies have investigated correlations between retrieval metrics and user performance [7] or user satisfaction [1, 5].

Turpin and Scholer [7] found that commonly reported measures, in particular mean average precision, do not correspond well with user performance on simple information-finding web search tasks. Their results suggest that measures such as precision at 1 are more likely to reflect actual user performance.

Huffman and Hochster [5] found that for navigational queries to a search engine there was a close correlation between the relevance of the first result returned (P@1) and user satisfaction. While relevance of second and third results in the ranked list contributed a little to user satisfaction for navigational queries, these latter ranked results contributed more to user satisfaction for non-navigational queries. Al-Maskari et al. [1] compare precision and three cumulative gain measures with user satisfaction of accuracy, coverage, and ranking of results. They were not able to find one measure that captured all these aspects of user satisfaction.

3 Comparison of Measures

In order to compare whether different evaluation measures are measuring similar or different things, we compare the relative performance of all 80 runs that were submitted for the 2006 Terabyte track *adhoc* retrieval task. This task consisted of 149 informational queries run on the GOV2 collection (TREC topics 701–850).

The runs are compared using 5 standard IR evaluation measures. These five measures are defined for a given run (that is a ranked list of answers retrieved by a system) over the 150 topics as follows.

Precision at 1 (P@1): is the mean (calculated over all topics) of the precision of the top ranked document retrieved.

Precision at 10 (P@10): is the mean (calculated over all topics) of the precision of the first ten documents retrieved.

Mean average precision (MAP): is the mean (calculated over all topics) of average precision, where the average precision of a single query is the mean of the precision scores at each relevant item returned in a search results list.

Mean reciprocal rank (MRR): is the mean (calculated over all topics) of the reciprocal rank of the highest ranking relevant document (zero for a topic if no relevant documents were returned by the system).

R-precision (RP): is the mean (calculated over all topics) of the precision after R_t documents have been retrieved for topic t, where R_t is number of relevant documents available for topic t.

For each of the retrieval measures, the set of submitted runs can be *ordered* from the best-performing run (highest value for a metric) to the worst (lowest value for the same metric). We compare the obtained orderings between different metrics using Kendall's τ correlation coefficient. This coefficient measures the agreement between two sets of ranked data [6].

The correlation between different pairs of retrieval metrics is shown in Table 1, and is also presented graphically in Figures 1 to 7. The straight line in the figures is the line of best fit (if the relationship between the metrics is assumed to be linear; note that Kendall's τ does not make this assumption, since it is based on ranks).

Metrics		Kendall's τ
P@1	MAP	0.386
P@1	MRR	0.865
P@1	P@10	0.659
P@1	RP	0.376
MAP	MRR	0.350
MAP	P@10	0.583
MAP	RP	0.899

Table 1: Kendall's τ correlation between different retrieval metrics for 80 TREC runs. All correlations are statistically significant at $\alpha = 0.01$.

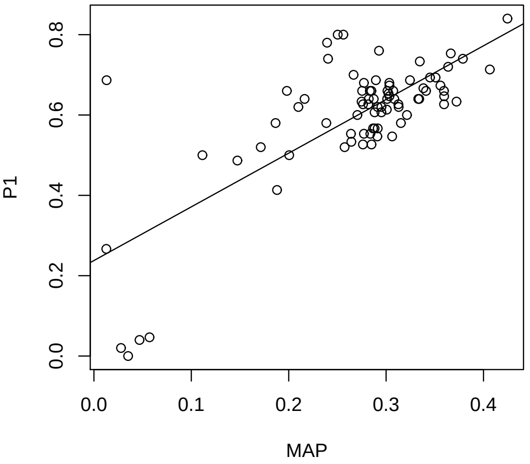

Figure 1: Correlation of P@1 with MAP

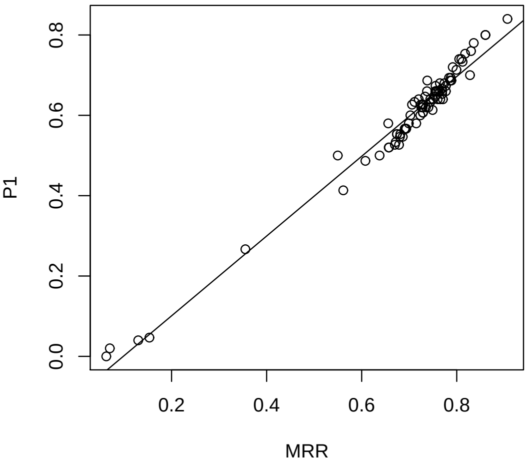

Figure 2: Correlation of P@1 with MRR

Figure 2 shows that there is a strong relationship between the two measures P@1 and MRR ($\tau = 0.865$); this is not surprising since both measures have a strong bias to systems that highly rank one relevant document.

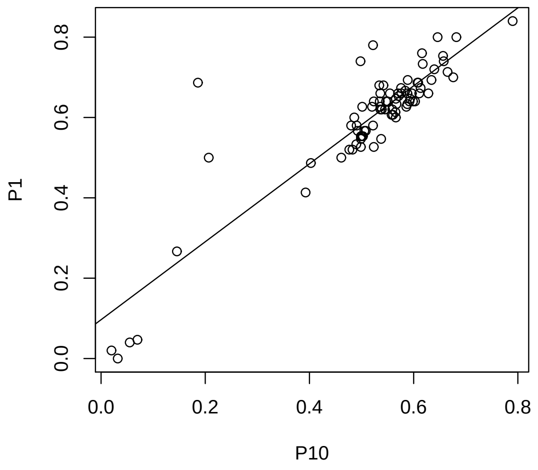

Figure 3: Correlation of P@1 with P@10

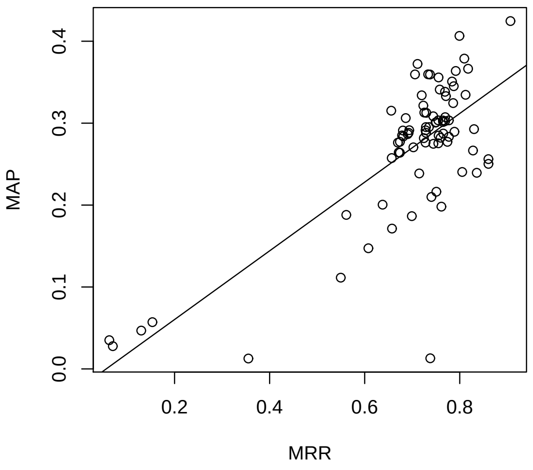

Figure 5: Correlation of MAP with MRR

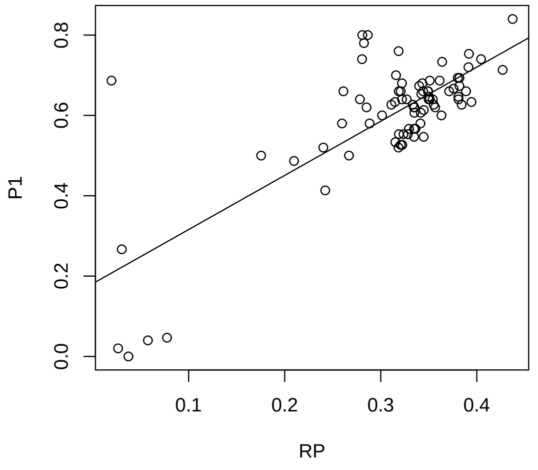

Figure 4: Correlation of P@1 with R-Precision

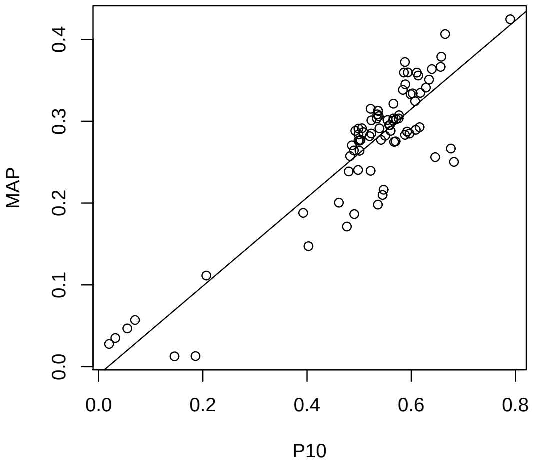

Figure 6: Correlation of MAP with P@10

Similarly, Figure 7 shows that there is a very close relationship between the two measures MAP and RP ($\tau = 0.899$); both of these measures incorporate a recall component.

As can be seen from Figures 3 and 6, there is not as strong a relationship between P@10 and either P@1 or MAP (the correlation coefficient is 0.659 and 0.583, respectively).

Figures 1 and 4 compare P@1 with MAP and RP and Figures 5 compares MAP with MRR; these figures show that systems that perform well at finding one relevant document either as the first answer (P@1), or highly ranked (MRR), do not necessarily perform so well in terms of measures of overall performance such as MAP or RP (while the correlations are statistically significant, τ is below 0.4 in each case, much lower than for other pairs of metrics).

We have also calculated correlations using the *bpref* measure, a retrieval metric that is intended for use in situations where only incomplete relevance judgements are available. When relevance judgements are mostly complete, bpref and MAP are closely correlated [3]. As a result, the correlations between bpref and other metrics closely reflect those between MAP and the other metrics, and are not reported here for brevity.

4 Discussion

Our results suggest at least two distinct categories of measures: those with a strong bias to highly ranking a relevant document (P@1 and MRR), and those that attempt to capture a broader summary of the performance (MAP, RP). The measure P@10 appears to share properties of both categories.

Although MAP has been widely accepted as the defacto standard for evaluation of information retrieval systems, it does not necessarily correspond to how users actually perform on search tasks. This is of particular concern because our results show that the correlations

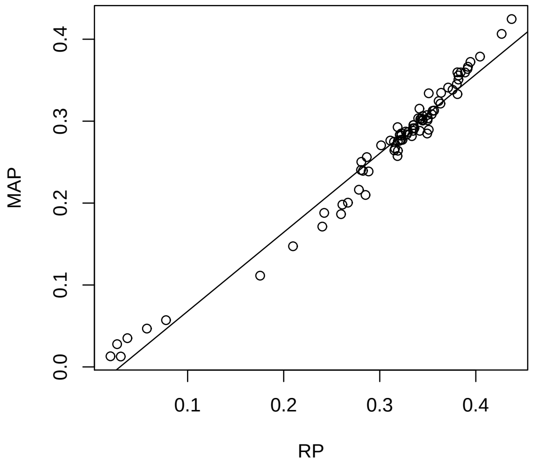

Figure 7: Correlation of MAP with R-Precision

between the two categories of metrics are weak – therefore, the relative ordering of systems that is commonly used in the TREC framework may not reflect user performance.

This preliminary analysis has considered only informational search tasks. Additional user studies are required to determine the wide range of situations—including navigational search tasks and question answering—in which measures such as P@1 and MRR are the more appropriate evaluation measures. Some of these different situations were investigated in the TREC Web track [4], which found that standard IR evaluation methodology did not adequately evaluate web search and different search tasks need specific evaluation metrics.

Furthermore, the retrieval metrics considered in this short paper are all based on binary relevance criteria (a document is either relevant, or it isn't). In future work, we intend to consider other evaluation metrics such as cumulative gain measures which take into account multi-level relevance judgements.

References

[1] Azzah Al-Maskari, Mark Sanderson and Paul Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 773–774, Amsterdam, The Netherlands, 2007.

[2] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In Emmanuel Yannakoudakis, Nicholas J. Belkin, Mun-Kew Leong and Peter Ingwersen (editors), *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 33–40, Athens, Greece, 2000.

[3] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In Kalervo Järvelin, James Allan, Peter Bruza and Mark Sanderson (editors), *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, UK, 2004.

[4] David Hawking and Nick Craswell. Very large scale retrieval and web search. In Ellen Voorhees and Donna Harman (editors), *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005. `http://es.csiro.au/pubs/trecbook_for_website.pdf`.

[5] Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 567–573, Amsterdam, The Netherlands, 2007.

[6] D. Sheskin. *Handbook of parametric and nonparametric statistical proceedures*. CRC Press, 1997.

[7] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In Charles L. A. Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij and Arjen P. de Vries (editors), *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 11–18, Seattle, Washington, August 2006.

[8] E. Voorhees. Evaluation by highly relevant documents. In Donald H. Kraft, W. Bruce Croft, David J. Harper and Justin Zobel (editors), *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 74–82, New Orleans, LA, 2001.