# A Comparative Study of Performance Measures for Information Retrieval Systems

Xiannong Meng
Department of Computer Science
Bucknell University
Lewisburg, PA 17837, U.S.A.

## Abstract

*Traditional performance measures of information retrieval systems include precision and recall and their variants. While these measures work well in closed-laboratory environments, they are not suitable for practical IR systems such as Web search systems. Many single-value measures were proposed to improve over the precision-recall measure, such as* expected search length *(ESL),* average search length *(ASL) and* RankPower*. We compare in this paper the measures of* ESL*,* ASL*, and* RankPower *applied to a set of real Web retrieval data. The results demonstrate that* RankPower *indeed is a feasible, effective, and easy-to-use single-value measure for performance of practical IR systems such as Web search engines.*

**Keywords***: information retrieval, performance measurement, Web search, user studies, user preference*

## 1 Introduction

While user perception is important in measuring the retrieval performance of search engines, quantitative analyses provide more "scientific evidence" that a particular search engine is "better" than the other. Traditional measures of *recall* and *precision* work well for laboratory studies of information retrieval systems. However, they do not capture the performance essence of today's Web information systems for three basic reasons. One reason for this problem lies in the importance of the rank of retrieved documents in Web search systems. A user of Web search engines would not go through the list of hundreds and thousands of results. A user typically goes through a few pages of a few tens of results. The *recall* and *precision* measures do not explicitly present the ranks of retrieved documents. A relevant document could be listed as the first or the last in the collection. They mean the same as far as recall and precision are concerned at a given recall value. The second reason that *recall* and *precision* measures do not work well is that Web search systems cannot practically identify and retrieve all the documents that are relevant to a search query in the whole collection of documents. This is required by the *recall/precision* measure. The third reason is that these *recall/precision* measures are a pair of numbers. It is not easy to read and interpret quickly what the measure means for ordinary users. Researchers (see a summary from [3]) have proposed many *single-value* measures such as estimated search length *ESL* [2], averaged search length *ASL* [4], *F harmonic mean*, *E-measure* and others to tackle the third problem.

We compare in this paper through a set of real-life Web search data the effectiveness of various single-value measures. The use and the results of *ASL*, *ESL*, average precision, F-measure, E-measure, and finally the *RankPower*, applied against a set of Web search results. The experiment data was collected by sending 72 randomly chosen queries to *AltaVista* and *MARS* [1, 5].

## 2 Results and Analysis

Table 1 shows the results of various single-value measures applied to the Web search data. We can draw the following observations from the data. Note that these observations demonstrate the effectiveness of single-value measures, especially, the *RankPower*. The focus was not on the comparison of the actual search engines since the experimental data is a few years old.

- In ESL Type 1 comparison, AltaVista has a value of 3.78 which means on the average, one needs to go through 3.78 irrelevant documents before finding a relevant document. In contrast, ESL Type 1 value for MARS is only 0.014 which means a relevant document can almost always be found at

Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)
0-7695-2497-4/06 $20.00 © 2006 IEEE

the beginning of the list. MARS performs much better in this comparison because of its relevance feedback feature.

- ESL Type 2 counts the number of irrelevant documents that a user has to go through if she wants to find *six* relevant documents. AltaVista has a value of 32.7 while MARS has a value of 25.7. Again because of the relevance feedback feature of MARS, it performs better than AltaVista.
- It is very interesting to analyze the results for ESL Type 3 request. ESL Type 3 request measures the number of irrelevant documents a user has to go through if she wants to find all relevant documents in a fixed document set. In our experiments, the document set is the 200 returned documents for a given query and the result is averaged over the 72 queries used in the study. Although the average *number* of relevant documents is the same between AltaVista and MARS (see the values of estimated ASL) because of the way MARS works, the *positions* of these relevant documents are different. This results in different values of ESL Type 3. In order to find all relevant documents in the return set which the average value is 29.8 documents, AltaVista would have to examine a total of 124 irrelevant documents while MARS would examine 113 irrelevant documents because MARS have arranged more relevant documents to the beginning of the set.
- ESL Type 5 requests examine up to a certain number of relevant documents. The example quoted in Cooper's paper [2] was five. For AltaVista, it takes about 26 irrelevant documents to find five relevant documents, while MARS requires only about 17.

**Table 1. Single-value Measures of Performance for AltaVista and MARS Averaged Over 72 Queries**

| | | AV | MARS |
|---|---|---|---|
| ESL | Type 1 | 3.78 | 0.014 |
| | Type 2 | 32.7 | 25.7 |
| | Type 3 | 124 | 113 |
| | Type 4 | 7.56 | 0.708 |
| | Type 5 | 25.7 | 17.3 |
| ASL | Measured | 82.2 | 77.6 |
| | Estimate | 29.8 | 29.8 |
| RankPower | | 3.29 | 2.53 |
| Revised RankPower | | 0.34 | 0.36 |

## 3 Conclusions

This paper compares the effectiveness of *RankPower* with other single-value performance measures, namely *ESL* and *ASL* using real data collected from Web search. The empirical results show the *RankPower* is a very effective measure. The result of applying *RankPower* to the given set of experiment data actually is consistent with conclusions drawn from using other single-value measures. Yet *RankPower* provides much more intuitive explanation and is much easier to measure.

## References


[1] Z. Chen & X. Meng, MARS: Applying Multiplicative Adaptive User Preference Retrieval to Web Search, *Proceedings of the 2002 International Conference on Internet Computing*, pp. 643-648, CSREA Press, Las Vegas, Nevada, June 24-27, 2002.

[2] W.S. Cooper, Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems, *Journal of the American Society for Information Science*, 19(1), 30-41, 1968.

[3] R.R. Korfhage, *Information Storage and Retrieval*, John Wiley & Sons, 1997.

[4] R.M. Losee, *Text Retrieval and Filtering: Analytic Models of Performance*, Kluwer Publisher, Boston, 1998.

[5] X. Meng & Z. Chen, MARS: Multiplicative Adaptive Refinement Web Search, in *Web Mining: Applications and Techniques*, Anthony Scime, pp. 99-118, Idea Group Publishing, Hershey, PA, U.S.A. 2005.