# Automatic Ranking of Retrieval Systems in Imperfect Environments

Rabia Nuray
Computer Engineering Department
Bilkent University
Bilkent, Ankara 06533, Turkey
rabian@cs.bilkent.edu.tr

Fazli Can
Computer Science and Systems Analysis Dept.
Miami University
Oxford, OH 45056, USA
canf@muohio.edu

## ABSTRACT
The empirical investigation of the effectiveness of information retrieval (IR) systems requires a test collection, a set of query topics, and a set of relevance judgments made by human assessors for each query. Previous experiments show that differences in human relevance assessments do not affect the relative performance of retrieval systems. Based on this observation, we propose and evaluate a new approach to replace the human relevance judgments by an automatic method. Ranking of retrieval systems with our methodology correlates positively and significantly with that of human-based evaluations. In the experiments, we assume a Web-like imperfect environment: the indexing information for all documents is available for ranking, but some documents may not be available for retrieval. Such conditions can be due to document deletions or network problems. Our method of simulating imperfect environments can be used for Web search engine assessment and in estimating the effects of network conditions (e.g., network unreliability) on IR system performance.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (effectiveness)*.

## General Terms
Measurement, Performance, Experimentation.

## Keywords:
IR Evaluation, Automatic Performance Evaluation

## 1. INTRODUCTION
For very large databases, creating relevance judgments is difficult, since it requires human labor and people usually disagree about the relevance of a document; furthermore, judging every document for every query topic in large collections is infeasible. To overcome the difficulty of creating relevance judgments, researchers proposed automatic methods to compare the retrieval effectiveness of IR systems [1, 3].

In this paper, we present a new automatic evaluation methodology that finds the effectiveness of IR systems without human intervention, and compare its performance with that of a human-based approach. We test our method in an imperfect environment (defined in the next section) and evaluate the performance of various IR systems using the relevance judgments formed by our methodology and the relevance judgments formed by human assessors. In the following section, we describe our evaluation methodology and report the experimental results on the consistency of automatic and human-based IR system performance rankings.


*SIGIR '03*, July 28-August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

## 2. EVAULATION METHOD
Our automatic evaluation method is based on heuristics. We first generate a pool of documents by using the top *b* documents returned by each IR system for a given query. If we have *n* retrieval systems, then the maximum number of documents in this pool would be (*n***b*); however, due to common documents that will be returned by different retrieval systems, the number of unique documents in the pool generally will be smaller than this maximum number. We then rank the pooled documents according to their similarity to the user query by using the vector space model. The top *s* documents of the ranking obtained for each topic are assumed to be (pseudo) relevant documents for that topic. The effectiveness of each retrieval system is computed using these automatic relevance judgments for each query and, finally, the overall system performance is obtained by finding the average for all queries.

### 2.1 Experiments
In the experiments, we used the data generated by the TREC project managed by NIST. For this purpose, we used the retrieval runs submitted to the ad hoc task of TREC-5 (there were 61 participants, for our purposes they represent *n*=61 different IR systems since each of them use a different IR algorithm) from the TREC Web site with the corresponding relevance judgments. The queries used in the experiments were the TREC topics 251-300. In TREC-5, each participating group returns a ranked list of documents from the databases TREC-1-4 for each topic. We only used the documents in the TREC-4 databases and assumed that the documents of the TREC-1-3 databases were inaccessible; their inaccessibility simulates an imperfect environment. To see the consistency of our method with the human-based methods, we also assumed that the human-relevance assessors could only access the TREC-4 documents, and that the relevance judgments used in actual TREC rankings were modified to include only the documents in TREC-4 databases, i.e., the inaccessible documents were assumed to be irrelevant. Thus, our approach simulates a Web-like imperfect environment: document indexing information is available for document ranking (so that retrieval systems can rank all documents); however, some of the top ranking documents are unavailable due to reasons such as network problems or document deletions.

Since the experiments were performed on TREC data, we followed the steps of the official TREC evaluation process [2]

with a small change. Our methodology replaces the steps that form the relevance judgments and perform the evaluations. Our evaluation process takes the top (*b*=) 30 or 200 accessible documents per topic from each official run to form the pool for that topic (we have tested other *b* values as well, but report only the results of these two due to limited space). Then we sort the pooled documents using a matching function based on the vector space model (by paying attention to document frequencies, *df*, and inverse document frequencies, *idf*, of the stemmed document words within the generated pool) to form the automatic relevance judgments (i.e., top *s* documents, or *pseudo qrels*). Finally in our automatic approach we evaluate all runs using the **trec_eval** package with the *pseudo qrels*.

## 2.2 Statistical Significance

To determine if our method is consistent with the actual TREC rankings, we measured the correlation of these two methods using Kendall's τ correlation coefficient. We computed the correlation of our method to the human-based rankings for the average precision and the precision at DCV (document cut-off value) documents retrieved (P@DCV=5, 10, 15, or 20).

## 2.3 Results

In the experiments, our purpose is to find the *b* and *s* values that will yield the highest level of correlation between automatic and human based evaluations. The Kendall's τ correlation of our method to the human-based rankings with a pool depth of *b*= 200 with various numbers of relevant documents (i.e., various *s* values) is given in Table 1, and are all significant for $\alpha = 0.01$.

**Table 1. Kendall's τ correlation of automatic method with human-based evaluations for different measures (*b*= 200)**

| s | Avg. Pre. | P@5 | P@10 | P@15 | P@20 |
|---|---|---|---|---|---|
| 100 | 0.377 | 0.373 | 0.353 | 0.328 | 0.316 |
| 200 | 0.351 | 0.351 | 0.362 | 0.339 | 0.326 |
| 300 | 0.340 | 0.380 | 0.349 | 0.338 | 0.315 |
| 500 | 0.335 | 0.511 | 0.331 | 0.333 | 0.311 |
| Exact | 0.325 | 0.348 | 0.387 | 0.408 | 0.408 |

We then tested our method with a pool depth of *b*=30 accessible documents of each run for each topic to see how the number of documents in the pool affects the performance of our method. The Kendall's τ correlation of both methods for various *s* values of relevant documents is given in Table 2. The correlations are all significant for $\alpha = 0.01$ and they are stronger than the correlations observed with a pool of top 200 accessible documents.

**Table 2. Kendall's τ correlation of automatic method with human-based evaluations for different measures (*b*= 30)**

| s | Avg. Pre. | P@5 | P@10 | P@15 | P@20 |
|---|---|---|---|---|---|
| 100 | 0.399 | 0.399 | 0.390 | 0.365 | 0.354 |
| 200 | 0.384 | 0.413 | 0.382 | 0.380 | 0.352 |
| 300 | 0.396 | 0.421 | 0.360 | 0.352 | 0.324 |
| 500 | 0.405 | 0.363 | 0.364 | 0.323 | 0.333 |
| Exact | 0.343 | 0.398 | 0.410 | 0.448 | 0.449 |

Figure 1 graphically shows the correlation of our method to the human based evaluations for P@20 values. The runs in the figure are sorted by their performance according to the human based evaluations. Observe that the shapes of the curves created by the rankings are very similar, especially in the middle to lower performance range. We also performed the experiments using the random selection process given in [3]. We selected the best performing combination of systems reported in [3]. We created a pool of top 10 accessible documents (with duplicated documents) and selected the exact number of relevant documents for each topic. We repeated the random experiments ten times; see Table 3 for the results. The correlations are all significant for $\alpha = 0.01$, but P@DCV values are not as strong as the correlations obtained with a pool of top 30 accessible documents with various number of relevant documents using our automatic method.

**Table 3. Kendall's τ correlation of automatic method with human-based evaluations for random selection method**

| τ | Avg. Pre. | P@5 | P@10 | P@15 | P@20 |
|---|---|---|---|---|---|
| Mean | 0.401 | 0.366 | 0.343 | 0.328 | 0.330 |
| Std. D. | 0.050 | 0.037 | 0.033 | 0.035 | 0.038 |

# 3. CONCLUSIONS

We have presented an automatic evaluation approach that finds the relative ranking of retrieval systems. Our method does not give the exact performance of individual systems; however, its results correlate significantly and positively to the human-based rankings. Our automatic approach is valuable in evaluating systems such as Web search engines. It has a stronger consistency with the human–based evaluations than the random selection process. Our method of simulating imperfect environments is also interesting and has practical implications; for example, it can be used in estimating the effects of the network conditions (e.g., network unreliability) on IR system performance.

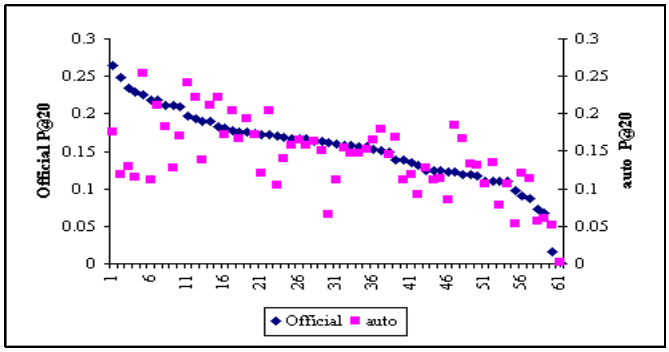


**Figure 1. The correlation of human based evaluations to our method with a pool of top 30 documents with exact number of relevant documents to each topic using P@20 values**

# 4. REFERENCES


[1] Chowdhury A., Soboroff I. Automatic evaluation of World Wide Web search services. *In the Proceedings of the 2002 ACM SIGIR Conference, 421-422*.

[2] Voorhees E.M., Harman, D. Overview of the Fifth Text Retrieval Conference (TREC-5). In E. M. Voorhees and D.K. Harman, editors, *The Fifth Text Retrieval Conference,* NIST Special Publication 500-238. National Institute of Standards and Technology, Gaithersburg, MD, November 1996.

[3] Soboroff, I., Nicholas,C., Cahan, P. Ranking Retrieval Systems without Relevance Judgments. *In the Proceedings of the 2001 ACM SIGIR Conference,* 66-73.