IOSR Journal of Engineering
Apr. 2012, Vol. 2(4) pp: 794-802

STATISTICAL SIGNIFICANCE IN MULTILINGUAL INFORMATION RETRIEVAL (MLIR) SYSTEM

Sadanandam Manchala[1] and Chandra Mohan D[2]
[1]*Department of Computer Science and Engineering, Kakatiya University, Warangal, India*
[2]*Department of Computer Science and Engineering, Matrix Institute of Technology and Engineering, Hyderabad, India*

ABSTRACT
The efficiency of retrieval system is precise by comparing performance on a regular set of queries in Information Retrieval (IR) and MLIR systems. Significance tests are often used to estimate the reliability of such comparisons. In this research paper, we revisit the question of how such significance tests should be used. We find that the t-test is highly reliable than the sign and Wilcoxon test, and is far more reliable than simply showing a large percentage difference in effectiveness measures between IR and MLIR systems. Our results show that previous experimental work on significance tests over-estimated the error of such tests. We also re-consider comparisons between the reliability of Average Precision(AP), Mean Average Precision(MAP), Average Mean Reciprocal Rank (AMRR) and Average Discounted Cumulative Gain (ADCG) arguing that past comparisons did not consider the assessor effort required to compute such measures. This research shows that judge effort would be better spent building test collections with more queries, each assessed in less detail.

KEYWORDS: SIGNIFICANCE, WILCOXON SIGNED-RANK TEST, SIGN TEST, T-TEST, MLIR, AMRR, ADCG.

1. INTRODUCTION

Test collections are the principal tool used for comparison and evaluation of retrieval systems. These collections typically comprised of queries and relevance judgments have been a key part of multilingual information retrieval research for decades; the use of such collections is based on research and practice in collection formation (Sparck Jones & Van Rijsbergen, 1975; Voorhees & Harman, 1999) and measurement of retrieval effectiveness (Van Rijsbergen 1979; Dunlop 1997; Jarvelin, 2000; Buckley, 2004). Effectiveness is computed by measuring the ability of systems to find relevant documents. The measured score is most often used as an indicator of the performance of one system relative to another; with an assumption that similar relative performance will be observed on other test collections. When researchers report results of a retrieval experiment and show, using some effectiveness measure, that one retrieval system is better than another, significance tests are often used to evaluate the result. The tests provide evidence that the observed difference in effectiveness is not due to chance. Significance tests such as the t-test or Wilcoxon are commonly used. The question of how likely it is that a significant result observed on a test collection will continue to be observed in other settings has not been as widely investigated as having other aspects of test collections. Significance tests require that the data being tested has certain properties. Among the assumptions of the Wilcoxon signed-rank test and the t-test are that the values being tested in this case, per-query effectiveness are distributed, respectively, symmetrically and normally (Van Rijsbergen, 1979); however, effectiveness rarely follows either distribution. The tests also assume that each set of query values being tested in a random sample from retrieved documents. What the tests determine, in comparing the runs of two retrieval systems, is whether the two samples are from the same retrieved documents of effectiveness the systems are equivalent or different retrieved documents i.e. one system gets better results than the other. In this context, where significance tests are widely used in MLIR experiments but their impact is little understood, that we undertook the work reported in this paper. Our results suggest new procedures for evaluation of retrieval systems and show that both a relative improvement in measured effectiveness and statistical significance are required for confidence in results. If significance is improvement is small as is the case in many papers – results are not reliable.

The remainder of this paper is organized as follows. Section 2 discusses related work to this study. Section 3 describes evaluation metrics of multilingual information retrieval system; Section 4 represents the experimental Results of Statistical tests in mlir system. Finally section 6 concludes the paper.

2. RELATED WORK

There is limited discussion of significance tests in IR literature. (Van Rijsbergen 1979) detailed the shape and form of data distributions to which the sign, Wilcoxon and t-tests can be applied. He showed that test collection data fails to meet all the requirements of any of the tests and warned that none can be used with confidence. Countering such caution, (Hull 1993) described past (non-IR) work showing that the t-test can be reliable even when data being tested is not distributed normally. Hull described a range of significance tests, but did not empirically test them.

(Savoy 1997) investigated several significance measures and echoed Van Rijsbergen's concerns. He proposed an alternative bootstrap method, based on sampling from a set of query outcomes; it is not clear whether this approach could be applied with the small sets of queries for which we have relevance judgments and to our knowledge it has not been used in practice to assess significance.

The Wilcoxon and t- test are common significance tests used in IR experiments. Both take a pair of equal-sized sets of per-query effectiveness values, and assign a confidence value to the null hypothesis: that the values are drawn from the

same population. If confidence in the hypothesis (reported as a p-value) is ≤0.05, it is typically rejected. Although such tests only consider the null hypothesis, it is common to assume rejection implies that values are from different retrieved documents with likelihood >0.95. Apart from correctly determining significance or a lack thereof, the tests also produce type I and type II errors. A type I error is a false positive; for a p-value of 0.05, one positive test in twenty is expected to be a type I error. A type II error is a false negative; the incidence of type II errors is unknown. In the statistical and medical communities, there has been concern that the theory underestimates the rate of type I error, but that the small samples used in typical studies mean that there is insufficient data to determine significance and thus that type II errors may be common, leading to useful methods being discarded prematurely (Matthews 2003).

The first IR-based work to measure the utility of significance tests was that of (Zobel 1998), who split the fifty topics of TREC-5 into two disjoint sets of 25: one set holding topics 251-275, the other topics 276-300. Taking the 61 runs submitted to TREC-5, Zobel compared each run with every other, resulting in 1,830 pair-wise comparisons. If a significant difference was observed between a pair of runs measured on the first 25 topics (that is, if the null hypothesis was rejected, p≤0.05), the ordering of the runs based on an effectiveness measure was noted and the same pair of runs was compared on the second 25 topics. If the ordering of runs on both sets was the same, the significance test was judged to be correct. If the ordering was different, a type I error was recorded. Zobel examined the 1,830 pairs under four effectiveness measures, including eleven-point average precision and P@10, resulting in a total of 7,320 comparisons.

The significance tests assessed were ANOVA, Wilcoxon and the t-test. Zobel found all three to be accurate at predicting system ordering in the second set of topics: depending on the effectiveness measure used, between 97%-98% for the t-test and ANOVA, and 94%-98% for Wilcoxon. Significance via the t-test was observed in 3,810 pairs; in all but 4 of these pairs ANOVA also found significance. Significance via the Wilcoxon test was not observed in 14 of the 3,810 pairs found by t-test, but significance was observed in an additional 724 pairs.

Expanding on Zobel's topic-partitioning methodology, (Voorhees & Buckley 2002) examined a simple form of significance: measuring the absolute difference in MAP between two systems. Their aim was to determine the size of difference observed for the first set of topics before it was possible to be confident that system ordering would be preserved in the second topic set. The runs used were those submitted to the ad hoc track of TRECs 3-10. The total number of runs was 476; the number of pair-wise comparisons made was 16,678 (comparisons were restricted to those pairs of runs submitted to the same year of TREC). Voorhees & Buckley randomly split the 50 topics in each TREC year into two disjoint sets of 25 and computed errors rates for 20 bins of MAP differences, 0%-1%, 1%-2%, up to 19%-20%. The whole procedure was repeated 50 times to ensure that any random variation in topic selection was smoothed out.

Evaluation methodology, and particularly its statistical tests associated, plays a central role in the information retrieval domain which maintains a strong empirical tradition. Moreover, this scheme may be used to assert the accuracy of virtually any statistic, to build approximate confidence interval, and to verify whether a statistically significant difference exists between two retrieval schemes, even when dealing with a relatively small sample size. This study also suggests selecting the sample median rather than the sample mean in evaluating retrieval effectiveness where the justification for this choice is based on the nature of the information retrieval data.

Moreover, when evaluating a retrieval scheme, we assume that the following three hypotheses are always respected (Tague-Sutcliffe & Blustein, 1992), (Hull, 1993). (i) all queries included in a test collection are independent or not obviously related. (ii) all documents contained in a test collection are judged either relevant or irrelevant to a given request. (iii) each relevant record is equally important in satisfying the user's information need. Thus, the relevance of a given document does not depend on the number of relevant and already retrieved documents.

Of course, these assumptions are not really realistic, but they can be adopted as a first approximation. For example, we recently suggested a retrieval scheme based on the relationships between past queries in order to enhance the ranking of related future requests (Savoy, 1994), and the underlying assumption of this retrieval scheme clearly contradicts the first hypothesis. The second supposition implies that a "good" test collection must include a list of pertinent documents obtained, ideally through a manual inspection of all documents contained in the corpus. Moreover, the relevance assessments given by users are subjective, as reported by (Saracevic, 1975), (Schamber, 1994) and (Harter, 1996). Finally, the third hypothesis is not totally realistic because a user will naturally attach a greater utility to the first retrieved and relevant document than to the 25th. However, these criticisms related to the design of test collections, and particularly to the method used to obtain relevance judgments, cannot be a well grounded argumentation to invalidate retrieval experiments (Salton, 1992).

These three hypotheses lead us to define a retrieval effectiveness measure on the basis of both the number of relevant documents and the number of retrieved documents. Respecting these two criteria, the average precision at eleven standard recall values can be considered as a good retrieval effectiveness measure (Tague-Sutcliffe, 1992), (Salton, 1992), (Tague-Sutcliffe & Blustein, 1994), and this means is widely accepted throughout information retrieval literature. However, other approaches can also be considered, see (van Rijsbergen, 1979), and for a broader perspective about evaluation of IR systems, see (Saracevic, 1995).

3. EVALUATION METRICS OF MULTILINGUAL INFORMATION RETRIEVAL SYSTEM

3.1. Propose MLIR Metrics

In the proposed MLIR metrics, we used dictionary based query translation using word to word translation. On the whole we have used 2700 documents which include English, Hindi, German and French languages. Here we have considered English as the source language and French, German and Hindi as the target languages. The Google language translator is used for the Query translation.

Average Precision (AP_{MLIR}): Average of the precision values at the points at which each relevant documents retrieved in some user specific languages.

$$AP_{MLIR} = \sum_{l=1}^{n} \frac{\left\{\frac{\sum_{r=1}^{|N|}(P_{MLIR}(r).B(r))}{|D_+|}\right\}}{n}$$

Where r=rank of the query in 'n' languages, N=the number of retrieved documents in n languages, B(r) =binary function on the relevance of a given rank r in n languages, $|\mathbf{D}_+|$=number of relevant documents, $P_{MLIR}(r)$ =Precision at a cut of rank r in n languages.

$$P_{MLIR}(r) = \sum_{l=1}^{n} \frac{|\text{relevant retrieved documents of rank r or less}|}{r}$$

Mean Average Precision (MAP_{MLIR}): Average of the average precision values for a number of queries in MLIR system.

$$MAP_{MLIR} = \frac{\sum_{q=1}^{NQ} AP_{MLIR}(NQ)}{NQ}$$

Where NQ=number of queries

The overall design of our system, which consists of an engine, and two post-processing modules, that is, re-ranking and clustering modules. The engine retrieves documents in response to user queries, and outputs those documents in the source (user) language. The MLIR, both the source and translated queries are used to search a multi-lingual collection for relevant documents. Then, only retrieved documents that are not in the user language are translated into the user language. In principle, we need only the engine to realize MLIR in the sense that users can retrieve/ browse foreign documents through their native language. However, to improve the quality of our system, two alternative post-processing modules can optionally be used.

Average Means Reciprocal Rank: Average mean reciprocal rank is a statistic for evaluating any process that produces a list of possible responses to a query, ordered by probability of appropriateness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first accurate answer. The mean reciprocal rank is the average of the reciprocal ranks of results for the given number of queries Q_T,

$$AMRR = \frac{1}{|Q_T|}\left[\sum_{i=1}^{|Q_T|}\left\{\frac{\sum_{j=1}^{|T_n|}\frac{1}{rank}}{|N_{rr}|}\right.\right.$$

Where Q_T is the total number of queries, T_n is top n documents in the retrieved documents, $rank_j$ is j^{th} position rank in the retrieved documents and N_{rr} is the total number of relevant retrieved documents in $rank_j$.

Average Discounted Cumulative Gain (ADCG): The Average Discounted Cumulative Gain score (Jarvelin and Kekalainen, 2002) is a popular evaluator for multi-level relevance judgments. In its indispensable form it has a logarithmic position discount: the benefit of considering a relevant document at position j is $1/\log_2(1 + j)$. Following (Burges et al, 2005), it became usual to assign exponentially high weight $2^{rel}{}_j$ to highly rated documents where rel_j is the grade of the j^{th} document going for instance from 0-irrelevant to 1- perfect relevant result. Thus the DCG for a ranking position j of a query having D_n associated documents is defined as

$$ADCG_{MLIR} = \frac{1}{|Q_T|}\left[\sum_{i=1}^{|Q_T|}\frac{\left\{\sum_{j=1}^{|D_n|}\frac{2^{rel_j}-1}{log_2(1+j)}\right\}}{|RRD_i|}\right]$$

Where Q_T is the total number of queries, D_n is the top n retrieved documents, $rel_j \in \{0, 1\}$, 0-irrelevant document, 1-relevant document; RRD_i is relevant retrieved documents for each query.

Traditional evaluation methodology compares the average precision, mean average precision, average mean reciprocal rank and average discounted cumulative gain values to determine whether a search strategy is better, equal of worse than another. However, we may wish to establish that the difference in retrieval effectiveness under two conditions is statistically different or that the difference does not simply occur by chance. To achieve this goal, we may base our decision rule on either parametric or nonparametric tests instead of applying the described informal rule.

3.2. Statistical Significance Tests of MLIR system

The aim of statistical tests is to know whether or not the difference between two retrieval systems is truly significant or if this difference could have occurred by chance. The resulting judgment is strengthened (i) when the difference values are relatively high; or (ii) when these values are, more or less, always in favour of the same retrieval scheme; and (iii) when the sample size grows. A null hypothesis plays the role of a devils promoter, and we hope that the resulting statistic will lead us to reject this hypothesis. Under H0, each test computes a statistic T and calculates the achieved significance level of this test which is the probability of observing a value at least as extreme as T when the null hypothesis H0 is true. If this probability is less than a specified significance level a, we may conclude that the search schemes are significantly different. The "paired t-test" represents the first statistical test that we might consider, under the assumption that the difference follows a normal distribution. The formulation of the statistic T-test is described by Equation3 (Conover, 1980, pp. 290-292). However, even if the distribution of the observations is not normally shaped but if the empirical distribution is roughly symmetric; the t-test can be still a useful test because it is relatively robust in the sense that the indicated significance level is not far from the true a level. However, testing symmetry is a more complex procedure, e.g., (Antille et al., 1982).
The Wilcoxon Signed-Ranks test is based on the statistic T computed according to Equation4 (Siegel, 1956, pp. 75-83), (Conover, 1980, pp. 280-288). If the null hypothesis H0 is true, the values of T tend to be large, and small values of T indicate that H0 is false. Therefore, our decision rule is to reject H0 if $T < Za$, where the percentile Za follows a standard normal distribution N (0, 1). However, the approximation, N (0, 1), is valid only if N>20, which is the case in the current study. Finally, for both the Wilcoxon and Sign tests, ties (if any) are removed before the underlying statistics are computed, and the resulting size N is indicated in Tables. According to van Rijsbergen (1979), we know that the conditions required for the application of these tests are not really met in the information retrieval context.

The normal distribution of the dissimilarity between two retrieval systems does not follow a normal distribution in all circumstances, and parametric tests are of doubtful value. Nonparametric tests stipulate hypotheses that may not hold in the context of multilingual information retrieval analysis and other statistical models may be based on unrealistic assumptions. Moreover, the IR metrics does not represent the only measure available to quantify the difference between two retrieval systems, and we may also consider the median, a more robust location statistic.
Hypothesis testing: After obtaining a numerical value for the point estimator accuracy and building confidence intervals for them, we then wish to test the null hypothesis H0 or the validity of the assumption of identical two medians (or means). This hypothesis will be accepted if two retrieval systems return statistically similar performances, and rejected if not. Such comparison of treatments or effects represents the major objective of a retrieval experiment.

4. EXPERIMENTAL RESULTS OF STATISTICAL TESTS IN MLIR SYSTEM

In the Statistical Significance of experimental results shown in the following tables, in that analysis we are applied four tests to our proposed MLIR metrics, these tests are Independent sample T-test, Wilcoxon Signed Ranks Test, Sign Test and the metrics are AP, MAP, AMRR and ADCG. These test are applied in both the systems IR and MLIR and these systems are decided to accepted or rejected based on the null hypothesis of significant values according to Confidence Interval (CI) i.e.95%, i.e. the significance value (p-value ≤ 0.05) then the null hypothesis is rejected otherwise p-value > 0.05 then the null hypothesis is accepted in all the test of IR and MLIR system metrics.
Table1 Precision in IR and MLIR

4.1. Table1 Precision in IR and MLIR

Independent sample T-Test groups=groups (1 2) /criteria=ci (.9500).

Groups	Kolmogoroy-Smirnoy[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Values1	0.081	30	.200*	.958	30	.270
2	.129	30	.200*	.952	30	.193

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance

Non Parametric Test:

Wilcoxon Signed Ranks Test: Test Statistics[b]

	MLIR-IR
Z	-.936[a]
Asymp.Sig.(2-tailed)	.349

a. Based on negative ranks
b. Wilcoxon Signed Ranks Test

Sign Test: Test Statistics[a]

	MLIR-IR
Z	-2.373
Asymp.Sig.(2-tailed)	.018

a. Sign Test

4.2. Table 1 Recall in IR and MLIR

Independent sample T-Test

Groups	Kolmogoroy-Smirnoy[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Values1	.127	30	.200*	.921	30	.029
2	.073	30	.200*	.969	30	.514

a. Lilliefors Significance Correction
*. This is a lower bound of the true signific

Non Parametric Test

Wilcoxon Signed Ranks Test: Test Statistics[b]

	MLIR-IR
Z	-1.873[a]
Asymp.Sig.(2-tailed)	.061

a.Based on negative ranks
b.Wilcoxon Signed Ranks Test

Sign Test: Test Statistics[a]

	MLIR-IR
Z	-2.739
Asymp.Sig.(2-tailed)	.006

a. Sign Test

4.3. Table 2 AP in IR and MLIR

Independent sample T-TEST

Groups	Kolmogoroy-Smirnoy[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Values1	0.110	30	.200*	.961	30	.331
2	.197	30	.005*	.869	30	.002

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance

Non Parametric Test

Wilcoxon Signed Ranks Test: Test Statistics[b]

	MLIR-IR
Z	-4.247[a]
Asymp.Sig.(2-tailed)	.000

a. Based on negative ranks
b.Wilcoxon Signed Ranks Test

Sign Test: Test Statistics[a]

	MLIR-IR
Z	-4.930
Asymp.Sig.(2-tailed)	.000

a. Sign Test

4.4. Table 3 MRR in IR and MLIR

Independent sample T-TEST

Groups	Kolmogoroy-Smirnoy[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Values1	0.134	30	.182*	.932	30	.057
2	.152	30	.075*	.922	30	.029

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance

Non Parametric Test:

Wilcoxon Signed Ranks Test: Test Statistics[b]

	MLIR-IR
Z	-3.383[a]
Asymp.Sig.(2-tailed)	.001

a. Based on negative ranks
b.Wilcoxon Signed Ranks Test

Sign Test: Test Statistics[a]

	MLIR-IR
Z	-2.739
Asymp.Sig.(2-tailed)	.006

a. Sign Test

4.5. Table 4 DCG in IR and MLIR

Independent sample T-TEST

Groups	Kolmogoroy-Smirnoy[a]			Shapiro-Wilk		
	Statistics	df	Sig.	Statistics	df	Sig.
Values1	.126	30	.200*	.936	30	.069
2	.097	30	.200*	.976	30	.707

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance

Non Parametric Test:

Wilcoxon Signed Ranks Test: Test Statistics[b]

	MLIR-IR
Z	-4.679[a]
Asymp.Sig.(2-tailed)	.000

a. Based on negative ranks
b.Wilcoxon Signed Ranks Test

Sign Test: Test Statistics[a]

	MLIR-IR
Z	-4.930
Asymp.Sig.(2-tailed)	.000

a. Sign Test

Table 1. Calculated values of Precision and and Recall

Table 2. Calculated values of AP

Query No.	**Precision$_{IR}$**	**Precision$_{MLIR}$**	**Recall$_{IR}$**	**Recall$_{MLIR}$**
1	0.3910	0.4615	0.21	0.30
2	0.2372	0.3571	0.16	0.25
3	0.4054	0.4706	0.32	0.40
4	0.4203	0.6111	0.39	0.55
5	0.5906	0.7059	0.43	0.60
6	0.6012	0.7368	0.49	0.70
7	0.5211	0.6957	0.32	0.48
8	0.4718	0.6098	0.29	0.50
9	0.4728	0.7528	0.29	0.67
10	0.5216	0.6988	0.41	0.58
11	0.6973	0.8934	0.51	0.69
12	0.7152	0.8152	0.53	0.75
13	0.2235	0.4375	0.12	0.35
14	0.2753	0.3571	0.13	0.20
15	0.5076	0.7067	0.37	0.53
16	0.4933	0.6333	0.21	0.38
17	0.6108	0.7180	0.42	0.56
18	0.5932	0.7959	0.43	0.78
19	0.3824	0.5844	0.26	0.45
20	0.2134	0.3181	0.15	0.28
21	0.9210	0.4568	0.89	0.12
22	0.3988	0.5741	0.64	0.09
23	0.8125	0.6128	0.73	0.64
24	0.7213	0.5123	0.12	0.38
25	0.6699	0.2416	0.34	0.42
26	0.9122	0.2341	0.69	0.34
27	0.2344	0.0274	0.15	0.10
28	0.5366	0.0910	0.34	0.82
29	0.2248	0.0713	0.91	0.36
30	0.4433	0.9192	0.46	0.12

Query num	AP_{IR}	AP_{MLIR}
1	0.8153	0.2614
2	0.7423	0.1370
3	0.3686	0.2584
4	0.8251	0.6076
5	0.2601	0.8809
6	0.5455	0.2359
7	1.0000	0.8466
8	0.7122	0.4765
9	0.9818	0.0957
10	0.6851	0.4378
11	0.7821	0.0289
12	0.5236	0.1264
13	0.6523	0.0928
14	0.9153	0.1583
15	0.8912	0.5546
16	0.3584	0.0729
17	0.4129	0.1030
18	0.5329	0.1158
19	0.1289	0.0728
20	0.3195	0.1349
21	0.7342	0.5246
22	0.9144	0.6137
23	0.6178	0.3482
24	0.4692	0.0158
25	0.3492	0.0617
26	0.8231	0.8001
27	0.6928	0.2068
28	0.8311	0.4309
29	0.8030	0.1439
30	0.5210	0.2169
MAP	**0.6403**	**0.1007**

5. CONCLUSIONS

This paper was to evaluate the retrieval effectiveness of search systems i.e. IR and MLIR and to form a firm theoretical basis for comparing retrieval schemes. After reviewing traditional statistical tests used in multilingual information retrieval system, we suggest rejecting the paired t-test to verify whether or not a retrieval scheme is better than another, because the underlying distribution of the data does not always follow a normal distribution. Moreover, since the hypotheses underlying nonparametric tests are not always strictly respected in the multilingual information retrieval system, this system suggests to both analyse the performance of an IR and MLIR retrieval mechanisms and to compare two search strategies and concludes that the Independent T-Test is reliable compare to the Non-parametric tests in MLIR system.

Table 3. Calculated values of MRR

Query no	*MRR-IR*	*MRR-MLIR*
1	0.1207	0.1314
2	0.1327	0.0963
3	0.1916	0.1138
4	0.1658	0.1987
5	0.0945	0.1829
6	0.3251	0.0958
7	0.2072	0.2126
8	0.1568	0.2728
9	0.2449	0.0419
10	0.2535	0.1994
11	0.1385	0.6571
12	0.9196	0.2889
13	0.8735	0.569
14	0.9839	0.1855
15	0.6614	0.7431
16	0.6255	0.5747
17	0.9004	0.3384
18	0.3913	0.3652
19	0.4491	0.1259
20	0.0283	0.0179
21	0.4690	0.2158
22	0.4636	0.2273
23	0.4917	0.3212
24	0.5098	0.1296
25	0.4274	0.2961
26	0.6019	0.1324
27	0.7823	0.3413
28	0.6928	0.0023
29	0.9412	0.5241
30	0.8125	0.4545
AMRR	**0.4686**	**0.2685**

Table 4. Calculated values of DCG

Query no	*DCG-IR*	*DCG-MLIR*
1	0.2909	0.2012
2	0.3043	0.1034
3	0.3549	0.3041
4	0.3354	0.1454
5	0.2754	0.0742
6	0.4692	0.296
7	0.3735	0.1753
8	0.3108	0.018
9	0.4039	0.1043
10	0.4075	0.1275
11	0.4498	0.2395
12	0.4889	0.2451
13	0.5120	0.2918
14	0.8103	0.1721
15	0.6102	0.1981
16	0.7102	0.1324
17	0.6124	0.1636
18	0.5120	0.3529
19	0.8024	0.3576
20	0.3941	0.1329
21	0.5421	0.2486
22	0.8912	0.3471
23	0.7412	0.1203
24	0.3946	0.3018
25	0.4168	0.3201
26	0.7123	0.2719
27	0.6423	0.0342
28	0.2861	0.4296
29	0.4628	0.0917
30	0.6127	0.2461
ADCG	**0.5043**	**0.2082**

REFERENCES

[1] Sakai, 2006b Sakai, T. (2006b). Give me just one highly relevant document: P-measure. In Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2006).

[2] Buckley and Voorhees, 2004 Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004) (pp. 25–32).

[3] Voorhees, 2005 Voorhees, E. M. (2005). Overview of the TREC 2004 robust retrieval track. In *Proceedings of the 13th text retrieval conference (TREC 2004)*.

[4] Sanderson and Zobel, 2005 Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 162–169).

[5] Sakai, 2005a Sakai, T. (2005a). The effect of topic sampling on sensitivity comparisons of information retrieval metrics. In *Proceedings of the 5th NTCIR workshop on research in information access technologies (NTCIR-5)*.

[6] Sakai, 2006a Sakai, T. (2006a). Evaluating evaluation metrics based on the bootstrap. In Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2006).

[7] Sakai, 2004 Sakai, T. (2004). Ranking the NTCIR systems based on multigrade relevance. In *Proceedings of Asia information retrieval symposium 2004* (pp. 170–177).

[8] Vu and Gallinari, 2005 Vu, H.-T., & Gallinari, P. (2005). On effectiveness measures and relevance functions in ranking INEX systems. In *Proceedings of Asia information retrieval symposium 2005. Lecture notes in computer science: Vol. 3689* (pp. 312–327).

[9] Jacques Savoy: Statistical inference in retrieval effectiveness evaluation. Inf. Process. Manage. 33(4): 495-512 (1997).

[10] Sparck Jones, K., Van Rijsbergen, C.J. (1975) Report on the need for and provision of an 'ideal' information retrieval test collection, British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.
[11] Voorhees, E.M., Harman, D. (1999) Overview of the 8th Text REtrieval Conference (TREC-8), in Proc. 8th Text REtrieval Conference.
[12] Van Rijsbergen, C.J. (1979) Information Retrieval, London: Butterworths.
[13] Dunlop, M.D. (1997) Time Relevance and Interaction Modeling for Information Retrieval, in Proc. ACM SIGIR, 206-213.
[14] Jarvelin, K. & Kekalainen, J. (2000) IR evaluation methods for retrieving highly relevant documents, in Proc. ACM SIGIR, 41-48.
[15] Buckley, C., Voorhees, E.M. (2004) Retrieval evaluation with incomplete information, in Proc. ACM SIGIR, 25-32.
[16] Hull, D. (1993) Using statistical testing in the evaluation of retrieval experiments, in Proc. of ACM SIGIR, 329-338.
[17] Savoy, J. (1997) Statistical inference in retrieval effectiveness evaluation, Information Processing & Management, 33(4):495-512.
[18] Matthews, R. (2003) the numbers don't add up, New Scientist, March, p. 28, issue 2385.
[19] Zobel, J. (1998) how reliable is the results of large-scale information retrieval experiments? In Proc. ACM SIGIR, 307-31.
[20] Voorhees, E.M., Buckley, C. (2002), the effect of topic set size on retrieval experiment error, in Proc. ACM SIGIR, 316-323.
[21] Tague-Sutcliffe, J., Blustein (1994) A Statistical Analysis of the TREC-3 Data, in Proc. TREC-3, 385-398.
[22] Saracevic, T. (1975). Relevance: A review of and a framework for thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321-343.
[23] Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3-48.
[24] Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science 47(1), 37-49.
[25] Salton, G. (1992). the state of retrieval system evaluation. Information Processing & Management, 28(4), 441-449.
[26] Tague-Sutcliffe, J. (1992). the pragmatics of information retrieval experimentation, revised. Information Processing & Management, 28(4), 467-490.
[27] Tague-Sutcliffe, J., & Blustein, J. (1994, November). A statistical analysis of the TREC-3 data. Proceedings of the 3rd Text REtrieval Conference TREC'3, Gaithersburg, MD, 385-398.
[28] Saracevic, T. (1995, July). Evaluation of evaluation in information retrieval. Proceedings of the 18th International Conference of the ACM-SIGIR'95, Seattle, WA, 137-146.
[29] Conover, W. J. (1980). Practical nonparametric statistics. 2nd ed., New-York, NY: John Wiley & Sons.
[30] Antille, A., Kersting, G., & Zucchini W. (1982). Testing symmetry. Journal of the American Statistical Association, 77(379), 639-646.
[31] Siegel, S. (1956). nonparametric statistics for the behavioral sciences, New-York, NY: McGraw-Hill.