Received 10 December 2014
Revised 13 May 2015
Accepted 18 May 2015

Document-based approach to improve the accuracy of pairwise comparison in evaluating information retrieval systems

Sri Devi Ravana, Masumeh Sadat Taheri and Prabha Rajagopal
Department of Information System, University of Malaya, Kuala Lumpur, Malaysia

Abstract

Purpose – The purpose of this paper is to propose a method to have more accurate results in comparing performance of the paired information retrieval (IR) systems with reference to the current method, which is based on the mean effectiveness scores of the systems across a set of identified topics/queries.

Design/methodology/approach – Based on the proposed approach, instead of the classic method of using a set of topic scores, the documents level scores are considered as the evaluation unit. These document scores are the defined document's weight, which play the role of the mean average precision (MAP) score of the systems as a significance test's statics. The experiments were conducted using the TREC 9 Web track collection.

Findings – The *p*-values generated through the two types of significance tests, namely the Student's *t*-test and Mann-Whitney show that by using the document level scores as an evaluation unit, the difference between IR systems is more significant compared with utilizing topic scores.

Originality/value – Utilizing a suitable test collection is a primary prerequisite for IR systems comparative evaluation. However, in addition to reusable test collections, having an accurate statistical testing is a necessity for these evaluations. The findings of this study will assist IR researchers to evaluate their retrieval systems and algorithms more accurately.

Keywords Information retrieval, Document-based evaluation, Information retrieval evaluation, Pairwise comparison, Significance test

Paper type Research paper

1. Introduction

To date, an overwhelming large number of evaluation approaches appraise the accuracy and effectiveness of an information retrieval (IR) system by using Cranfield paradigm, which is a system-based evaluation method. An overview of a system-based experiment using a test collection is depicted in Figure 1. It is worth noting that in large-scale IR evaluation experimentation, the researchers use system-based evaluation approaches instead of user-based evaluation method due to the high number of human participants and retrieval systems required. Such requirements make the user-based evaluation method time consuming and costly (Moghadasi *et al.*, 2013). Besides, it would require a controlled environment in running the experiments and a very carefully designed experiment (Voorhees, 2002). The system-based evaluation method deploys a test collection which includes a document corpus, a batch of predefined users' information requests, known as queries, and the relevancy judgments pointing out which document is related to what topic (Carterette *et al.*, 2006; Moghadasi *et al.*, 2013).

This research is supported by UMRG Program RP028E-14AET. Also, this work was supported by the Exploratory Research Grant Scheme (ERGS) ER027-2013A.

Aslib Journal of Information Management
Vol. 67 No. 4, 2015
pp. 408-421

2050-3806
DOI 10.1108/AJIM-12-2014-0171

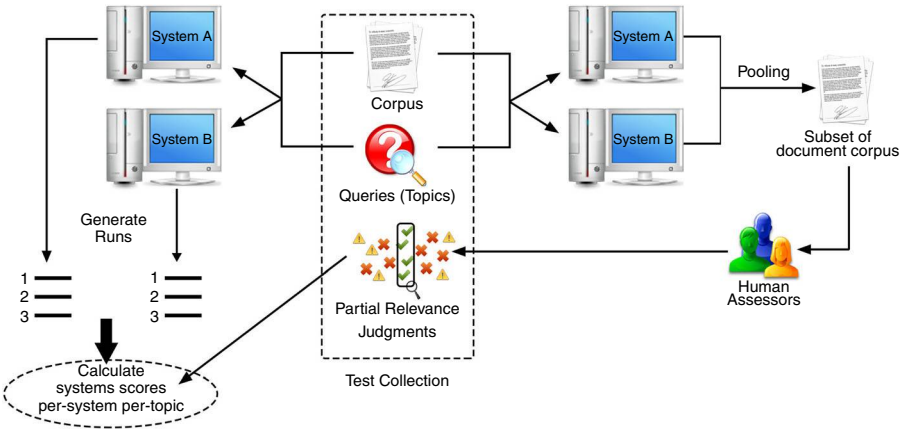

Figure 1.
Schematic view of the classic TREC retrieval evaluation

The Text REtrieval Conference (TREC) prepares the essential infrastructure for voluminous evaluation of text retrieval systems; namely, reusable test collections in couple with the large sets of relevance judgments. According to the time consuming and costly part of preparing relevance judgments that should be carried out by human assessors, judging the whole document corpus is hardly possible. Hence, pooling method is used through which each query is submitted to the participant IR systems, also known as runs, and a set of top k retrieved documents by each run is chosen for further assessments. After all, the IR systems effectiveness is defined as the ability of retrieving most relevant documents to a user's query (Carterette and Voorhees, 2011) which is evaluated over a set of topics via expedient evaluation measures. However, the existence of intrinsic noise such as different levels of difficulty for the chosen topics and documents corpus as well as human errors as part of the assessors' judgments in the evaluation is not ignorable. Hence, using statistical significance tests plays a main role in detecting IR algorithms or systems which truly have better performance rather than by chance. Significant improvements can be observed via a powerful and accurate statistical test, even if that improvement is small (Smucker *et al.*, 2007).

Regularly, the difference in the mean score of an IR metric, typically, average precision, is used by researchers as a test statistic for significance tests. This is because it is considered that the overall performance of a system is not detectable by examining only one topic (Hull, 1993). However, summing up the topic scores via each and every metric such as mean average precision (MAP) does not always lead to an accurate effectiveness evaluation result when comparing systems executed on the distinct topics due to two main reasons. First, according to various studies, different topics have different grades of difficulty and each topic has different impact on document retrieval process (Harman and Buckley, 2004; Webber *et al.*, 2008a). Second, incurring data loss caused by the inherent trait of aggregation techniques (Bendat and Piersol, 2011).

In this paper, instead of considering topics as units of evaluation, IR systems are evaluated from the document level with assumption that both concerns of difficulty of varying topics; and loss of information due to aggregation of topic scores are addressed. In order to score the retrieved documents, for each topic of a system, we measured the probability of a retrieved document being relevant in a retrieval

process. This is done by counting the retrieval frequency of a document in couple with their rank per-system per-topic. Hence, having this assumption that we have a total of ten systems, if for topic t, document D is retrieved by all the ten systems, the probability for document D to be relevant to topic t is one ($p = 1$). On the other hand, if only five systems retrieve document D, the probability for document D to be relevant is half ($p = 0.5$). In this way, each document for a topic is given a value or relevance weight that is then used to conduct a significance test to compare the effectiveness difference between a pair of retrieval systems. This document weight indicates the effectiveness of the systems in retrieving the relevant documents to a specific topic and how these systems retrieve these documents early.

The rest of this paper is organized as follows. The next section provides an overview on the existing literature on evaluation of IR systems in general and specifically by pairwise comparisons method. Section 3 explains the processes involved in the pairwise comparisons, and the significance tests used in the IR systems evaluation experimentation. Section 4 outlines the experimental design and the test environment used in this paper, in couple with the obtained results and related discussions, while the final section proposes the ways to extend the work for our future work.

2. Methods used for evaluating IR systems

In the last 20 years, IR analysis procedure has been experiencing a noticeable development. Re-usable, high quality test collections mainly created by TREC, in couple with correlation coefficient of system rankings as well as utilizing significance tests to distinguish the real improvements of IR algorithms from those acquired by chance, are three main underlying reasons for these developments (Carterette *et al.*, 2006; Cormack and Lynam, 2007; Sakai, 2006; Sanderson and Zobel, 2005; Smucker *et al.*, 2007, 2009). Accordingly, this section is mainly focussed on these three categories as discussed in the following.

2.1 Scoring of system effectiveness

In a test collection experiment, various aggregation methods such as geometric, arithmetic, harmonic mean (Robertson, 2006) and median are used to obtain an IR system effectiveness score by aggregating the system's topic scores.

Revealing another approach, Sanderson and Zobel (2005) investigated the reliability of a system's performance measurement while utilizing a large number of topics. While investigating the compatibility of evaluation measures for evaluating significance tests, they also examined P@10 and MAP. They found that creating test collections which includes more topics rather than relevance judgments is more practical than considering more documents for assessment when evaluate systems by fewer number of queries.

Other approach is to enhance the progress of comparability of different topic scores acquired from different systems. To achieve this aim, Zobel (1998) tried to divide the scores of each system by the highest score which is obtained by either system considering a same topic to normalize the metric scores of the systems. Meanwhile, Jarvelin and Kekalainen (2002) suggested that only considering the obtained highest score for normalizing the scores is not adequate but instead the highest achievable score should be noted, having the admitted distribution of relevance.

The other researchers, Buckley and Voorhees (2000) and Sanderson and Zobel (2005) for evaluating IR measurement metrics, randomly divided a topic set to examine how

many times system pairs are ordered adversely by the obtained results. Consequently, they were enabled to measure the metrics error. Buckley and Voorhees (2000) measure the mean error rate for variant system score deltas and different topic set sizes over TREC systems. Their goal was to admit that an average AP delta of 0.06 is at least 90 percent dependable on 50 topics. In contrast with the statistical analysis, their proposed approach did not consider the variability of score deltas. Hence, it needs to assume that the obtained results from the previous TREC systems are also suitable for new collections and systems. On the other hand, Sanderson and Zobel (2005) used statistical significance in couple with absolute deltas while measuring the error rates.

Mizzaro and Robertson (2007) intended to normalize the achieved topic scores per-system by deducting the mean achieved score for a specific system or topic. However, they did not modify their method for variance.

2.2 Correlation coefficient of system rankings

The relationship between variables under evaluation is measured by correlation coefficients. Similar to significance tests, they can be divided into two groups, parametric tests and the nonparametric tests (30). Correlation coefficient methods in IR test collection based experiments are mostly used to measure the similarity of the two system rankings. Pearson, Kendall's (1938) τ and Spearman rank correlation coefficient (Wackerly *et al.*, 2007) are the three rank correlation statistics which are usually used by researchers. Among these tests, Kendall's τ has become a standard statistic, the so-called "gold standard," to compare the correlation between two system runs. When the correlation between the generated runs and the gold standard is high, that system can be judged as a better system. Soboroff *et al.* (2001) introduced a method for evaluating IR systems without relevance judgements. They measured the quality of their proposed method by using Kendall's τ. Another evaluation measure, bpref, was proposed by Buckley and Voorhees (2004). With the help of Kendall's τ, they showed that ranking the systems by their measure is very close to average precision. Similar examples of Kendall's τ usage in IR can be found in Aslam *et al.* (2003), Melucci (2007) and Voorhees (2003).

In addition to comparing system rankings, correlation coefficient methods have been used to observe the correlation between system effectiveness and users satisfaction (Al-Maskari *et al.*, 2007, 2008).

Researchers have put some efforts to explore new correlation coefficient methods in the experiments. For instance, an AP-based correlation, τ_{AP}, was proposed by Yilmaz *et al.* (2008) which penalizes the errors which occurs among ranked documents at the high place of the list and at the bottom of the list differently.

Recently, a similarity measure, rank-biased overlap (RBO), was proposed by Webber *et al.* (2010) by the help of which indefinite rankings can be handled.

2.3 Significance tests for pairwise comparison

Recently, in addition to assessing the effectiveness of IR systems by utilizing various measurement metrics, statistical tests have been executed to verify the significance of the differences in system effectiveness measured (Ravana, 2011). To determine that an obtained score for an IR system is significant is as important as to distinguish the superiority of system A over system B on a predefined metric and collection.

Zobel (1998) empirically evaluated the effectiveness of statistical significance tests such as *t*-test, Wilcoxon, Sign test, Bootstrap and ANOVA. He found that the two halves of a topic set are highly expected to have same significance findings. The rate of the *t*-test

confirmation at significance level of 0.05 was considered to be 0.97 and 0.98. *t*-Test is found to be more accurate than the other two Sign and Wilcoxon tests. Besides, it is proved that the Bootstrap and *t*-test produce approximately identical outcomes. The theoretical foundation of hypothesis experiments in IR context was examined by Savoy (1997). He proposed to account the Bootstrap statistical test.

Likewise, the bootstrap test is suggested by Sakai (2006) through which a metric distinction can be determined by the number of system pairs which are significantly different under the test hypothesis. Meanwhile, an entropy analysis is proposed by Aslam *et al.* (2005) to determine a metric quality. The better metric is the one that can provide more information about its obtained score.

Moreover, Smucker *et al.* (2007) who compared the Wilcoxon, Sign, Bootstrap, Randomization permutation and *t*-test and proposed the randomized permutation test as requiring less assumptions. In addition, he demonstrated that the Sign and Wilcoxon tests are unreliable but the other tests produce almost identical results.

Assessing the correlation between evaluation metrics and user experience is an alternative method of evaluating metrics. Huffman and Hochster (2007) discovered that the three top documents or indeed the very top ranked documents relevance correlates more or less greatly with the satisfaction of human judges. This result is obtained in spite of utilizing qualified assessors for analyzing the information essentials as well as users satisfaction who offered the queries. Conversely, just a little correlation between users' satisfaction and most metrics is found by Al-Maskari *et al.* (2007).

Alternatively, two distinct tasks were given to the users by Turpin and Scholer (2006). First was to find one relevant document in the minimum time (precision) and second, was to find as many related document as feasible in only five minutes (recall). They found that there is no significant correlation between a user performance and the average precision score of a system in the first task and just an unsteady correlation on the latter task.

Previous studies have relied on two-sample paired comparison tests, such as the paired *t*-test, the sign test, or the Wilcoxon matched-pairs test. Of these approaches, only the sign test is certain to be valid when applied to the standard IR measures. However, we can examine the validity of the *t*-test and the Wilcoxon with a few simple diagnostic data plots. In addition, the paired-comparison tests can be generalized to more than two samples, which can be useful for simultaneously analyzing a large number of different retrieval methods.

To the best of our knowledge, our proposed approach is the first effort in IR research community to examine the significance tests from this point of view, which is considering document level instead of topic level to find the significant difference in evaluating the effectiveness of IR systems.

3. Pairwise comparisons experimentation

Pairwise systems comparisons

Considering two IR systems, namely system A and system B which are to be scored and compared utilizing the traditional method and test collection. Each IR system runs several predefined queries and for each query retrieves a ranked list of documents that it considers as relevant to that specific query. The retrieved documents evaluated for relevancy utilizing the relevance judgments for the queries and an evaluation metric is used to produce an effectiveness score for that system. If $s.t_i$ denotes the system's score on topic t_i, for $i = 1, \ldots, m$, where m is the number of considered topics, the mean score s_A for system A is $\sum_i s.t_i/n$, where n is the number of queries and it is also true for system B. The difference between two systems mean scores, $\overline{s_A} - \overline{s_B} = d_{A,B}$ which indicates the two systems observed delta. The per-topic delta for each topic t is $d_t = s_{A,\,t} - s_{B,\,t}$. Hence,

$d_{A,\,B} = \sum_t (d_t)/(n)$. Having $d_{A,\,B} > 0$, it can be considered that system A performs better than system B on topic set t, at least in favour of the selected metric. However, a verification of obtaining a real difference between systems is necessary since this result may have produced by chance (Carterette, 2012).

Repeating the experiments in an IR environment to explore confidence intervals on reliability of the obtained results from test scores is not useful. This is because IR systems are inflexible and same test conditions always lead them to set up the same results. In addition, diversifying the document corpus or topic set is also useless, since these have a direct impact on the system's retrieval behaviour (Buckley and Voorhees, 2000).

Significance test

A significance test is used to verify how likely it is that the obtained results of retrieval effectiveness difference between a System A and a System B has emerged by chance rather than a real superiority. The output of a significance test is a score known as *p*-value. Obtaining smaller *p*-values leads to verify that it is more likely to have a significant difference between systems. Normally, 0.05 and 0.01 values are considered as significance tests *p*-value's thresholds for demonstrating the systems differences are statistically significant. In addition to *p*-values, formulating a null hypothesis *H0* which indicates that the two systems have equivalent effectiveness encompasses a significance testing. If the *p*-value is below the defined threshold, then *H0* is rejected and the substitute hypothesis which indicates that the two systems are not equal is accepted.

Generally, paired or correlated statistical tests are more desired in the evaluations of IR systems. Choosing the tests depends on general principles (Carterette and Voorhees, 2011). Table I shows some of the statistical tests recommended for different class of statistics and test categories.

As it is discussed, there are a variety of significance tests in IR evaluation environment. In this paper, two tests, namely Student's paired *t*-test and Mann-Whitney test are selected for the experiments. The underlying reasons for this selection are more discussed in Section 4.

Student paired t-test

IR researchers, mostly, utilize the Student's paired *t*-test, also called as *t*-test, in their experimentations. This statistical test is used by pairing the obtained results vs the *t* distribution quantiles which is the distribution of the mean normal variable sampling (Webber *et al.*, 2008b). This test is widely used when the population has a normal distribution. The following equation is formulated and the *t* score as:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \sqrt{n}\,\frac{(\overline{x} - \mu_0)}{s} \tag{1}$$

	Parametric	Non-parametric
Distribution	Normal	Any
Dependent groups	Student's paired *t*-test	Sign test, Wilcoxon signed-rank test, Binomial test
Independent groups	Student's Independent (unpaired) *t*-test	Mann-Whitney test, Wilcoxon rank-sum test

Table I. Different statistical tests classification

where sample size is demonstrated by n, $\overline{x}$ is the mean of the n observations, the hypothesized population mean is designated by μ_0, and the sample standard deviation is measured by the following equation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2}$$

After computing the t score, it is compared to a predefined p-value threshold to examine the result of the significance test.

Mann-Whitney test

The Mann-Whitney test, which is also known as Wilcoxon rank-sum test, is another statistical test to examine the mean of two independent populations. It is a nonparametric complement of paired t-test and widely used when the sample data is not normally distributed (Carbno, 2007). It is worth to note that the Mann-Whitney test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and include aggregation of ranks. Measuring a statistic, normally known as U, is the main part of this test. The distribution of U under null hypothesis is clear. The test is applied differently on the populations based on their small or large sample. Having small samples, the U statistic is calculated from the following equations:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \tag{3}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2 \tag{4}$$

where n is the rank of comparing groups data, W_1 is the sum ranks of values from group 1 and W_2 is the sum ranks of values from group 2. The smallest of the obtained U values produce the final test statistics as comes in the following equation (30):

$$U' = n_1 \times n_2 - U \tag{5}$$

4. Experimental design and results

The proposed method is experimentally evaluated on TREC 9 Web track collection, where 50 topics 451-500 are used and the overall number of documents is 70,070. This TREC data sets statistics is listed in Table II. There are 105 systems that participated in this TREC, but, after the initial phase of the experiment, data cleaning, just 103 systems are considered for the rest of the analysis. This selection is taken part due to data inconsistency in the discarded runs. For instance, the number of retrieved documents for all topics was beneath

Table II. Summary of TREC 9 web track statistics

TREC-9 Web track Document corpus	VLC2, WT10g
Number of participated runs	105
Overall number of documents	70,070
The mean number of relevant documents over 50 topics	52.34

our experiment expectations. Hence, the total number of paired wise comparisons of the 103 systems will be $103 \times 102/2 = 5{,}253$. Three varying depths for retrieved documents were considered in this experiment with depth, $k = 100$, 500 and 1,000. Each depth implies the number of documents to be selected for the experiments. For instance depth, $k = 100$ indicates that the top 100 retrieved documents will be selected.

In this study, the main aim is to alleviate the concern of having more reliable paired comparison of retrieval systems by considering the role of each document which is retrieved by each system for each topic. Instead of considering the topic scores (i.e. aggregation of precision scores of a number of retrieved relevant documents per topic) for each system, the frequency of each retrieved document per-topic per-system in couple with its rank in that particular system is examined in the paired evaluation.

As depicted in Table III, F_{ij}, a matrix for frequency of retrieved documents, associated to retrieved documents per-system (s_i) per-topic (t_j), for $i = 1, \ldots, n$, where n is the number of systems and $j = 1, \ldots, m$, where m is the number of considered topics. Each cell of the matrix contains the retrieval frequency of each document per-system $df_{d_l t_j}$, which defines the number of times that document d_l , where l in d_l is the documents number, has been retrieved across all 103 evaluated systems, as well as their associated rank r_l in the corresponding system, where l in r_l is the document's retrieval order per-system per-topic. In this study, the number of retrieved documents per-system is conceded to be a fixed number of k, where k is the top 100, 500 and 1,000 retrieved documents per-system per-topic.

Our assumption is, as $df_{d_l t_j r_l}$ value for each document increases, the probability of that document to be relevant becomes higher. In other words, if a document is retrieved by a high number of systems, most likely it is a relevant document to the considered topic. Accordingly, a system that gives higher rank to a retrieved document with higher frequency is superior to a system which gives lower rank to that particular document. Here, superior is defined in terms of the effectiveness of a system in retrieving a relevant document earlier than other systems. For instance, in comparison of a system A and a system B, where system A assigns rank r_3 to a document with frequency value of 98 (out of 103), its retrieval performance is superior to system B which assigns rank r_{12} to that particular document.

To quantify the relevancy of documents in the document-level pairwise comparisons, the $df_{d_l t_j r_l}$ value of each document, which is retrieved by system, s_n, is divided by its corresponding rank r_l to get the documents weight, as given in DW formula in the following equation:

$$DW(d_l s_n) = \frac{df_{d_l t_j r_l}}{r_l} \tag{6}$$

For instance, assuming that there is a total five systems and each system retrieves five documents. The retrieval frequency of each document per system for one given topic in

Retrieved document	Systems S_1	S_2	S_3	…	S_n
d_1	$df_{d_1 t_j r_1}$	$df_{d_1 t_j r_1}$	$df_{d_1 t_j r_1}$	…	$df_{d_1 t_j r_1}$
d_2	$df_{d_2 t_j r_2}$	$df_{d_2 t_j r_2}$	$df_{d_2 t_j r_2}$	…	$df_{d_2 t_j r_2}$
d_3	$df_{d_3 t_j r_3}$	$df_{d_3 t_j r_3}$	$df_{d_3 t_j r_3}$	…	$df_{d_3 t_j r_3}$
⋮	⋮	⋮	⋮	⋮	⋮
d_l	$df_{d_l t_j r_1}$	$df_{d_l t_j r_1}$	$df_{d_l t_j r_1}$	…	$df_{d_l t_j r_1}$

Table III. Matrix for frequency of retrieved documents

couple with their retrieval rank in each system is illustrated in Table IV. For example, $9_{d_1 t_{451} r_1}$ in the first cell indicates that the document frequency of d_1 in system s_1 for topic 451 is nine and its rank in s_1 is one. In other words, document d_1 is retrieved by nine systems out of ten and its rank in system s_1 is one. Now, applying Equation (6), document weight for d_1 is:

$$DW(d_l s_1) = \frac{9}{1} = 9 \tag{7}$$

By considering the documents weight as a test statistic for significance tests, we can alleviate the disadvantages of aggregating the obtained data in IR systems comparison.

A topic-by-topic comparison of five retrieval systems from TREC-9 based on effectiveness scores using average precision as the evaluation metric with depth of top 1,000 retrieved documents is illustrated in Figure 2.

In the proposed document-level pairwise system evaluation approach, before applying a significance test to a pairwise comparison, the two systems that are under observation will be compared together in order to find their common documents. The documents with equal document name and topic number are considered as common documents. In essence, this allows us to work on the same parameters of a population, in order to use

Table IV.
Example of retrieved documents frequency matrix

Retrieved document	Systems S_1	S_2	S_3	S_4	S_5
d_1	$9_{d_1 t_{451} r_1}$	$9_{d_1 t_{451} r_5}$	$9_{d_1 t_{451} r_3}$	$9_{d_1 t_{451} r_1}$	$9_{d_1 t_{451} r_2}$
d_2	$5_{d_2 t_{451} r_2}$	$5_{d_2 t_{451} r_1}$	$5_{d_2 t_{451} r_1}$	$5_{d_2 t_{451} r_4}$	$5_{d_2 t_{451} r_1}$
d_3	$8_{d_3 t_{451} r_4}$	$8_{d_3 t_{451} r_2}$	$8_{d_3 t_{451} r_5}$	$8_{d_3 t_{451} r_2}$	$8_{d_3 t_{451} r_3}$
d_4	$3_{d_4 t_{451} r_3}$	$3_{d_4 t_{451} r_4}$	$3_{d_4 t_{451} r_4}$	$3_{d_4 t_{451} r_3}$	$3_{d_4 t_{451} r_4}$
d_5	$6_{d_5 t_{451} r_5}$	$6_{d_5 t_{451} r_3}$	$6_{d_5 t_{451} r_2}$	$6_{d_5 t_{451} r_5}$	$6_{d_5 t_{451} r_5}$

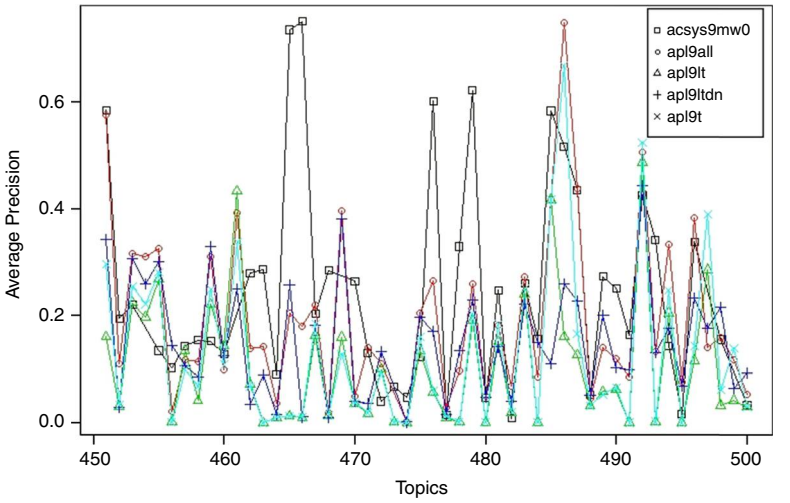

Figure 2.
Topic by topic comparison of five TREC-9 retrieval systems

paired *t*-tests in the document-level experiments. Besides, our observations for the two retrieval systems are assumed independent, since the topic scores are considered as arbitrary samples from the population of all captivating queries (Ravana, 2011). With these in mind, paired *t*-test is chosen as a significance test to evaluate the effectiveness of the systems in the pairwise comparison experiment. However, Mann-Whitney test is also utilized in the experiments in order to examine a probable difference in the power of the tests in determining the significant difference between systems.

In Figure 3, the set of obtained *p*-value results from a *t*-test on 103 systems (i.e. 102 pairs) at the 0.01 confidence level for one topic (topic = 451) is illustrated in a closer look. A randomly selected system (here it is input.acsys9mw0) against the other 102 systems composes the system pairs in this figure. A plotted point depicts the *p*-value generated from the *t*-test done on a system pair. Hence, there are 102 plotted points representing 102 *p*-values. The density of plotted points is mostly under 0.4 and only a few of them are below 0.01. Consequently, these results show that only a few pairs of systems are significantly different and can be considered as separable systems at significance grade of 1 percent.

To compare the systems' rankings which are acquired using our proposed method, we randomly selected five systems out of 103 systems in TREC-9 Web track comprising ten system pairs. Figure 4 shows the distribution of the paired *t*-test *p*-values in this experimental environment. As mentioned before, three different depths $k = 100$, 500 and 1,000 is considered for selecting documents per-system in this experiment which are compared to the benchmark systems ranking, i.e. the obtained *t*-test *p*-values at topic level scores of the same systems.

Similarly, Figure 5 illustrated the results of applying significance test on the same ten system pairs which are previously selected. But now, Mann-Whitney test is applied for examining the probable difference between the selected significance tests.

Moreover, as it can be seen in both figures, in the proposed method, in high depths in which the number of selected documents increases, more accurate results are obtained in the significance tests. In other words, the power of the tests in indicating the significance difference between system pairs increases with the experiment sample size.

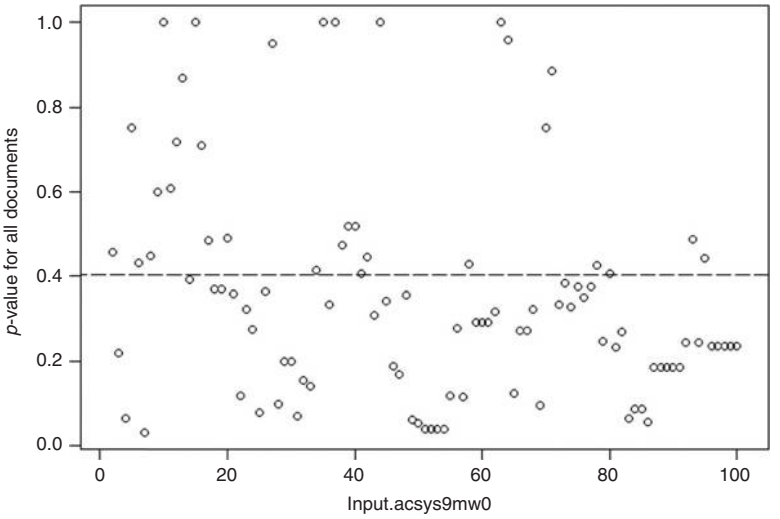

Figure 3.
p-Values from the *t*-test done on 103 systems – 102 pairs

Although there is a notable difference in the measurement when depth $k = 100$ in both tests, there is a slight contrast when depth k increases by 500 and 1,000.

It is worth mentioning that due to existence of tie differences or zero differences, the experiments are involved ties which are usually ignored in the measurements(Conover, 1973; Emerson and Simon, 1979; Putter, 1955; Randles, 2001; Rayner and Best, 1999).

5. Conclusion

IR systems evaluation has been in the centre of attention due to its viable role in finding the best retrieval algorithms or systems in IR research field. Refinement of the ranking systems and enhancing the results quality is achievable through a systematic evaluation mechanism. Even after a proper system evaluation, there is a need to be assured of the obtained results. Utilizing statistical tests is widely used in this stage. However, the current significance tests are based on averaging the systems scores that leads to incur data loss. In this paper, a new

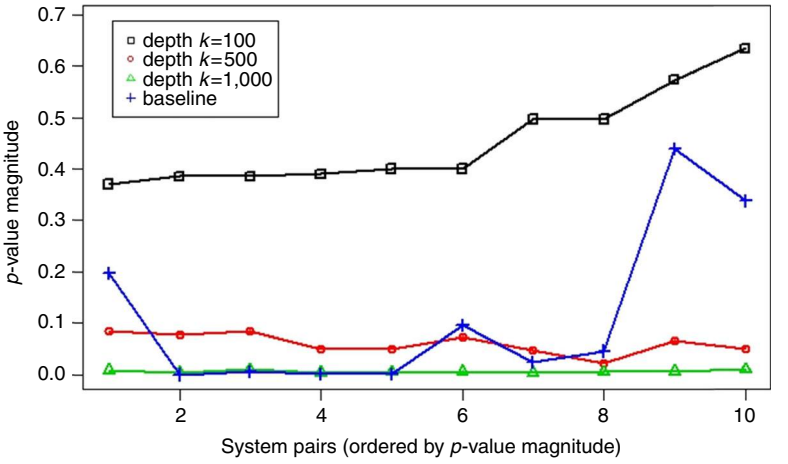

Figure 4.
A comparison of *p*-value ranges of ten system pairs in three different depths using documents weight against systems mean AP scores by *t*-test

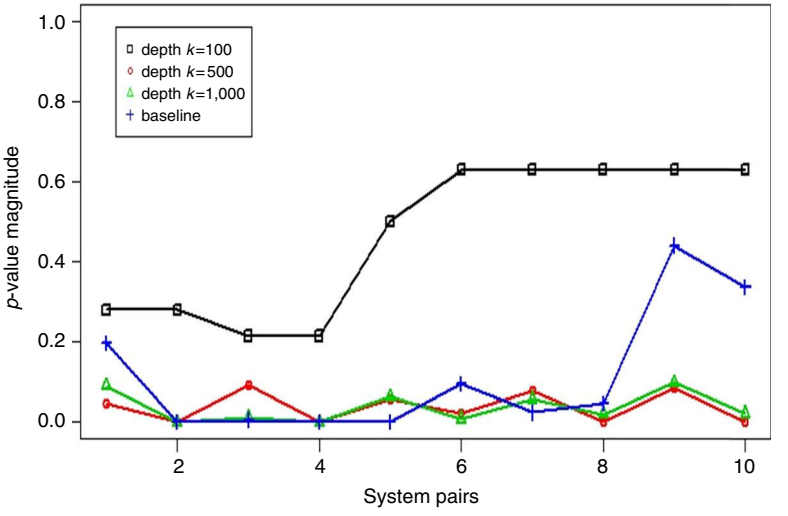

Figure 5.
A comparison of *p*-value ranges of ten system pairs in three different depths using documents weight against systems mean AP scores by Mann-Whitney test

approach that considers the systems retrieved documents instead of average systems scores as a measurement unit is proposed. The experimental results obtained from two applied tests, namely *t*-test and Mann-Whitney, indicate an improvement in detecting how likely the difference between systems is significant. Moreover, increase in the experiment sample (documents) size leads to a boost in the power of the tests. There are still a lot of works remaining for future studies to validate this novel approach by performing more experiments on different document depths and more various TREC data collections.

References

Al-Maskari, A., Sanderson, M. and Clough, P. (2007), "The relationship between IR effectiveness measures and user satisfaction", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 773-774.

Al-Maskari, A., Sanderson, M., Clough, P. and Airio, E. (2008), "The good and the bad system: does the test collection predict users' effectiveness?", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 59-66.

Aslam, J.A., Pavlu, V. and Savell, R. (2003), "A unified model for metasearch, pooling, and system evaluation", *Proceedings of the Twelfth International Conference on Information and Knowledge Management, ACM, New York, NY*, pp. 484-491.

Aslam, J.A., Yilmaz, E. and Pavlu, V. (2005), "The maximum entropy method for analyzing retrieval measures", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 27-34.

Bendat, J.S. and Piersol, A.G. (2011), *Random Data: Analysis and Measurement Procedures*, Vol. 729, John Wiley & Sons, Hoboken, NJ.

Buckley, C. and Voorhees, E.M. (2000), "Evaluating evaluation measure stability", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 33-40.

Buckley, C. and Voorhees, E.M. (2004), "Retrieval evaluation with incomplete information", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 25-32.

Carbno, C. (2007), "Business statistics: contemporary decision making", *Technometrics*, Vol. 49 No. 4, pp. 495-496.

Carterette, B. and Voorhees, E.M. (2011), "Overview of information retrieval evaluation", *Current Challenges in Patent Information Retrieval*, Vol. 29 No. 4, pp. 69-85.

Carterette, B., Allan, J. and Sitaraman, R. (2006), "Minimal test collections for retrieval evaluation", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 268-275.

Carterette, B.A. (2012), "Multiple testing in statistical analysis of systems-based information retrieval experiments", *ACM Transactions on Information Systems (TOIS)*, Vol. 30 No. 1, p. 4.

Conover, W.J. (1973), "On methods of handling ties in the wilcoxon signed-rank test", *Journal of the American Statistical Association*, Vol. 68 No. 344, pp. 985-988.

Cormack, G.V. and Lynam, T.R. (2007), "Validity and power of t-test for comparing MAP and GMAP", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 753-754.

Emerson, J.D. and Simon, G.A. (1979), "Another look at the sign test when ties are present: the problem of confidence intervals", *The American Statistician*, Vol. 33 No. 3, pp. 140-142.

Harman, D. and Buckley, C. (2004), "The NRRC reliable information access (RIA) workshop", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 528-529.

Huffman, S.B. and Hochster, M. (2007), "How well does result relevance predict session satisfaction?", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 567-574.

Hull, D. (1993), "Using statistical testing in the evaluation of retrieval experiments", *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 329-338.

Jarvelin, K. and Kekalainen, J. (2002), "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems (TOIS)*, Vol. 20 No. 4, pp. 422-446.

Kendall, M.G. (1938), "A new measure of rank correlation", *Biometrika*, Vol. 30 Nos 1/2, pp. 81-93.

Melucci, M. (2007), "On rank correlation in information retrieval evaluation", *ACM SIGIR Forum*, Vol. 41 No. 1, pp. 18-33.

Mizzaro, S. and Robertson, S. (2007), "Hits hits TREC: exploring IR evaluation results with network analysis", *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY*, pp. 479-486.

Moghadasi, S.I., Ravana, S.D. and Raman, S.N. (2013), "Low-cost evaluation techniques for information retrieval systems: a review", *Journal of Informetrics*, Vol. 7 No. 2, pp. 301-312.

Putter, J. (1955), "The treatment of ties in some nonparametric tests", *The Annals of Mathematical Statistics*, Vol. 26 No. 3, pp. 368-386.

Randles, R.H. (2001), "On neutral responses (zeros) in the sign test and ties in the wilcoxon–mann–whitney test", *The American Statistician*, Vol. 55 No. 2, pp. 96-101.

Ravana, S.D. (2011), "Experimental evaluation of information retrieval systems", PhD thesis, University of Melbourne, Melbourne, VC.

Rayner, J. and Best, D. (1999), "Modelling ties in the sign test", *Biometrics*, Vol. 55 No. 2, pp. 663-665.

Robertson, S. (2006), "On GMAP: and other transformations", *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, New York, NY*, pp. 78-83.

Sakai, T. (2006), "Evaluating evaluation metrics based on the bootstrap", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 525-532.

Sanderson, M. and Zobel, J. (2005), "Information retrieval system evaluation: effort, sensitivity, and reliability", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 162-169.

Savoy, J. (1997), "Statistical inference in retrieval effectiveness evaluation", *Information Processing & Management*, Vol. 33 No. 4, pp. 495-512.

Smucker, M.D., Allan, J. and Carterette, B. (2007), "A comparison of statistical significance tests for information retrieval evaluation", *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, New York, NY*, pp. 623-632.

Smucker, M.D., Allan, J. and Carterette, B. (2009), "Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes", *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 630-631.

Soboroff, I., Nicholas, C. and Cahan, P. (2001), "Ranking retrieval systems without relevance judgments", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 66-73.

Turpin, A. and Scholer, F. (2006), "User performance versus precision measures for simple search tasks", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 11-18.

Voorhees, E.M. (2002), "The philosophy of information retrieval evaluation", *Evaluation of Cross-Language Information Retrieval Systems*, Springer-Verlag, London, pp. 355-370.

Voorhees, E.M. (2003), "Overview of the TREC 2003 robust retrieval track", in Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (Eds), *TREC*, pp. 69-77.

Wackerly, D., Mendenhall, W. and Scheaffer, R. (2007), *Mathematical Statistics with Applications*, Cengage Learning, Belmont, CA.

Webber, W., Moffat, A. and Zobel, J. (2008a), "Score standardization for inter-collection comparison of retrieval systems", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 51-58.

Webber, W., Moffat, A. and Zobel, J. (2008b), "Statistical power in retrieval experimentation", *Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM, New York, NY*, pp. 571-580.

Webber, W., Moffat, A. and Zobel, J. (2010), "A similarity measure for indefinite rankings", *ACM Transactions on Information Systems (TOIS)*, Vol. 28 No. 4, pp. 1-34.

Yilmaz, E., Aslam, J.A. and Robertson, S. (2008), "A new rank correlation coefficient for information retrieval", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 587-594.

Zobel, J. (1998), "How reliable are the results of large-scale information retrieval experiments?", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 307-314.

About the authors

Dr Sri Devi Ravana received her Bachelor of Information Technology from the National University of Malaysia in 2000. Followed by the Master of Software Engineering from the University of Malaya, Malaysia and PhD Degree from the Department of Computer Science and Software Engineering, The University of Melbourne, Australia, in 2001 and 2012, respectively. Her research interests include information retrieval and Web 2.0 in education. She received a couple of best paper awards in international conferences within the area of information retrieval. She is currently a Senior Lecturer at the Department of Information Systems, University of Malaya, Malaysia. She is also the recipient of her University's Excellent Service Award in 2013. Dr Sri Devi Ravana is the corresponding author and can be contacted at: sdevi@um.edu.my

Masumeh Sadat Taheri is currently a Research Assistant in the Information System Department, Faculty of Computer Science and Information Technology at the University of Malaya, Malaysia. She received her degree in Master of Computer Science in 2013 at the same university. Her current research interests include information retrieval evaluation, scientometrics and game-based learning techniques.

Prabha Rajagopal is a Graduate Research Assistant in the University of Malaya, Malaysia and attached to the Department of Information Systems. She is pursuing her PhD studies in the field of information retrieval.