MEASURING THE RELIABILITY OF RANKING IN INFORMATION RETRIEVAL SYSTEMS EVALUATION

***Prabha Rajagopal*[1], *Sri Devi Ravana*[2*]**

[1,2] University of Malaya, 50603 Kuala Lumpur, Malaysia

Email: prabz13@yahoo.com[1], sdevi@um.edu.my[2*](corresponding author)

DOI: https://doi.org/10.22452/mjcs.vol32no4.1

ABSTRACT

A reliable system is crucial in satisfying users' need, but the reliability is dependent on the varying effects of the test collection. The reliability is usually evaluated by the similarities of a set of system rankings to understand the impact of variations in relevance to judgments or effectiveness metrics. However, such evaluations do not indicate the reliability of individual system rankings. This study proposes a method to measure the reliability of individual retrieval systems based on their relative rankings. The Intraclass Correlation Coefficient (ICC) is used as a reliability measure of individual system ranks. Various combination of effectiveness metrics according to their clusters, selection of topic sizes, and Kendall's tau correlation coefficient with the gold standard are experimented. The metrics average precision (AP) and rank-biased precision (RBP) are suitable for measuring the reliability of system rankings and generalizing the outcome with other similar metrics. Highly reliable system rankings belong mostly to the top and mid performing systems and are strongly correlated with the gold standard system ranks. The proposed method can be replicated to other test collections as it utilizes relative ranking in measuring reliability. The study measures the ranking reliability of individual retrieval systems to indicate the level of reliability a user can consume from the retrieval system regardless of its performance.

Keywords: Information Retrieval, System Evaluation, Reliability Testing, Intraclass Correlation Coefficient, TREC, Information Systems

1.0 INTRODUCTION

The performance of an information retrieval system can be measured using various metrics. Each metric evaluates the systems differently according to the user model implementation [1]. The effectiveness of the system is measurable using retrieved ranked documents, metrics, and relevance judgments. The retrieval is an important aspect of the system resulting in studies thet focuses on indexing [2]–[5] and query formulation [3], [4], [6]. The evaluation of the retrieval system is one of the other important aspects of information retrieval. The evaluation usually provides information about the performance of the system.

A system that performs well in a test collection does not necessarily has the same performance wit another test collection [7]. Meanwhile, the selection of topics is also important in the quality of system ranking [8], [9]. It is inevitable that a system would perform differently for different topics or metrics. Due to various effects in the evaluation of retrieval system, the reliability of the test collections [10], relevance assessment [11]–[13], and effectiveness metrics [14], [15] were studied.

Topic size selection is dependable to the test collection and task [10]. The aspect of test collection is the relevance judgments. The reliability of relevant judgments was measured through the variation in system rankings. Relevance judgments from different expert judges [13] or crowdsourced judges [11] did not impact the reliability of the system rankings. System rankings are in fact translations from effectiveness metrics scores. Suitable metrics was suggested based on the binary and graded relevance [14], and metrics that represent clusters of similar metrics based on some mathematical properties [15]. The experimentation in [15] used effectiveness scores instead of the translated ranks to identify the metrics clusters for IR system evaluation. However, in [15], it did not mention which of those metrics are suitable for measuring the reliability of the system rankings.

A reliable system is crucial in satisfying users' need. In evaluating the IR systems, different metrics could be based on user models, fulfilling different user needs [1]. While the studies [13], [15] mentioned above have measured reliability for a set of systems through experimentations of differences in scores or ranks obtained from a test-retest setting, none have explored the individual system's ranking reliability. Evaluating similarities of a set of system rankings were explored due to variations from relevance judgments, metrics or test collections. When system rankings are evaluated in a set of systems, the results such as Kendall's tau only indicates the overall strength of correlation coefficient but not for the individual systems.

As an example, consider a scenario in Table 1 where the effectiveness scores are different but not the ranks. The ranks are from different metrics, which measure the systems according to their user model representation. However, a system's reliability can be observed by their consistency in rankings.

Table 1: Example of consistency in system ranking.

System	Topic	P@5	Rank	AP@5	Rank
Sys A	Topic 1	0.8	1	0.81	1
Sys B		0.6	2	0.45	2
Sys A	Topic 2	0.2	1	0.04	1
Sys B		0	2	0	2
Sys A	Topic 3	0.4	1	0.28	1
Sys B		0.2	2	0.25	2

The consistency in a system's ranking can be translated as the reliability of the system's performance. It is not dependent on good or poor performing system but rather its ability to sustain its performance in regards to other systems. With that, this study aims to propose a method to measure the reliability of the retrieval systems based on their rankings.

The next sections discuss the past studies, reliability measures, methodology and the proposed method. Following these are results and discussions, and finally the conclusion.

2.0 LITERATURE REVIEW

This section provides information about the past studies related to measuring reliability, the differences between inter-rater, and intra-rater correlation in measuring reliability, and details about intra-rater correlation known as the intraclass correlation coefficient (ICC) as a suitable measure of reliability.

2.1 Past Studies

Various studies have explored measuring reliability with regards to the relevance judgments, test collection, and effectiveness metrics. The reliability of relevance judgments were measured for intra-assessor consistency [12], and inter-assessor agreement [11]. As part of their study, intra-assessor consistency measure was used to examine the frequency of an assessor making the same decision multiple times [12]. The assessors needed to select or rate expansion terms from a supplied list. These expansion terms appeared in multiple forms while the same assessors repeat the assessment of expansion terms on the same topic. Such setup is to allow comparison on the consistency of the term selection. The Spearman's rho was used as a measure of reliability [12].

In the year 2000, NIST initiated a study to identify the changes in system rankings due to the differences in the relevance judgment by different expert judges. High Kendall's tau correlation coefficients proved the system rankings of the retrieval systems are reliable despite the differences in the relevance judgments [13]. Similarly, [11] conducted experiments on relevance judgments using crowdsource and showed that the judgments are reliable in repeated experiments. Three different experiments; two crowdsourced judgments and one expert judges' judgments were used in measuring inter-assessor agreement using Fleiss's kappa. The Fleiss's kappa is suitable for use with categorical data when there are more than two raters [16]. The study has shown that although the reliability of the crowdsourced judges is lower than expert judges, the system rankings from crowdsourced judges were same as that obtained from expert judges [11].

Another study mentioned that a test collection is reliable if the conclusions can be replicated with other test collection [10] although a system that perfroms well with one test collection does not always perform well with another test collection [7]. To overcome the difficulties of interpreting the Generalizability Theory, the study [10] proposed a tool based on interval estimates of the stability indicators. Through experimentation, it was concluded that some test collections were not reliable. Also, query set size selection depends on the task and test collection as there was significant variation among the test collections.

In [17], an 'Adaptive' method was experimented to select queries and determine the reliability of system rankings using Kendall's tau and Pearson correlation coefficient. The results suggest that the 'Adaptive' method employed can produce reliable system rankings.

Variations in system evaluation could also occur due to effectiveness metrics. A study proved that some metrics could be grouped together in terms of mathematical properties [15]. As such, seven groups of metrics were determined, and it was also stated that any one of the metrics within the group could be used as an evaluation metric. The mean average precision (MAP) as a measure of average precision, the precision at ten (P@10) to measure satisfaction of user with the first retrieved document, and the exact_recall to measure the ability of the system to retrieve most of the relevant documents has been suggested [15]. The study used seven ad hoc test collections from TREC from the year 1993 to 1999. The average precision (AP) is favorable due to its stability and robustness whereby if the differences are statistically significant, AP-based differences between systems on one set of topics can be observed on another set of topics [18]. The RBP, on the other hand, measures the rate at which utility is gained by a user at a given degree of persistence, *p* [18].

2.2 Difference between Inter-rater correlation (interclass) and Intra-rater (intraclass) correlation

Reliability can be measured using specific measures such as Fleiss's Kappa [11], Spearman's rho [12], Kendall's tau and Pearson correlation coefficient [17]. These reliability measures can be divided into inter-rater and intra-rater correlation. The relationship between variables is defined as the correlation while the strength of the correlation is presented as the correlation coefficient [19].

In the information retrieval evaluation field, Pearson correlation, Kendall's tau, and Spearman rho are the usual correlation coefficients [20], which is also known as inter-rater correlations. In short, the Pearson measures linearity between two sets of variables, Kendall's tau measures the rank swaps between two sets of system rankings, while Spearman's rho/rank correlation measures the strength and direction of the association between two variables for non-normal distribution of data.

The inter-rater and intra-rater are closely related to correlation, which sometimes more useful and accurate to be measured in regards to analysis of variance [21]. The intra-class correlation coefficient (ICC) is an intra-rater correlation, common in social sciences [22]. The ICC is the correlation between one measurement on a target and another measurement obtained on that same target [23]. The difference between inter-rater (Pearson) and intra-rater (ICC two-way random effects model) correlation is shown using an example in Table 2.

Table 2: Example of data consistency and agreement.

Topic	Rater 1	Rater 2	Rater 3
1	1	6	1
2	2	7	2
3	3	8	3
4	4	9	4
5	5	10	5

The Pearson correlation coefficient between Rater 1 and Rater 2 is 1, and the correlation coefficient between Rater 1 and Rater 3 is also 1. The exact correlation coefficient occurs when the variance-covariance matrix is unchanged from the original dataset [24]. The Pearson correlation may reflect agreement by underestimating or overestimating the inter-rater agreement [25]. Hence the Pearson correlation is not recommended to measure reliability [24]. Similarly, the Kendall's tau correlation is also not suitable to measure realibility. Meanwhile, the ICC using two-way random effects model between Rater 1 and Rater 2 is 0.167, and between Rater 1 and Rater 3 is 1. The ICC measures the agreement between the two raters. The more consistent the measurement between the two raters, the higher will be the reliability [26].

The ability of ICC in measuring the agreement or uniformity in data is suitable for measuring the reliability of system rank in this study. The ICC has a few criteria of model selection and is detailed in the next section.

2.3 Models of Intraclass Correlation Coefficient (ICC)

The technique of finding ICC is a basis of variance analysis and estimation of various variance components known as reliability index. The ICC can only be interpreted as correlation coefficient if the denominator includes the total variance [22]. The selection of ICC depends on the decomposition of a rating made by the ith judge on the jth target in regards to various effects [23]. An ICC can take values between 1 and -1/(*k*-1) [24]. However, a negative ICC is usually taken to be zero reliability [27]. There are three important decisions to select a proper ICC; (1) whether the ANOVA should be one way or two way, (2) whether raters' effect is considered random or fixed, and (3) whether the analysis unit is a single measurement or mean of few measurements.

If the study has non-consistent raters, whereby different raters will rate each of the targets, the ANOVA model is one-way random Model 1. If the study uses consistent raters, it is necessary to determine if the raters are a sample or population. Consistent raters mean the same raters would rate each target. Use Model 2 if the raters are samples from a larger population, and Model 3 for population raters or the only raters of interest [23], [26].

Model 2 has both raters and target effects while assuming both raters and targets are drawn randomly from a larger population [26]. The two-way random model using single measurement unit is defined in equation [1], whereby MSB$_{\text{targets}}$ is mean square between targets, MSE is mean square error, MSB$_{\text{rater}}$ is mean square between raters, *k* is the number of raters rating each target and *n* is the number of targets.

$$ICC(2,1) = \frac{MSB_{targets} - MSE}{MSB_{targets} + (k-1)MSE + \frac{k(MSB_{rater} - MSE)}{n\prime}} \qquad [1]$$

Model 3 assumed that the raters as fixed [23], [26] while removing the between raters variance [22], [23]. When the raters variance is ignored, the correlation coefficient is interpreted as rater consistency rather than rater agreement, which leads us to the second decision related to the effects due to judges [23]. The Model 2 allows generalization to other raters within the population. Meanwhile, the fixed raters in Model 3 indicate interest in a single rater or a fixed set raters and there are no other raters of interest.

The final decision is related to the analysis unit. The analysis unit could take single or mean measures. The single measure applies to single measurements such as individual scores, and mean measure applies to average measurements such as the average score for a *k*-item test [28] or an average of 2 or more measurements taken by different raters.

There are six different combinations of models and forms in ICC. In the ICC representation, the first number indicates the model (1,2 or 3) while the second represents the form (1 - single or *k* – average of *k* raters). The different combinations and forms in ICC are as below.

- ICC(1,1) - Each subject is assessed by a different set of randomly selected raters, and the reliability is calculated from a single measurement.
- ICC(1,*k*) - As above, but reliability is calculated by taking an average of the *k* raters' measurements.
- ICC(2,1) - Each subject is measured by each rater, and raters are considered representative of a larger population of similar raters. Reliability calculated from a single measurement.
- ICC(2,*k*) - As above, but reliability is calculated by taking an average of the *k* raters' measurements.
- ICC(3,1) - Each subject is assessed by each rater, but the raters are the only raters of interest. Reliability calculated from a single measurement.
- ICC(3,*k*) - As above, but reliability is calculated by taking an average of the *k* raters' measurements.

3.0 METHODOLOGY

This study proposes to measure the reliability of individual system rank with the use of ICC as a measure of reliability. However, before conducting the experimentation, it is important to decide the appropriate ICC model and unit of measurement.

The experimentation represents the targets by topics and the raters by effectiveness metrics. Both the topics and effectiveness metrics are data samples. The matching model is Model 2. Information retrieval systems can be evaluated using various effectiveness metrics. Hence the selection of few metrics indicates a sample from a larger population. Similarly, a large number of topics are available. Therefore, it is also suitable to say the sample topics are drawn from a population.

The second decision is to determine the ICC measurement; agreement (random raters, Model 2) or consistency (fixed raters, Model 3). As seen earlier in the literature review section, agreement and consistency are different. This study focuses on measuring the agreement to determine the reliability of system ranks. In other words, agreement measures data uniformity.

The final decision is about the measurement unit. The topic rank represents the ratings for each target or topic. These individual ranks will be used in the ICC computation. Hence, the experimentation uses ICC(2,1) whereby the number 2 represents agreement or random raters while number 1 represents the single unit of measurement.

The experimentation uses dataset from the TREC-2004 Robust track which consists of 110 systems and 249 topics. A larger number of topics than the usual 50 topics per test collection is needed for multiple combinations of topics in this experimentation. TREC provides the system input runs and the relevance judgment. Based on these input runs, the reliability of the systems' ranks will be measured.

Fig. 1 shows the steps involved in computing the reliability of system rank and the validation of the proposed method with the gold standard. The gold standard is the system ranks using MAP scores from all topics for each system. Evaluation metrics from groups of AP, rank biased precision (RBP p=0.95), and Precision@k (P@k) will be used. For simplicity, Fig. 1 shows only two metrics, but the steps apply for all metrics.

The first step involves calculating the effectiveness scores for each topic per system according to the selected metrics. For ease of understanding, assume the metric is AP@1000. Choose one topic, for example, topic 1 and gather the topic scores for all the 110 systems. Based on these topic scores, rank the systems. Repeat the ranking of topics per system for the remaining topics. Each system will now have 249 ranks from each topic using metric AP@1000. Similarly, compute the topic ranks using the other metrics.

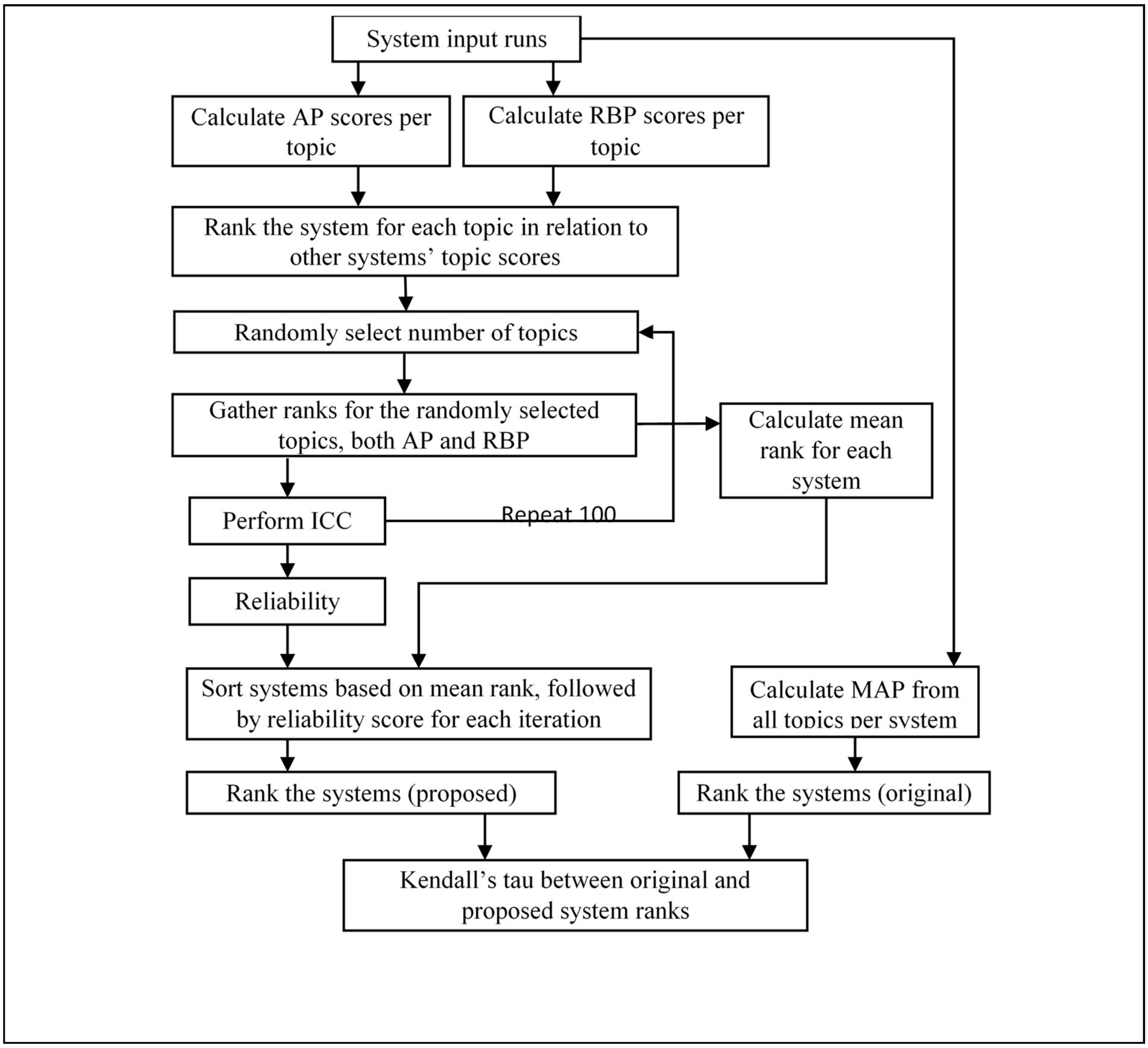
System input runs
Calculate AP scores per topic
Calculate RBP scores per topic
Rank the system for each topic in relation to other systems' topic scores
Randomly select number of topics
Gather ranks for the randomly selected topics, both AP and RBP
Calculate mean rank for each system
Perform ICC
Repeat 100
Reliability
Sort systems based on mean rank, followed by reliability score for each iteration
Calculate MAP from all topics per system
Rank the systems (proposed)
Rank the systems (original)
Kendall's tau between original and proposed system ranks

Fig. 1: Steps involved in proposed method for measuring the reliability of system ranking.

Next, randomly select some topics from the total 249 topics. Topic size variations are 10, 20, 30, 40, and 50. The purpose of topic size variation is to identify the number of topics needed to obtain the best reliability coefficient. For example, gather the ranks of a system obtained from both metrics for 10 randomly selected topics. Perform ICC (2,1) using these ranks. Repeat the same for other systems. The experimentation uses R language for the ICC computation.

Steps from a random selection of topics till the calculation of ICC is repeated 100 times to allow the combination of various topics for each topic size selection. Though the ICC provides reliability of the ranks for each system, it is important to determine the closeness of this output to the gold standard. The gold standard is the system ranks from MAP scores obtained from all topics.

The mean rank is calculated from the selected random topics for each system separately (see Fig. 1). The mean rank best suits to represent the reliability index or coefficient since the ANOVA components calculate the variance by utilizing the mean rank. Some systems appear to have tied mean ranks. Hence, the reliability score is used to determine the rank among the tied systems. The systems are ranked ascendingly by mean rank, followed by descending reliability score. High-reliability score from ICC indicates good results.

Finally, Kendall's tau correlation coefficient between gold standard system rankings and the proposed method's system rankings is performed. High Kendall's tau correlation coefficient indicates the system ranks from the proposed method is close to that of original or gold standard system ranks.

4.0 RESULTS AND DISCUSSION

4.1 Agreement of system rankings as a measure of reliability

The ICC(2,1) measures the agreement between two metrics using a single measurement of the topic ranks. The more uniform the ranks between the metrics, the higher will be the reliability [26]. A reliability coefficient of 0.8 and above indicates high reliability [10]. Four different groups of metrics experimented are AP, RBP, low P@*k* (*k* = 10 or 30), and high P@*k* (*k* = 100 or 200). The combination of these metrics can be grouped as 'outside cluster' and 'within cluster'. The outside cluster means both the metrics belong to different groups and within cluster means both the metrics belong to the same group. The classification of clusters was taken from [15]. Outside cluster determines if the system is reliable in satisfying different user needs. Meanwhile, within cluster determines the reliability of the system rank for a specific user model.

Table 3 shows the number of highly reliable systems measured using different metrics and is divided according to the clusters. The rows in the table are in pairs. Any one of the row in the pair represents the initial experiment for measuring the reliability of system rankings while the other row represents the generalizability using metrics from the same population. The rows can be interchanged. Recall that ICC Model 2 allows generalization to other samples within the population. In accordance to this, other samples of metrics within the same population were attempted to determine the generalization of system rankings' reliability.

The similarity in results for outside cluster from Table 3 between the initial metrics combination and its generalization metrics is an indication of the consistency of the particular metrics in measuring system ranking reliability. If the numbers between the initial and generalization metric (the pairs of rows from outside cluster) have small variation, it would indicate the consistency in the measuring the system ranking reliability using that particular combination of metrics. Large variations between the initial and generalization metrics would mean the metrics combination is not consistent in measuring the system rankings' reliability. In other words, small variation means using a particular combination of metrics, similar results could be observed with any of the metrics combination from the same cluster, and significant variations mean results may vary with that combination of metrics. As for the within cluster combinations, a consistent number is favorable to indicate the capability of the metrics to be consistent in its output to measure the system rankings.

The combination of AP@100 and RBP@100 metrics can determine significant numbers of highly reliable system rankings. When another set of metric samples, AP@1000 and RBP@1000 experimented, a similar outcome was attained. For this group of metrics, the generalization holds true. Another sample of metrics from these population should yield similar system rankings' reliability. This group of metrics is also the best among the remaining outside cluster metrics experimented. Nonetheless, with these numbers, it is not possible to understand the level of robustness and user gain. These systems could represent good or poor performing systems but are reliable in their relative ranking measurements. The relative ranks of AP and RBP are similar to produce high reliability, which could also mean the effectiveness of the systems from both these metrics are somewhat similar.

As the Table 3 shows, a combination of AP with low P@*k* and AP with high P@*k* have low numbers of reliable system rankings. Their generalized metrics combinations also show similar results. These systems produce the varying level of ranks between both the metrics. The low numbers in reliable systems could be due to the P@*k* metrics. The effectiveness scores among few systems tend to be the same. The ranks are numbered by effectiveness scores followed by system name. Similar effectiveness scores could have caused a larger rank variance between AP and P@*k*.

The combination of RBP@100 and P@30 has large numbers of highly reliable system rankings. These ranks from RBP@100 and P@30 must be mostly in agreement to be able to achieve high reliability. When compared with another set of the sample from the same population, RBP@1000 with P@10, the results are not alike. The combination of RBP with low P@*k* metrics is not consistent in measuring the reliability of system ranks. The generalization of the system rankings' reliability may produce varying outcome when different metric samples are used.

Table 3: Number of highly reliable systems (averaged from 100 iterations) measured using different pairs of metrics.

Cluster	Topic size	10	20	30	40	50
Outside cluster	AP@100-RBP@100	101	99	87	87	85
	AP@1000-RBP@1000	78	87	86	88	87
	AP@100-P@30	6	5	5	6	6
	AP@1000-P@10	0	0	0	0	0
	AP@100-P@100	0	0	0	1	0
	AP@1000-P@200	4	13	12	13	13
	RBP@100-P@30	64	71	73	69	72
	RBP@1000-P@10	3	2	4	3	3
	RBP@100-P@100	1	5	23	36	33
	RBP@1000-P@200	0	0	0	0	0
Within cluster	P@30-P@10	0	0	0	0	0
	P@100-P@200	109	109	109	109	109
	AP@1000-AP@100	67	71	73	73	74
	RBP@1000-RBP@100	83	83	83	83	83

When RBP is combined with high P@*k*, the number of highly reliable system rankings remain low, although 33% of systems had high reliable system rankings by RBP@100 with P@100. The performance of these systems is different in measuring user gain and high P@*k*. It is the reason behind low numbers of reliable system rankings.

The within cluster metrics only measures the reliability of the systems in satisfying a single user need. The reliability is measured using ranks from sample metrics within the same population. Low P@*k* is unable to identify any reliable systems. In contrast, high P@*k*, AP, and RBP metrics have large numbers of reliable systems. The numbers suggest that evaluating system ranks with any one of the metrics within the cluster would still produce highly reliable results. The outcome for within cluster metrics is similar to [15]'s study which aimed at identifying small numbers of metrics to represent a large number of metrics available for evaluation. Conversely, this study focuses on measuring the reliability of the individual system ranks.

For the combinations of metrics from outside cluster, the topic size is not conclusive. The use of smaller or larger topic sizes will impact the number of reliable system rankings identified. For most of the metric combinations within the cluster, the topic sizes produce similar numbers of systems. Hence, the usage of any topic size will not impact the number of reliable system rankings.

4.2 Representation of highly reliable system rankings with the original system ranks

This paper further questions if these systems with highly reliable rankings represent the top or bottom ranked systems from the original MAP or gold standard system ranks. Fig. 2 shows the plot of average reliability score from the 100 iterations against the system ranks from original MAP scores. The plots are from topic size 30 for the various combinations of metrics. Only plots for topic size 30 is shown in the graphs for the other topic sizes are similar. The plots represent metric combinations outside the cluster. The labels A, B, C, D, and E is to differentiate the metric pairs and its generalization metric.

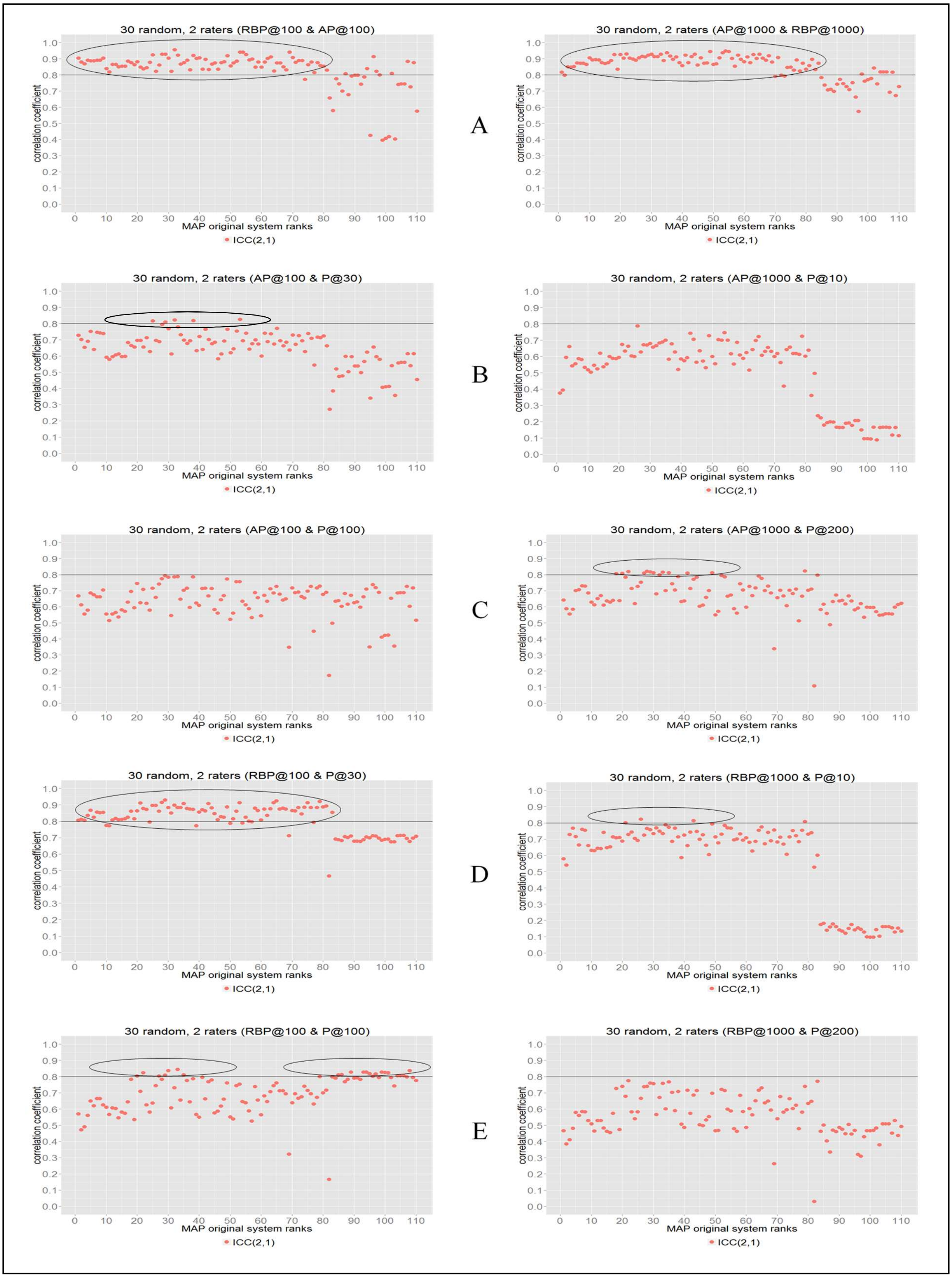

Fig. 2: Original system rank against average ICC(2,1) from 100 iterations for different combination of metrics (outside cluster).

For label A, the plot shows systems with highly reliable rankings consist mostly of the top and middle ranked systems. Similar observations for sample metrics from the same population as well. For label B, mostly the systems do not have reliable rankings except very few mid-ranked systems. Meanwhile, for metrics combination in label C, a small number of mid-ranked systems have reliable rankings. However, it cannot be differentiated if the top or low ranked systems are better or worse in any way. It is because some low ranked systems appear to have reliability scores as good as top-ranked systems.

When observing the left plot for label D, clear groupings can be seen. The top-ranked systems have better reliability compared to those of low ranked systems. The right plot for label D appears to have shifted lower in magnitude compared to the left plot. The metric P@10 could have been the cause of such shift. The similar effect of P@10 could be observed with the low ranked systems on the right plot of label B, but it is not the case when another sample metric from low P@*k* (P@30) was combined with AP or RBP. The P@10 is not a good choice in measuring system rank reliability due to its low evaluation depth causing multiple systems to have similar effectiveness scores. For label E, the left plot shows small numbers of middle and low ranked systems with high ranking reliability. Generalizing the reliability of system rankings using similar metrics does not produce consistent results as observed on the right plot for label E.

For the combination of within cluster metrics, the highly reliable system rankings consist mostly of top-ranked systems. This indicates the system ranks from metrics within a population could produce a similar ranking for top performing systems. It appears that some combinations of metrics show high ranking reliability for top-ranked systems, while others for middle and low ranked systems. With that, it can be observed that reliable system rankings can be achieved by low ranked systems as well although their performance is not as good as other top performing systems.

4.3 Kendall's tau correlation between gold standard and proposed method

The system ranks from the gold standard, and their reliability scores were analyzed in the previous section to understand which systems from the gold standard had a highly reliable ranking. The proposed method, however, represents the reliability score for a different set of ranking as per ICC computation. Therefore, it is necessary to understand the correlation coefficient of the system ranks between the gold standard and the proposed method. A strong positive correlation will indicate the true representation of the reliability score to the gold standard system ranks.

Fig. 3 and Fig. 4 show the density plot of Kendall's tau correlation of system ranks between the proposed method and the gold standard for metrics combination outside cluster and within cluster respectively. The proposed method system ranks were obtained from the mean rank, and tied ranks were resolved using their reliability scores for each iteration (see Fig. 1). It justifies using the mean rank as ICC computation largely involves the use of overall mean rank. A general measure of Kendall's tau is also done between system ranks from the gold standard and MAP scores using 50 topics. This general measure Kendall's tau correlation is referred as a base in Fig. 3 and Fig. 4. The base is plotted as a guide for the proposed method and has a mean value of 0.88 while the standard deviation is 0.03.

The correlation coefficient of system ranks between the gold standard and proposed method has a mean value of more than 0.6 for all the metrics combination. While most correlation coefficients are moderate, some combination of metrics has strong correlation coefficient. These are AP@100-RBP@100, AP@1000-RBP@1000, RBP@1000-P@200, AP@1000-AP@100, and RBP@1000-RBP@100 metric pairs. Strong positive correlation coefficient indicates most system pair ranks to be concordant.

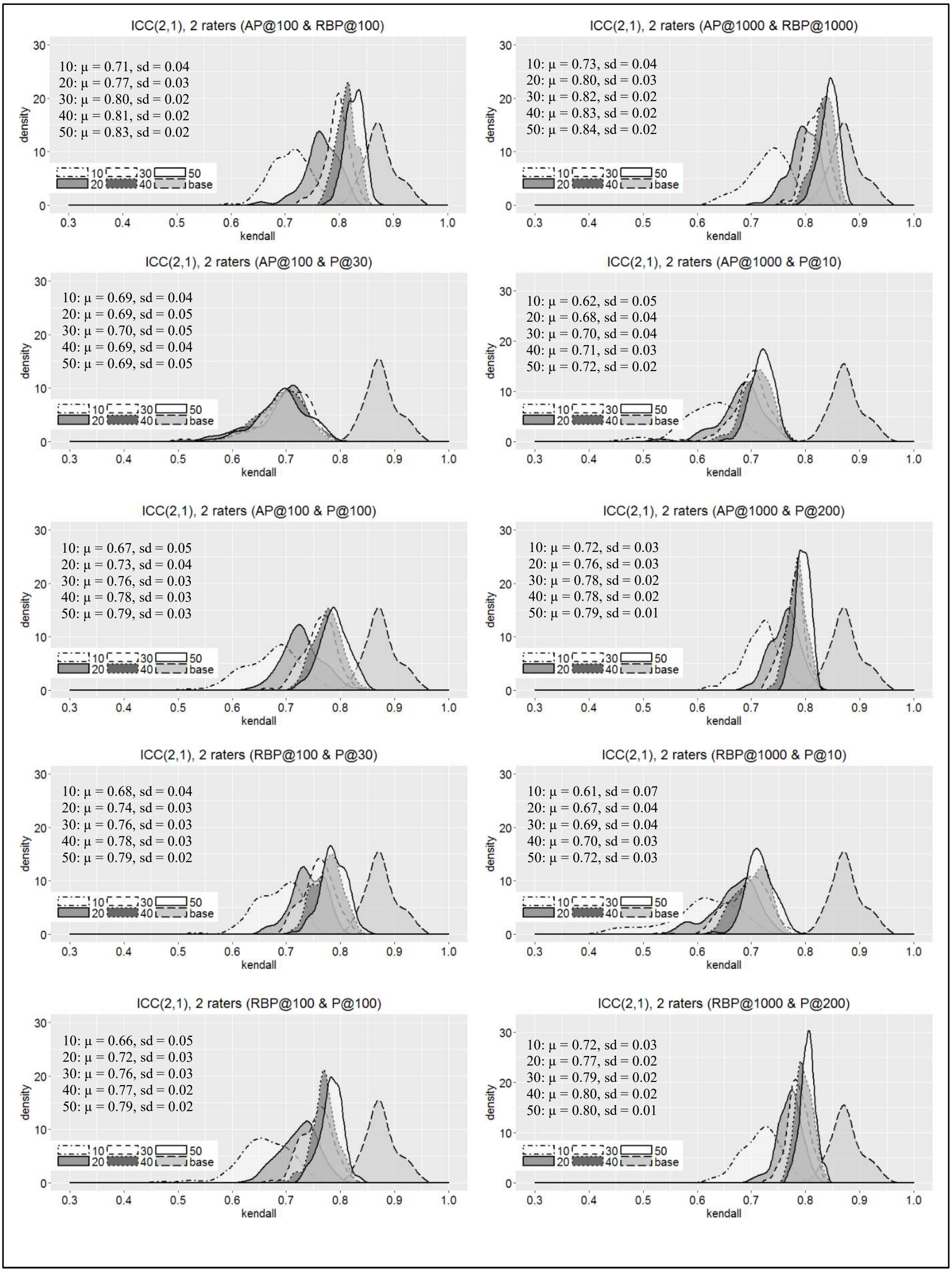

Fig. 3: Density plot of Kendall's tau correlation coefficient between proposed method and gold standard system ranks (outside cluster).

The standard deviations of the mentioned metrics are smaller than that of the base, suggesting that most tau values fall close to the mean tau value. Based on Kendall’s tau analysis, the reliability score can be accepted as the true representation of the gold standard system ranks.

Based on the experimental results and analyses, a specific combination of metrics such as AP@100-RBP@100, AP@1000-RBP@1000, AP@1000-AP@100, and RBP@1000-RBP@100 are suitable for measuring the reliability of system ranks. The highly reliable system ranks are mostly top-ranked systems from the original system ranking. In addition, they have a strong correlation coefficient of the system rankings between the gold standard and the proposed method.

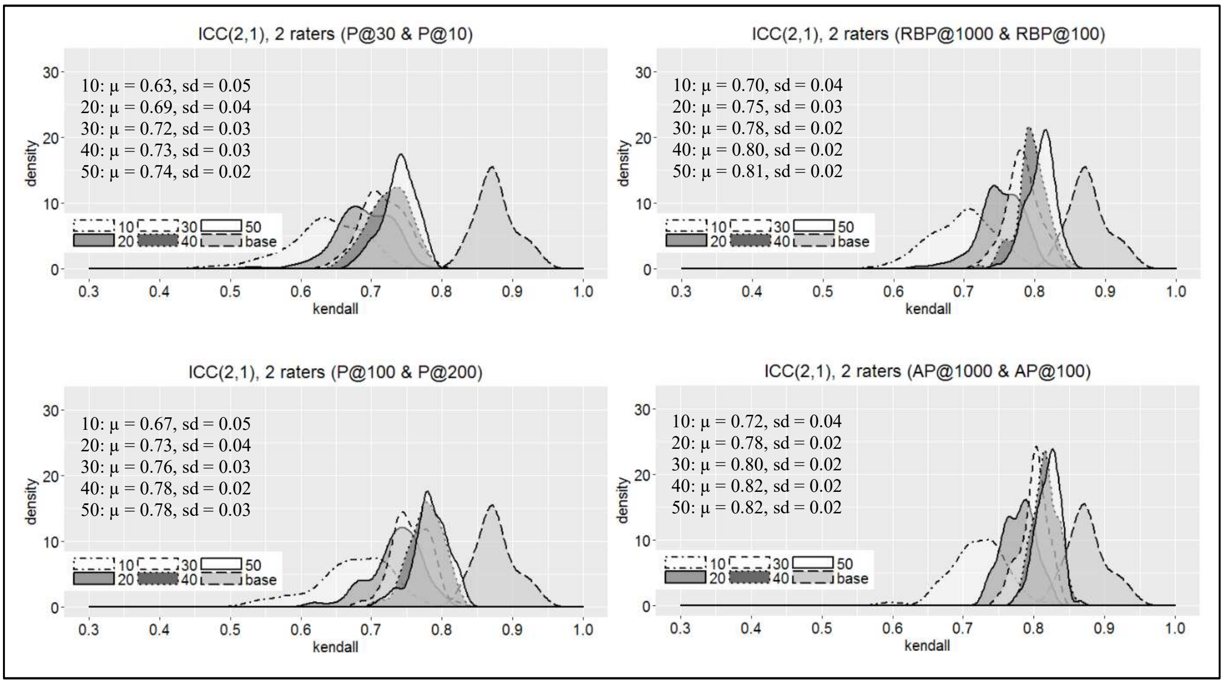

Fig. 4: Density plot of Kendall’s tau correlation coefficient between proposed method and gold standard system ranks (within cluster).

4.4 Validation of the proposed method with another test collection

This study aimed at measuring the reliability of individual systems’ ranking as opposed to the evaluation of a set of system rankings between a before-and-after experiment. The initial experiment evaluated reliability using ICC(2,1) using the ranks from two different metrics. The experimentation also explored various combinations of metrics pairs outside cluster and within the cluster. Based on the initial experimental result, an extension of the experiment is conducted on another test collection to validate the findings.

A test collection from the TREC-2005 Robust track (to be referred as dataset2 moving forward) was used for the validation experiment. This collection contains 74 system runs and 50 topics. These topics are said to be difficult topics taken from another collection [29]. Compared to the TREC-2004 Robust track (to be referred as dataset1 moving forward), the topic size is much smaller for dataset2. Nevertheless, the test collection will be used since the experimentation with dataset1 shows topic size smaller than 50 could still yield positive results.

The extension experiment is conducted with a perception that an individual is now interested to know the level of reliability a user can consume from the systems using the proposed method. Due to that, further experimentation will only select random topics once (not 100 times as in dataset1 experimentation) to compute the reliability scores of the systems. Then the number of systems that have high reliability (>=0.8) will be counted to determine if a similar trend is observed for the TREC-2005 Robust track.

Table 4 shows the number of highly reliable systems measured with ICC(2,1) using the topic ranks. The table also shows the various combination of metrics pair used that is similar to the initial experimentation using dataset1. Similar metric pairs allow for easy comparison. Since the test collection from dataset2 only contains 50 topics, the last column (50) in Table 4 is using all topics instead of random selection. The remaining topics were randomly selected once to compute the reliability score.

Table 4: Number of highly reliable systems measured using different pairs of metrics from the TREC-2005 Robust track.

Cluster	Topic size	10	20	30	40	50
Outside cluster	AP@100-RBP@100	55	41	45	42	50
	AP@1000-RBP@1000	47	45	39	36	34
	AP@100-P@30	37	23	27	23	25
	AP@1000-P@10	7	3	0	1	0
	AP@100-P@100	55	46	45	40	41
	AP@1000-P@200	53	56	53	47	49
	RBP@100-P@30	69	72	73	73	73
	RBP@1000-P@10	23	37	26	24	22
	RBP@100-P@100	47	36	50	56	49
	RBP@1000-P@200	20	26	20	18	11
Within cluster	P@30-P@10	12	14	16	12	10
	P@100-P@200	65	60	69	68	67
	AP@1000-AP@100	38	51	44	39	38
	RBP@1000-RBP@100	73	73	73	73	73

Most of the metrics pair yield similar results to dataset1. Although in the dataset1, AP and RBP combination produced the highest numbers of reliable systems, the same is not observed in this test collection. The change in numbers could be because of the single random selection of topics in dataset2, in contrast to the numbers shown in Table 3, which is an average of 100 iterations. However, the initial and generalization metric have small variation like those observed in dataset1. Such observation indicates the combination of AP and RBP metric is still suitable for measuring the reliability of systems‘ rankings.

Another difference observed between the results from both test collections involve metrics AP and low P@*k*. In dataset2, there are differences between the initial and generalization metrics, which was not observed in the dataset1. Additionally, AP and high P@*k* metrics combination also show differences between the test collections. For dataset2, there are larger numbers of systems identified as highly reliable as opposed to dataset1. Nonetheless, the initial and generalization results for the metrics combination remains consistent. Consistent here means if the initial metric identifies small or large numbers of the highly reliable system, the generalization also results in the same. The dissimilarities observed between the results from both test collections raise doubts about the usage of these metrics pairs in future for measuring the reliability of system rankings. The results of the remaining metrics combinations, outside cluster and within a cluster are alike between the test collections.

For dataset2, Kendall’s tau correlation coefficient was measured between the system ranks of proposed method and the gold standard for metrics combination outside cluster and within cluster respectively. The proposed method system ranks were obtained from the mean rank, and tied ranks were resolved using their reliability scores, similar to dataset1. Table 5 shows Kendall’s tau correlation for the randomly selected topics sizes (topic size 50 uses all available topics for the collection). The tau values are mostly above 0.6, parallel dataset1. Only 6 of the tau values in dataset2 are below 0.6. The Kendall’s tau values are mostly moderate while some strong correlation coefficients were observed when the topic sizes are above 30. Then again, the tau values shown in Table 5 are from single random selection in contrast to the mean tau value shown in Fig.3 and Fig.4. There are possibilities of obtaining better or worse correlation coefficients with other random topic selections. Nonetheless, both datasets show comparable results for the proposed method.

Table 5: Kendall's tau correlation coefficient between system ranks of proposed method and original MAP for the TREC-2005 Robust track.

Cluster	Topic size	10	20	30	40	50
Outside cluster	AP@100-RBP@100	0.67	0.71	0.75	0.76	0.76
	AP@1000-RBP@1000	0.69	0.79	0.77	0.77	0.81
	AP@100-P@30	0.68	0.58	0.74	0.72	0.76
	AP@1000-P@10	0.57	0.72	0.68	0.71	0.74
	AP@100-P@100	0.46	0.78	0.74	0.80	0.81
	AP@1000-P@200	0.71	0.75	0.81	0.85	0.83
	RBP@100-P@30	0.70	0.60	0.60	0.67	0.73
	RBP@1000-P@10	0.51	0.63	0.65	0.67	0.65
	RBP@100-P@100	0.78	0.65	0.79	0.83	0.82
	RBP@1000-P@200	0.64	0.81	0.79	0.78	0.82
Within cluster	P@30-P@10	0.51	0.57	0.56	0.65	0.65
	P@100-P@200	0.66	0.71	0.80	0.82	0.85
	AP@1000-AP@100	0.42	0.71	0.73	0.81	0.82
	RBP@1000-RBP@100	0.63	0.68	0.70	0.72	0.72

5.0 CONCLUSIONS

A reliable retrieval system is crucial in satisfying users' need and this study has explored in measuring the reliability of system rankings. From this research, it was the metrics, AP and RBP are suitable choices in measuring the reliability of system ranks. These metrics pair when combined or used individually with different depth of evaluation determines highly reliable system ranks. Additionally, these metrics have the capability of generalizing and producing an equally comparable outcome with other similar metrics.

This study suggests the use of topic size 30 and above to obtain a strong correlation coefficient in addition to measuring large numbers of highly reliable systems. A strong positive correlation will indicate the true representation of the reliability score to the gold standard system ranks. Meanwhile, users can benefit from this retrieval systems due to their consistent performance.

The proposed method in this study is capable of measuring the reliability of individual retrieval systems using its ranking, and the versatility of the systems in satisfying multiple user needs. The study highlights the combination of effectiveness metrics which are suitable for measuring the reliability of ranking in information retrieval systems. Also, the reliability score is measured for individual systems from their topic ranks but the evaluation is dependable on relative ranking data among the systems. Nevertheless, ranking only makes sense if it is relatively measured. Even so, the study can be easily replicated and suited to other test collections for it utilizes relative ranking in measuring reliability. There lies an opportunity to investigate more than two metrics combination that would determine the diversity of a system's performance in satisfying multiple user need.

ACKNOWLEDGEMENTS

This material is based upon work supported by the University of Malaya and Ministry of Education, Malaysia under Grant No. ER027-2013A.

REFERENCES

[1] A. Moffat, F. Scholer, and P. Thomas, “Models and metrics:IR Evaluation as a User Process,” in *Proceedings of the Seventeenth Australasian Document Computing Symposium on - ADCS ’12*, 2012, pp. 47–54.

[2] G. De Melo and K. Hose, “Searching the web of data,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. LNCS 7814, pp. 869–873, 2013.

[3] K. Golub, D. Soergel, G. Buchanan, D. Tudhope, M. Lykke, and D. Hiom, “A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval,” *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 1, pp. 3–16, 2016.

[4] D. Hiemstra, “Information Retrieval Models,” *Inf. Retr. Search. 21st Century*, pp. 1–20, 2009.

[5] K. D. Varathan, R. Tengku Sembok, Tengku Mohd Abdul Kadir, and N. Omar, “Semantic Indexing For Question Answering System,” *Malaysian J. Comput. Sci.*, vol. 27, no. 4, pp. 261–274, 2014.

[6] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, “User Variability and IR System Evaluation,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’15*, 2015, pp. 625–634.

[7] V. Pavlu, S. Rajput, P. B. Golbus, and J. a. Aslam, “IR system evaluation using nugget-based test collections,” in *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM ’12*, 2012, pp. 393–402.

[8] C. Hauff, D. Hiemstra, F. Jong, and L. Azzopardi, “Relying on topic subsets for system ranking estimation,” *Proceeding 18th ACM Conf. Inf. Knowl. Manag. CIKM 09*, p. 1859, 2009.

[9] E. M. Voorhees and C. Buckley, “The Effect of Topic Set Size on Retrieval Experiment Error,” in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR’02*, 2002, pp. 316–323.

[10] J. Urbano, M. Marrero, and D. Martín, “A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation,” in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 925–928.

[11] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, and H. S. Thompson, “Repeatable and Reliable Search System Evaluation using Crowdsourcing,” *J. Web Semant.*, vol. 21, pp. 923–932, 2013.

[12] I. Ruthven, “Relevance behaviour in TREC,” *J. Doc.*, vol. 70, no. 6, pp. 1098–1117, 2014.

[13] E. M. Voorhees, “Variations in relevance judgments and the measurement of retrieval effectiveness,” *Inf. Process. Manag.*, vol. 36, no. 5, pp. 697–716, 2000.

[14] T. Sakai, “On the reliability of information retrieval metrics based on graded relevance,” *Inf. Process. Manag.*, vol. 43, no. 2, pp. 531–548, 2007.

[15] A. Baccini, S. Déjean, L. Lafage, and J. Mothe, “How many performance measures to evaluate information retrieval systems?,” *Knowl. Inf. Syst.*, vol. 30, no. 3, pp. 693–713, 2012.

[16] “Inter-rater reliability,” 2016. [Online]. Available: http://www.cookbook-r.com/Statistical_analysis/Inter-rater_reliability/. [Accessed: 28-Nov-2016].

[17] M. Hosseini, I. J. Cox, N. Milic-Frayling, M. Shokouhi, and E. Yilmaz, “An uncertainty-aware query selection model for evaluation of IR systems,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’12*, 2012, pp. 901–910.

[18] A. Moffat and J. Zobel, “Rank-biased precision for measurement of retrieval effectiveness,” *ACM Trans. Inf. Syst.*, vol. 27, no. 1, pp. 1–27, 2008.

[19] Y. P. Chua, *Mastering Research Statistics*, no. May. 2013.

[20] J. Hauke and T. Kossowski, “Comparison of Values of Pearson’s and Spearman’s Correlation Coefficients on the Same Sets of Data,” *Quaest. Geogr.*, vol. 30, no. 2, pp. 87–93, 2011.

[21] R. A. Fisher, *Statistical Methods for Research Workers*, Fourteenth. Oliver and Boyd, 1969.

[22] J. J. Bartko, “The intraclass correlation coefficient as a measure of reliability.,” *Psychol. Rep.*, vol. 19, no. 1, pp. 3–11, 1966.

[23] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: Uses in assessing rater reliability.,” *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.

[24] J. J. Bartko, “On Various Intraclass Correlation Reliability Coefficients,” *Psychol. Bull.*, vol. 83, no. 5, pp. 762–765, 1976.

[25] M. L. McHugh, “Interrater reliability: the kappa statistic.,” *Biochem. medica*, vol. 22, no. 3, pp. 276–282, 2012.

[26] R. Landers, “Computing (ICC) as Intraclass Estimates Correlations of Interrater Reliability in SPSS,” *The Winnower*, 2015. [Online]. Available: http://neoacademic.com/2011/11/16/computing-intraclass-correlations-icc-as-estimates-of-interrater-reliability-in-spss/. [Accessed: 26-Sep-2016].

[27] J. Algina, “Comment on Bartko’s ‘On various intraclass correlation reliability coefficients.,’” *Psychological Bulletin*, vol. 85, no. 1. pp. 135–138, 1978.

[28] P. Barrett, “Assessing the Reliability of Rating Data,” 2001. [Online]. Available: http://www.pbarrett.net/presentations/rater.pdf. [Accessed: 15-Mar-2016].

[29] E. M. Voorhees, “Overview of the TREC 2005 Robust Retrieval Track,” in *ACM SIGIR Forum*, 2006, vol. 40, no. 1.