RELEVANCE JUDGMENTS EXCLUSIVE OF HUMAN ASSESSORS IN LARGE SCALE INFORMATION RETRIEVAL EVALUATION EXPERIMENTATION

Prabha Rajagopal[1], Sri Devi Ravana[2], and Maizatul Akmar Ismail[3]
[1, 2, 3] Department of Information Systems
Faculty of Computer Science and Information Technology
University of Malaya, Kuala Lumpur, Malaysia

[1]prabz13@yahoo.com,[2]sdevi@um.edu.my,[3]maizatul@um.edu.my

ABSTRACT

Inconsistent judgments by various human assessors' compromises the reliability of the relevance judgments generated for large scale test collections. An automated method that creates a similar set of relevance judgments (pseudo relevance judgments) that eliminate the human efforts and errors introduced in creating relevance judgments is investigated in this study. Traditionally, the participating systems in TREC are measured by using a chosen metrics and ranked according to its performance scores. In order to generate these scores, the documents retrieved by these systems for each topic are matched with the set of relevance judgments (often assessed by humans). In this study, the number of occurrences of each document per topic from the various runs will be used with an assumption, the higher the number of occurrences of a document, the possibility of the document being relevant is higher. The study proposesa method with a pool depth of 100 using the cutoff percentage of >35% that could provide an alternate way of generating consistent relevance judgments without the involvement of human assessors.

Keywords: Information retrieval, relevance judgments, retrieval evaluation, large scale experimentation

1.0 INTRODUCTION

Information Retrieval (IR) is a way of obtaining information that is most relevant or related to a user's query from a collection of information. The IR evaluation is divided into two categories; user-based and system-based evaluation. IR user-based evaluation emphasizes on satisfaction of the user to the retrievals from their query while the IR system-based evaluation emphasizes on system effectiveness. In this study, we focus on the IR system-based evaluation which was popularized through the Cranfield methodology [4]. As the IR field continues to evolve, more techniques have arose. In the 90s, Web search engines used the partial match models and displayed the retrievals in a sorted ranking starting from the best matched [7]. While previous experiments had been conducted with smaller test collections, the Web has increased theneed for research in the IR field using larger test collections. The Web being a vast pool of test collection that is volatile makes it difficult for any sort of comparative research.

TREC is the first large scale test collection initiative undertaken by the National Institute for Standard and Technology (NIST) and U.S. Department of Defense. In TREC, researchers are provided with the document corpus and topic statements that are to be used to produce their own queries using automated or manual methods. These queries are then run against the collection of documents that have been provided. The output of these runs are then submitted back to TREC as official runs. Each participant is allowed to submit up to 1000 retrieved documents for each topic, ranking them from the most relevant to least relevant. The correct answers on the relevant documents for each topic are not known at the time of conducting the experiments. A document known as relevance judgments containing a list of all relevant documents to the topics given is created by experts of the topics. The relevance judgments is then used to evaluate the performance of each participating system. Fig. 1presents the evaluation cycle in TREC and the methodproposed in this study. The numbers in the figure indicates the sequence of steps in the evaluation cycle where 3a is the sequence from original TREC evaluation cycle while 3bis the sequence using the proposed method in replacement to human assessors in the original evaluation cycle.

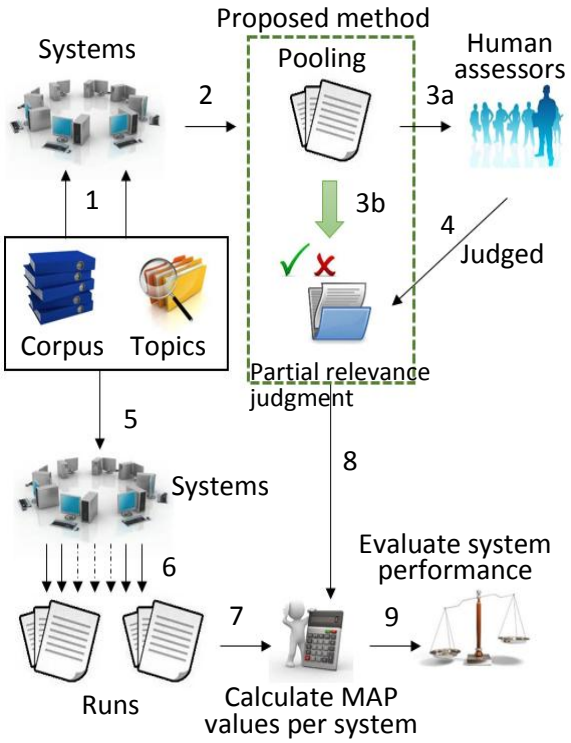


Fig.1: Evaluation cycle in TREC test collection and proposed method(designated with shaded arrow)

Due to large amount of documents, *pooling* is incorporated to reduce the number of documents that are to be judged by the judges, but the level of reliability when using such method is definitely questionable. Pooling only utilizes some documents to be judged for relevance, resulting in partial relevance. Zobel [17] had indicated in his experiment that pooling does not produce significant impact on the relevance judgment when system performances are judged. He also stated at least 50% to 70% of the relevant documents are identified when pooling is performed [17]. Although pooling does not consider all participating systems to generate the partial relevance judgments, results show reliable output when a sufficient pool depth of 100 is used [17].

The relevance judgment completes a test collection when matches are discovered between topics and documents. Generally, there are two ways of classifying relevancy:1) binary relevance judgments' uses the indication of 0 for irrelevant and 1 for relevant document; and 2) graded relevance judgments' indicates very relevant, relevant or not relevant [2]. Initially, the relevance judgments were complete but started to adopt to pooling methodrecently when the collection of documents start to increase in size [15]. In this study, relevancy of documents in the pseudo relevance judgments are classified based on binary relevance judgment due to the nature of the experiment conducted without the knowledge of the contents of the documents and topics from TREC.

To determine the performance of each retrieval system, there are some common measures that can beused, namely Precision at depth *k* (P@*k*) [14], average precision (AP) indicated by formula (1), normalized discounted cumulative gain (NDCG) [6] and rank-biased precision (RBP) [8]. Inthis study, mean average precision (MAP) is used which denotes the quality of a system in a single value whereby it is calculated based on the average precision value at each level of recall [7], indicated by formula (2).

$$AP@k = \frac{1}{R}\sum_{i=1}^{k} r_i \frac{\sum_{j=1}^{i} r_j}{i} \quad (1)$$

$$MAP@k = \frac{\sum_{q=1}^{|Q|} AP(q)}{|Q|} \quad (2)$$

In this paper, we proposed a method to automate the creation of relevance judgments by eliminatingthe involvement of human judges to create the relevance judgments in large scale test collections. The proposed method could result in eliminatingthe errorsinvolved through human assessment and produce a reliablemethod in IR system evaluation.

Starting with a review of the previousworks on errors introduced by human judges or assessors and relevance judgments created without human assessors, the paper focuses on using the number of occurrences of documents and marking relevancy based on first, cutoff percentage, second, exact count and third, different depth *k.* Finally, conclusions are drawn through the results and discussions and future work is proposed.

2.0 ERRORS INTRODUCED BY JUDGES (HUMAN ASSESSORS)

In the process of obtaining the relevance judgments, the human expertise is a reliable source of judgment. However, there are always the possibilities of errors introduced by humans during this process. Many studies have been conducted to analyze the level of errors and the tolerable levels injected by humans. Analyzing the human assessment error, Scholer, Turpin, & Sanderson [11] have stated inconsistency exists between the various topics, displaying assessment error at a high level. As time increases while performing assessment, the possibilities of introducing errors increase [12]. While judging consumes a lot of time, Scholer et al. [11] have associated the distance between two documents' matches with the amount of time between the judgments made and inconsistency increases as the distance between the duplicate pair increases as well. It has been investigated by Carterette & Soboroff [3] that judging relevant documents need more time compared to judging irrelevant documents.On another experiment, it has been stated that an error judgment made took longer time when compared to making correct judgments [12]. In either scenarios, the judges are prone to induce error judgments.

In the analysis where at least one of the document judged as relevant, the fraction of duplicates that were inconsistently judged was rather similar, ranging from 15% to 24%. According to Scholer et al. [11],"semantically similar documents in the relevance judgments are just as likely to be at least as inconsistently judged". In their study, multiple assessors were used to analyze the impact of the errors that could be introduced.Engaging different groups of assessors show low level of agreement in judging the relevancy [1]. Same documents that are being judged by different assessors tend to cause disagreement on the relevancy, whereby a low ranked document that has been marked relevant and high ranked document marked irrelevant cause disagreement from the second assessor [16]. It has been noted, the level of details provided in the topic specification does not seem to affect the errors introduced by the judges, instead previously judged similar documents have significant impact on the errors [3] [11].

3.0 RELEVANCE JUDGMENT WITHOUT HUMAN ASSESSOR

Due to variable judgments by human assessors, studies have been conducted to reduce or eliminate the involvement of human judges in generating the relevance judgments. Soboroff, Nicholas, & Cahan [13] had suggested a random selection of documents from the pool. The selection is done based on the average number of relevant documents for each topic in the pool. Mean and standard deviation for a particular year of TREC is then used to randomly select the number of documents to form the pseudo relevance judgments, whereby pseudo relevance judgments is a document created similar to the relevance judgments in TREC. A total of 50 repetition and system rankings were performed and each was correlated with the system rankings using the original relevance judgments from TREC. Refer to Fig. 2 for illustrations of steps explained above.

In another method, the exact-fraction sampling of relevant document occurrences in each topic was used to populate the pseudo relevance judgments [13]. This method draws exact numbers of documents per topic based on the percentage calculated from the original relevance judgments (Refer to Fig. 3). It is also known that each topic has varied amount of relevancy. The outcome of the latter method performed slightly better than the random selection method.

While TREC removes duplicate documents from its pool [15] before the judgment process, Soboroff, Nicholas, & Cahan [13] experimented with sustaining them in the pool to allow higher probabilities of being chosen during the

random selection. It was mentioned that when more than one system has retrieved a document, it could most likely be relevant as well.

Nuray and Can [9] performed an experiment to generate the relevance judgment automatically using the heuristics method. The method was performed to replicate the imperfect web environment where the original relevance judgment used was modified to suit the web like scenario. They performed pooling and ranked the documents based on the similarity scores using the vector space model. Their experiment resulted in Kendall's tau correlation for average precision and precision at DCV (Document Cutoff Value) appearing to be better for pool depth of 30 compared to pool depth of 200 [9]. The average precision correlation for pool depth of 200 between the automatic method and human judged relevance judgment ranges between 0.325 and 0.377 [9]. The computed correlation did not produce a strong correlation when compared to the methods proposed by Soboroff et al. [13]. The authors had also randomly selected top 10 documents from some systems to form the pool and repeated the selection 10 times before computing the AP correlation which resulted Kendall's tau correlation that was not as strong [9] as that proposed by Soboroff et al. [13].

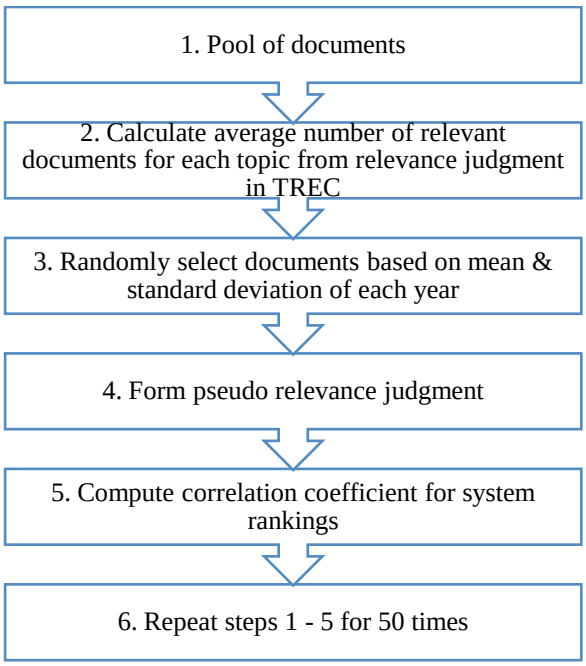


Fig.2: Graphical illustrations ofsteps taken in random selection method

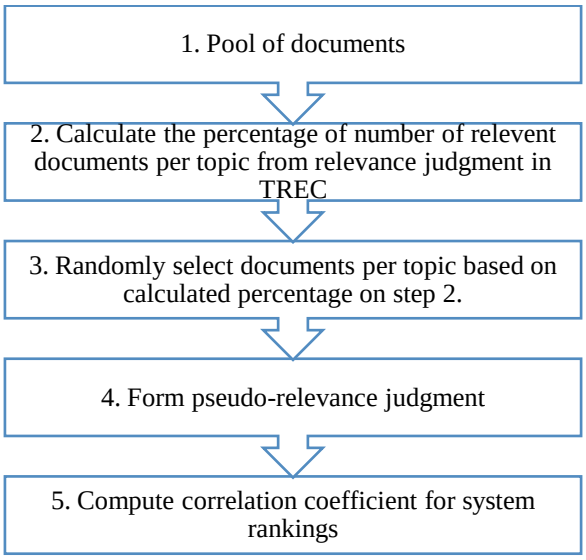


Fig.3: Graphical illustrations ofsteps taken in exact-fraction sampling method

Cormack, Palmer, & Clarke [5] have experimented a method known as move-to-front (MTF) to estimate the likelihood of relevancy of a document. It is assumed that if a particular submission produces a relevant document, the subsequent document from the same submission would also be relevant. Cormack et al. [5] considered two variations in MTF; global and local whereby the earlier uses all submissions for all topics while the latter uses a depth of *k* to each topic. It is concluded that both the global and local MTF method are more effective compared to the common pooling method [5].

4.0 MOTIVATION

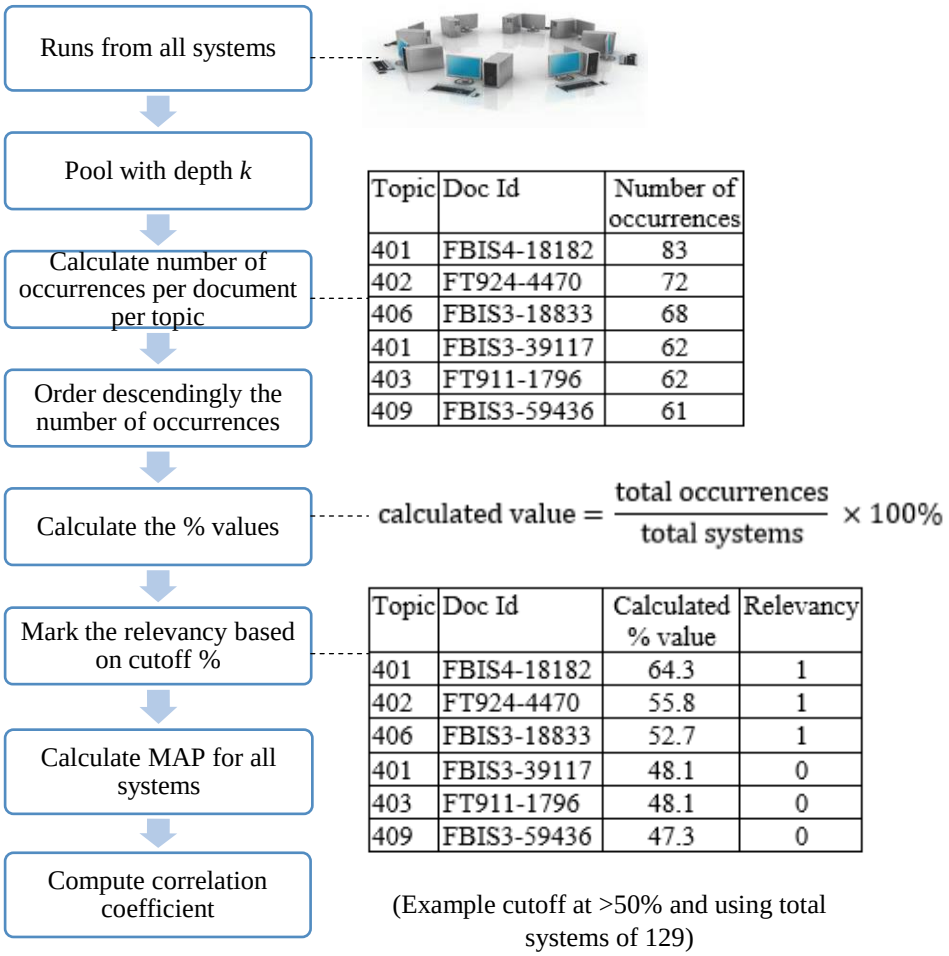
The web provides large amount of information and various search engines to retrieve documents according to user's query. Although evaluation of retrieval systems in the web could be ideal but the volatile nature of ever-changing web contents and retrieval systems makes it challenging for comparative studies and repeatability. Large test collections that are static such as TREC provides a platform for web-like scenarios and facilitate comparative system evaluations. To determine the effectiveness of the retrieval systems, human judges are used to identify the relevant

documents for each topic similar to that done by TREC. It is assumed that human assessment is a reliable source but the involvement of human assessors in generating the relevance judgments could be inconsistent and may contain errors that affect the reliability of the relevance judgment created.

While eliminating human assessors could avoid inconsistencies and human errors, an alternate method that produces a reliable set of relevance judgments similar to that generated by human assessors is needed. It is the motivation of this study to propose a method to generate a reliable set of relevance judgments without human judges through automation. The automated method would be consistent in evaluating information retrieval systems in large scale experimentation.

5.0 EXPERIMENT

The method employed in this experiment uses one important parameter from the system runs which is the number of occurrences of a document.



Topic	Doc Id	Number of occurrences
401	FBIS4-18182	83
402	FT924-4470	72
406	FBIS3-18833	68
401	FBIS3-39117	62
403	FT911-1796	62
409	FBIS3-59436	61

Topic	Doc Id	Calculated % value	Relevancy
401	FBIS4-18182	64.3	1
402	FT924-4470	55.8	1
406	FBIS3-18833	52.7	1
401	FBIS3-39117	48.1	0
403	FT911-1796	48.1	0
409	FBIS3-59436	47.3	0

Fig.4: Graphical illustrations of steps taken in computing number of occurrences per topic from pooling and marking of relevancy using cutoff percentage method

Using a pool depth of k, occurrences of each document per topic is obtained and used as the main measure for identifying relevant or irrelevant documents. It is assumed that the higher the number of occurrences of a document, the probability of it being relevant is high. In generating the pseudo relevance judgments, TREC runs from the ad-hoc track in TREC-8 and Web-track from TREC-9 were used whereby both selected test collections have 50 topics each. In TREC, documents are pooled from each topic per system runs, de-duplicated, sorted by topic then

alphabetically by document ID before being presented to human assessors for judgment. A variation to the TREC practice is done in this experiment to generate the relevance judgments. Once pooled, the documents are directly used to create the pseudo relevance judgments using the proposed methods without being assessed by human assessors (Refer to Fig. 1). There are two main methods conducted in a laboratory based experiment to generate the pseudo relevance judgments; (i) cutoff percentage and (ii) exact count. In general, a pool depth of 100 was used but it was interesting to know if pool depth of 200 could affect the system performance. Fig. 4 indicates the steps undertaken to create the pseudo relevance judgment using the cutoff percentage method where the documents are marked based on the calculated percentage values. Once the documents have been sorted based on the calculated percentage values, the cutoff percentage value is used to determine which documents will be judged as relevant or irrelevant. The pseudo relevance judgment created using cutoff percentage of >50% calculated percentage value marks all documents that have calculated percentage value of >50% as relevant while the remaining are judged as irrelevant. Whereas for the pseudo relevance judgment created using cutoff percentage of >35% calculated percentage value marks all documents with the calculated percentage value of >35% as relevant while the remaining as irrelevant. The marking of these documents are independent of topic but stresses on the usage of calculated percentage value based on number of occurrences of each document.

Fig. 5 indicates the steps taken to create the pseudo relevance judgment using the exact count method. Exact count per topic occurrences was used whereby for each topic, exact number of relevant documents from the original relevance judgment were counted and used to mark relevancy of documents for the pseudo relevance judgment. First, pooling is done with a depth of 100. Then the number of occurrences per document per topic is calculated. The pooled documents are then ordered ascending by topic, followed by descending order of number of occurrences. Then the percentage value for each of the documents is calculated. Finally, the exact count based on each topic derived from the original relevance judgment is identified and the same numbers are selected for each topic to form the pseudo relevance judgment.

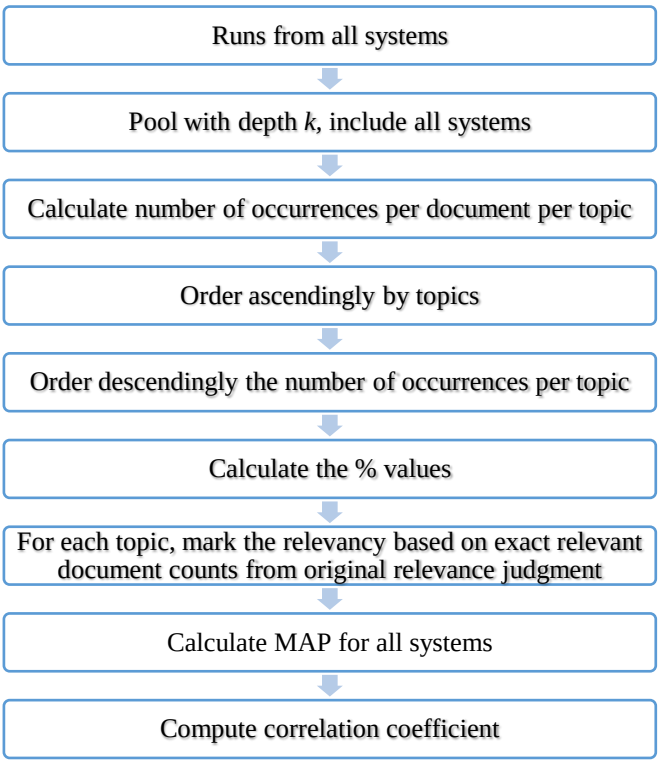


Fig.5: Illustration of steps taken for exact count method

Table *1*shows an example of judging exact number of documents based on number of relevant document occurrences from the original relevance judgment. The first column indicates the number of documents for each topic and the colon indicates more documents symbolically. The documents for each topic is sorted descending based on the calculated % value column. For example, topic 401 contains 40 relevant documents in the original relevance judgment. Hence, the column relevancy indicates the first 40 documents of topic 401 as 1 for relevant and remaining documents as 0 for irrelevant in the pseudo relevance judgment. Similarly, topic 402 has 33 relevant documents in original relevance judgment while the pseudo relevance judgment too marks the same numbers as shown in relevancy column but the document that is judged relevant or irrelevant may be different when compared to the original relevance judgments.

With the generated pseudo relevance judgment, the MAP is calculated for all the systems and compared with the MAP values calculated from the original relevance judgment. Two different correlation coefficients are used to measure the outcome of the experimentation done. Kendall's tau utilizes the generated system orderings or rankings but not the scores that steered to that ordering and it allows the "closeness" of system ranking pairs to be quantified [10].Kendall's τ is based on the number of concordant and discordantpairs between pairs of observations. A pair is a concordant pair, if the relativerank ordering for the pair is the same in both rankings (for example, $X_i>X_j$ and $Y_i>Y_j$ or $X_i<X_j$and $Y_i<Y_j$). In contrast, if the ranks disagree, thepair is considered as a discordant pair (for example, $X_i>X_j$ and $Y_i<Y_j$ or if $X_i<X_j$and $Y_i>Y_j$). If n is the number of observations in each randomvariable, Kendall's τ is:

$$\tau = \frac{(number\ \ of\ \ concordant\ \ \ pairs\ ) - (number\ \ of\ \ discordant\ \ \ pairs\ )}{\frac{1}{2}n(n-1)} \qquad (3)$$

Whereas, Pearson correlation coefficient is a measure of linear correlation between two variables x and y. This correlation is based on scores rather than the system rankings [10]. Pearson correlation oftendenoted as r. To compute r, let $i\in\{1,...,n\}$ represent the ranked items, and X and Y the two system orderings, such that $X_i$is the score that item i achieves in ranking X, and Y_i its score in ranking Y. Pearson's r is:

$$r = \frac{\sum_{i=1}^{n}(\bar{X} - X_i)(\bar{Y} - Y_i)}{\sqrt{\sum_{i=1}^{n}\left((\bar{X} - X_i)\right)^2}\sqrt{\sum_{i=1}^{n}\left((\bar{Y} - Y_i)\right)^2}} \qquad (4)$$

Table 1: Example of marking relevancy of documents in pseudo relevance judgment based on number of relevant documents from original relevance judgment

Number of documents	Topic	Document	Calculated % value	Relevancy
1	401	FBIS4-18182	61.97	1
:	401	FBIS3-18833	53.52	1
:	401	FBIS3-39117	53.52	1
40	401	FBIS4-55036	50.70	1
41	401	FBIS3-817077	47.89	0
:	401	FBIS3-59436	45.07	0
1	402	LA101290-0115	83.10	1
:	402	LA101690-0121	81.69	1
33	402	FT921-11140	80.28	1
34	402	LA062590-0042	80.28	0
:	402	LA080190-0099	78.87	0
:	402	FT923-7735	77.46	0

Both correlation coefficients can hold values between -1 and 1. It is known that the perfect correlation is 1.0, hence the closer the correlation values to 1.0 is better and a value of 0.8 and above is usually sufficient to accept the

reliability of the proposed method in IR evaluation experiments. Based on the previous experiments that have been conducted to automate the creation of relevance judgment, a correlation of approximately 0.5 [13] is an acceptable correlation.

6.0 RESULTS AND DISCUSSION

In the following sections, the correlation between original relevance judgment and groups of systems with similar performance level has been calculated for all participating systems that has been divided into 3 subsections; good, moderate and low performing systems based on retrieval effectiveness scores of the systems using original relevance judgments. The grouping of systems were done to identify if the methods proposed perform better for a specific group of systems. For TREC-8, there were 129 systems with each groupconsisting of 43 systems. For TREC-9, there were 104 systems used with goodand lowperforming groupswith35 systems while the moderate performing grouphas 34 systems.

6.1. Cutoff Percentages

6.1.1. Marked Relevant For Percentage >50% Occurrences

The correlation coefficient results of TREC-8 and TREC-9 were compared to system rankings of original relevance judgment where both has similar performance level. Table 2 displays the Kendall's tau and Pearson correlation for TREC-8 and TREC-9 for the cutoff percentage (>50% and >35%) method that have been performed in this experiment.Kendall's tau correlation for cutoff percentage >50% has a reasonably strong Kendall's tau correlation, 0.638 and a strong Pearson correlation, 0.822.This is possible because the non-contributing systems are contributing to the set of relevance judgments which may change the rankings of the systemsanddid not contribute to the pool for original relevance judgments.

Table 2: Kendall's tau and Pearson correlation for MAP values for depth 100 using cutoff percentage method for TREC-8 and TREC-9

Methods	**Kendall's tau**		**Pearson**	
	TREC-8	TREC-9	TREC-8	TREC-9
>50% occurrences	0.506	**0.638**	0.739	**0.822**
>35% occurrences	0.515	**0.663**	0.736	**0.836**

According toTable 3, Kendall's tau correlation for 3 subsections indicates that the low performing systems having high correlation coefficient of 0.8 and above. Low performing systems are systems which have scored low in performance measurement. The nature of low performing systems in TREC is that they have very less relevant documents retrieved that match the relevance judgments. It would have been the case when using the pseudo relevance judgments as well where the proposed method did not mark documents as relevant unnecessarily to give these low performing systems high scores. On the other two subsections, good and moderately performing systems for the cutoff percentage of >50% falls lower than 0.4. The correlation is not as strong as the low performing systems. This could be the case because the non-contributing systems are contributing to the pseudo relevance judgments and impacting the goodand moderately performing systems ranking a lot.

Table 3: Kendall's tau correlation for 3 subsections for depth 100 using cutoff percentage method for TREC-8 and TREC-9

	Methods	**Good performing systems**	**Moderately performing systems**	**Low performing systems**
TREC-8	>50% occurrences	-0.296	0.409	**0.835**
	>35% occurrences	-0.276	0.402	**0.822**
TREC-9	>50% occurrences	-0.035	0.144	**0.916**
	>35% occurrences	0.098	0.184	**0.915**

As presented in Table 4, the Pearson correlation computed for the low performing systems displays a very strong correlation of 0.8 and above for the proposed method of cutoff percentages >50%. The moderate performing systems did not score high as the system rankings using the pseudo relevance judgment has affectedtheir performance. This could be the results of pooling documents from the contributing and non-contributing systems in TREC. In TREC, all non-contributing systems documents are marked as irrelevant while using the proposed method in this experiment, those documents could have been marked as relevant due to higher retrieval occurrences. The good performing systems for TREC-8 has a strong negative correlation using the cutoff percentage method with >50% occurrences. The negativity indicates that while system scores using original relevance judgment was decreasing, system scores using pseudo relevance judgment was increasing. This pattern can be noticed in Fig. 6. When compared to correlation coefficient from TREC-9 for the good performing systems, similar output was not obtained. It can be clearly stated that cutoff percentage of >50% occurrences is able to rank the low performing systems well since the correlation coefficient for both test collections are strong.

Table 4: Pearson correlation for 3 subsections for depth 100 using cutoff percentage method for TREC-8 and TREC-9

	Methods	**Good performing systems**	**Moderately performing systems**	**Low performing systems**
TREC-8	>50% occurrences	-0.840	0.592	**0.947**
	>35% occurrences	-0.846	0.637	**0.949**
TREC-9	>50% occurrences	-0.344	0.160	**0.949**
	>35% occurrences	-0.232	0.251	**0.952**

For both TREC-8 and TREC-9, system rankings computed using the pseudo relevance judgments appear to produce a close linear correlation with the original relevance judgments as shown in Fig. 6 and Fig. 7. The space between the original system rankings and the linear line from the proposed cutoff percentage (>50%) method is closer for TREC-9 compared to TREC-8, which produced a higher correlation coefficient in TREC-9.

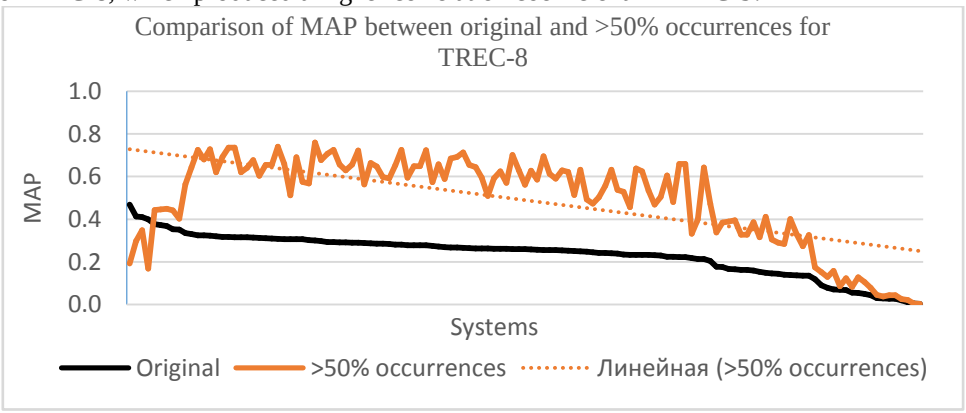


Fig.6: Comparison of MAP values between original and pseudo relevance judgment using cutoff percentage of >50% occurrences with pool depth 100 for TREC-8

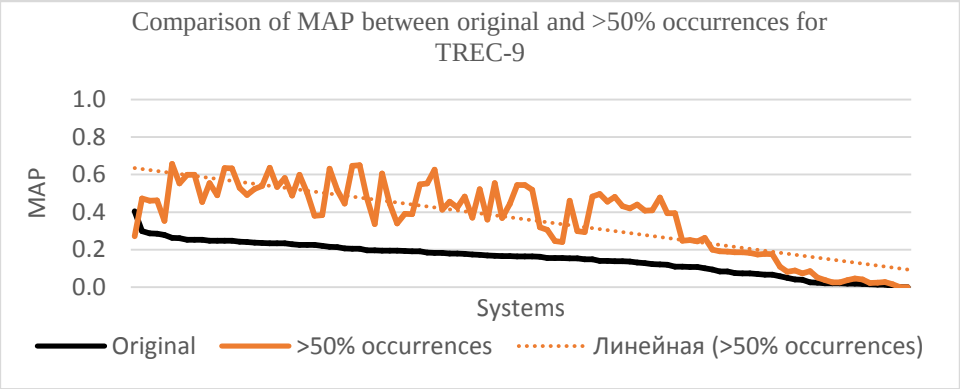


Fig.7: Comparison of MAP values between original and pseudo relevance judgment using cutoff percentage of >50% occurrences with pool depth 100 for TREC-9

6.1.2. Marked Relevant For Percentage >35% Occurrences

Overall, the proposed method with cutoff percentage of >35% occurrences calculation has a better Kendall's tau and Pearson correlation when compared to the cutoff percentage of >50% for TREC-8 and TREC-9 respectively (Refer to Table 2 above).The probable reason for such results could be due to the better identification and higher numbers of relevant documents in the pseudo relevance judgments. Although the documents marked as relevant in pseudo relevance judgment could be different from those in the original relevance judgment, it does not impact the correlation coefficient of the system rankings. It is more important to rank the systems in the same order as obtained from original relevance judgment to show that the proposed methods to create pseudo relevance judgment is effective and reliable.

As shown inTable 3 above, Kendall's tau correlation coefficient for the low performing systems appears strong between 0.8 and 0.9, while the other two subsections, good and moderately performing systems, for both TRECs, the values range lower than 0.4 for cutoff percentage of >35%. Similar to cutoff percentage of >50% occurrences, the Kendall's tau correlation is not strong when compared to the low performing systems which could have been impacted with the additional contributing documents from the non-contributing systems.

As presented inTable 4 above, the Pearson correlation computed for the low performing systems display a very strong correlation for both TREC-8 and TREC-9, scoring at a high 0.949 and 0.952 for cutoff percentage of >35%. When compared to the cutoff percentage of > 50%, it can be noted that more documents marked as relevant in the cutoff percentage of >35% produces better Pearson correlation. In addition, the correlations for the good and moderately performing systems show a similar trend to cutoff percentage of >50% occurrences although their correlations are better. It can be concludedthat having more documents marked as relevant could produce a better correlation. Fig. 8and Fig. 9 presented the plotted graphs and note the space between the original system rankings and the linear line. It is found that the proposed cutoff percentage (>35%) method is closer for TREC-9 compared to TREC-8.

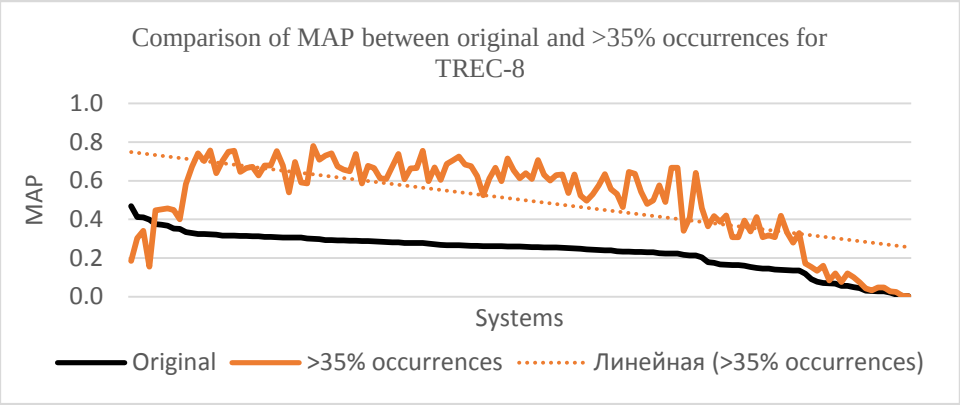


Fig.8: Comparison of MAP values between original and pseudo relevance judgment using cutoff percentage of >35% occurrences with pool depth 100 for TREC-8

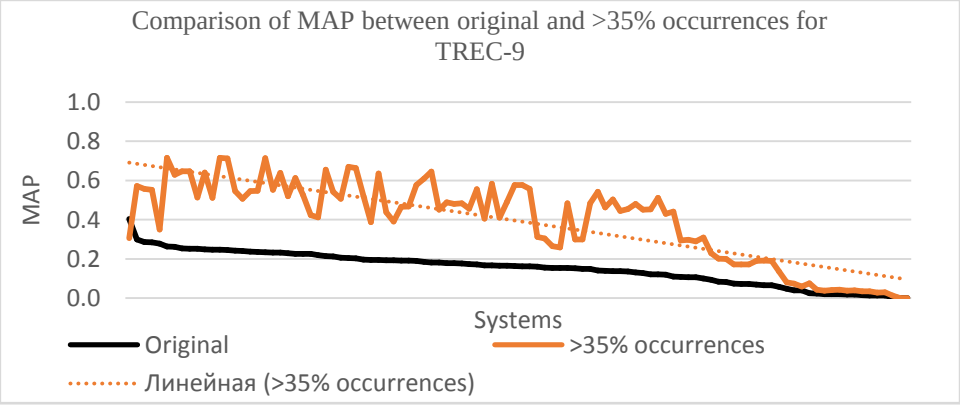


Fig.9: Comparison of MAP values between original and pseudo relevance judgment using cutoff percentage of >35% occurrences with pool depth 100 for TREC-9

6.2. Exact Count Of Number Of Relevant Documents Per Topic Occurrences

Another method using exact count of number of relevant documents per topic occurrences was experimentedbecause it is known that not all the topics have the same number of relevant documents. Similar experiment using the exact-fraction sampling of relevant documents from original relevance judgment was performed by Soboroff et al. (2001). However, in this study, an attempt to select exactly the same numbers of relevant documents were experimented to identify if the system performance is different. Using TREC-8, a depth of 100 was used for pooling and calculated percentage values were used to mark the relevancy into forming the pseudo relevance judgment as indicated in Fig. 5. Using this method, the Kendall's tau correlation is 0.504 and Pearson correlation is 0.733 for TREC-8, whereby both the correlations are lower compared to the two cutoff percentage methods discussed earlier (>35% and 50%).Although there are equal number of relevant documents, the matching relevant documents with the relevance judgments from TREC is 36.5% from a total of 4728 relevant document. This indicates that it is not necessarily the same number of relevant documents could result in better correlations.

According to Table 5, the low performing systems have strong correlations similar to the previous method using the cutoff percentages. Meanwhile, the Pearson correlation for the good performing systems appears to be decreasingly linear but strong. From Fig. 10, it can be noted that as the plotted graph for original MAP values decreases, the graph for pseudo relevance judgment MAP values are increasing, hence giving a negative linear correlation. The moderately performing systems have a relatively close linear Pearson correlation. Although from the graph, the MAP values using pseudo relevance judgment appears to be not correlated, it has a decreasing linear trend that matches with the original.

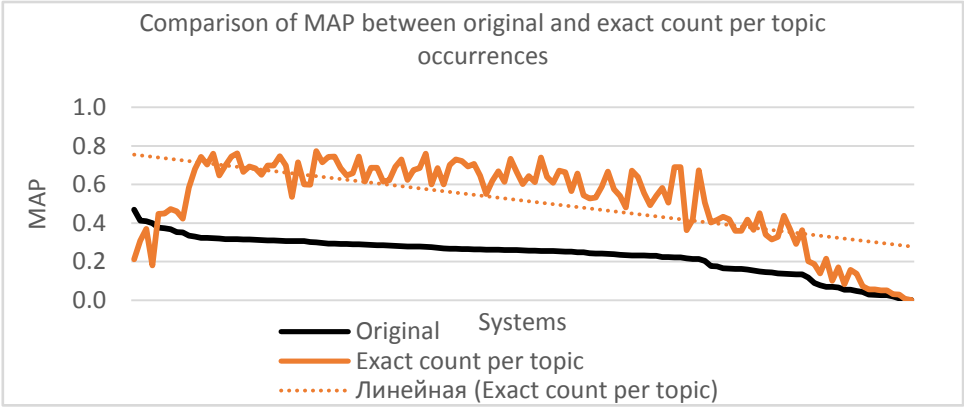


Fig.10: Comparison of MAP values between original and exact count per topic occurrences for pool depth 100

Table 5: Kendall's tau and Pearson correlation for 3 subsections for depth 100 using TREC-8

Correlation method	**Good performing systems**	**Moderately performing systems**	**Low performing systems**
Kendall's tau	-0.307	0.415	**0.820**
Pearson	-0.851	0.605	**0.951**

6.3. Different Depth At Cutoff *k*

Using different depth in pooling could include additional documents and affect the computation of the occurrences of each document. Hence, a depth of 200 was experimented in addition to the standard depth of 100 using TREC-8 for the cutoff percentages of >50% occurrences. Based on Fig. 11, the graph for different depth is almost overlapping each other with slight differences in the system rankings scores. According to Table 6, the Kendall's tau correlation using pool depth 100 and 200 is almost equal and shows no significant impact when the depth is increased. Meanwhile, the Pearson correlation is lower from 0.739 to 0.729 when the depth for pooling was increased to 200, showing a very slight drop in linear correlation but does not show much significant impact.

Table 6: Kendall's tau and Pearson correlation for different depth of pooling

Methods	**Depth 100**	**Depth 200**
Kendall's tau	0.506	0.507
Pearson	0.739	0.729

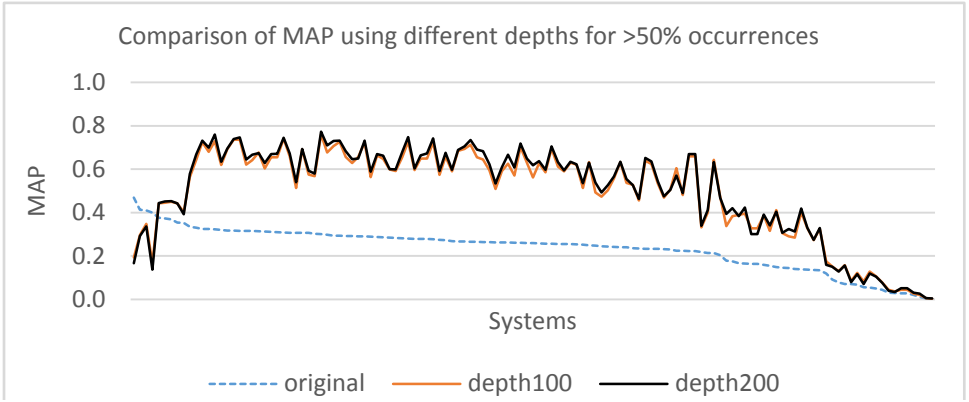


Fig.11: Comparison of MAP values using original and pseudo relevance judgment with different pool depths for cutoff percentage of >50% occurrences

(Note: The graph for pool depth 100 and depth 200 is overlapping and not clearly differentiable)

As presented in Table 7, the 3 subsections for both the pool depth 100 and 200 does not deviate much and is within 0.04 deviation. Although there are some differences in the correlation values, the depth does not seem to bring significant impact to the computed correlation.

Table 7: Kendall's tau and Pearson correlation for 3 subsections for different pool depths using TREC-8

Methods	**Depth *k***	**Good performing systems**	**Moderately performing systems**	**Low performing systems**
Kendall's tau	100	-0.277	0.415	**0.815**
	200	-0.296	0.409	**0.835**
Pearson	100	-0.849	0.637	**0.949**
	200	-0.840	0.592	**0.947**

7.0 CONCLUSION

Two main methods have been explored into automating the creation of relevance judgment to eliminate the human effort and errorsinvolved through human judgments.

Based on the Kendall's tau correlation for pooling depth 100 the cutoff percentage value method has higher correlations with the benchmarked current approach compared to exact count method.Comparing withthe previous study which uses the random selection method in generating the relevance judgments(conducted by Soboroff et al., (2001)),it produced a Kendall's tau value of 0.459 using TREC-8. Using similar TREC collection, the proposed method in this paper produced a much better Kendall's tau value of 0.515. This proposed method is called the cutoff percentage method.

In the subdivision of similar performing systems, low performing systems correlate positively with the original systems ranks and the proposed methods having high values of more than 0.8. The good and moderately performing systems didn't correlate as well as the low performing systems. Pooling that was done with the non-contributing and contributing systems from TREC has now impacted a change in the system rankings of the good and moderately performing systems, while the scores of low performing systems remained somewhat similar because the relevant documents were not marked unnecessarily in pseudo relevance judgments.

Meanwhile, increasing the pool depth to 200 did not produce significant difference in the correlation coefficientas the results seconds previous findings that pool depth of 100 is sufficient to produce reliable output [17]. On the other

hand, comparisons with the previous study shows improvement in Kendall's tau to 0.507 as opposed to 0.351 obtained by Nuray and Can (2003). Correlations for the 3 different groups of similar performance systems shows the low performing systems for depth 200 has high correlation of 0.8 and above similar to pool depth of 100. This could have been contributed due to the additional documents that appear in the pseudo relevance judgment when the depth was increased.

From the findings it can be concluded that generating a set ofrelevance judgments without involving human assessors using the proposed cutoff percentage methods could function as an alternative technique in measuring the performance of the retrieval systems (correlation coefficient between system rankings generated using traditional and proposed method was 0.6 and above in this study). Theexperiments could be expanded further to explore if this new method proposed functions better when paired with smaller number of topics or even when other evaluation metrics are used such as standardized metrics. In addition to that, different pool depths could also be experimented to observe the correlations.

ACKNOWLEDGEMENTS

This research was funded through University of Malaya research grant – UMRG (RG093/12ICT).

REFERENCES

[1] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., Vries, A. P., & Yilmaz, E.,"Relevance Assessment: Are Judges Exchangeable and Does it Matter?"in *The 31st Annual International ACM SIGIR Conference*,Singapore, ACM, 2008,pp. 667-674.

[2] Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I., "Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements", in*The 30th Annual International ACM SIGIR Conference*, Amsterdam, ACM, 2007, pp. 63 - 70.

[3] Carterette, B., & Soboroff, I.,"The Effect of Assessor Errors on IR System Evaluation" in*The 33rd Annual ACM SIGIR Conference*,Geneva, ACM, 2010, pp. 539 - 546.

[4] Cleverdon, C., "The Significance of the Cranfield Tests on Index Languages", ACM, 1991, pp. 3 - 12.

[5] Cormack, G. V., Palmer, C. R., & Clarke, C. L.,"Efficient Construction of Large Test Collections" in*The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, ACM, 1998, pp. 282 - 289.

[6] K. Järvelin and J. Kekäläinen., "Cumulated gain-based evaluation of IR techniques"*ACM Transactions on Information Systems, 20(4)*, 2002, pp. 422–446.

[7] Mandl, T.,"Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance",*Informatica 32*, 2008, pp. 27 – 38.

[8] Moffat, A., & Zobel, J., "Rank-biased precision for measurement of retrieval effectiveness",*ACM Transactions on Information Systems*, 2008, pp. 1-27.

[9] Nuray, R., & Can, F., "Automatic Ranking of Retrieval Systems in Imperfect Environments", in*Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, ACM, 2003, pp. 379 – 380.

[10] Ravana, S. D., & Moffat, A.,"Score Aggregation Techniques in Retrieval Experimentation",in *Conferences in Research and Practice in Information Technology (CRPIT), Vol. 92*, New Zealand, 2009, pp. 57 – 66.

[11] Scholer, F., Turpin, A., & Sanderson, M., "Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements", in*The 34th Annual ACM SIGIR Conference*, Beijing, ACM, 2011, pp. 1063 – 1072.

[12] Smucker, M. D., & Jethani, C. P., "Time to Judge Relevance as an Indicator of Assessor Error", in*Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Oregon, ACM, 2012, pp. 1153-1154.

[13] Soboroff, I., Nicholas, C., & Cahan, P., "Ranking Retrieval Systems without Relevance Judgments", in*The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, ACM, 2001, pp. 66 – 73.

[14] T. Joachims, L. Granka, B. Pan, H.Hembrooke, and G. Gay.,"Accurately interpreting click through data as implicit feedback",*The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, ACM, 2005, pp. 154–161.

[15] Vorhees, E. M., "Overview of TREC 2007", in*The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2007.

[16] Webber, W., Chandar, P., & Carterette, B., "Alternative Assessor Disagreement and Retrieval Depth", in*The 21st ACM International Conference on Information and Knowledge Management*, Maui, ACM, 2012, pp. 125-134.

[17] Zobel, J., "How Reliable are the Results of Large-Scale Information Retrieval Experiments?" in*The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, ACM, 1998, pp. 307 – 314.