Aslib Journal of Information Management

Ranking retrieval systems using pseudo relevance judgments
Sri Devi Ravana Prabha Rajagopal Vimala Balakrishnan

Received 25 March 2015
Revised 13 September 2015
Accepted 23 September 2015

Ranking retrieval systems using pseudo relevance judgments

Sri Devi Ravana, Prabha Rajagopal and Vimala Balakrishnan
Department of Information Systems,
Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia

Abstract

Purpose – In a system-based approach, replicating the web would require large test collections, and judging the relevancy of all documents per topic in creating relevance judgment through human assessors is infeasible. Due to the large amount of documents that requires judgment, there are possible errors introduced by human assessors because of disagreements. The paper aims to discuss these issues.

Design/methodology/approach – This study explores exponential variation and document ranking methods that generate a reliable set of relevance judgments (pseudo relevance judgments) to reduce human efforts. These methods overcome problems with large amounts of documents for judgment while avoiding human disagreement errors during the judgment process. This study utilizes two key factors: number of occurrences of each document per topic from all the system runs; and document rankings to generate the alternate methods.

Findings – The effectiveness of the proposed method is evaluated using the correlation coefficient of ranked systems using mean average precision scores between the original Text REtrieval Conference (TREC) relevance judgments and pseudo relevance judgments. The results suggest that the proposed document ranking method with a pool depth of 100 could be a reliable alternative to reduce human effort and disagreement errors involved in generating TREC-like relevance judgments.

Originality/value – Simple methods proposed in this study show improvement in the correlation coefficient in generating alternate relevance judgment without human assessors while contributing to information retrieval evaluation.

Keywords Information retrieval, TREC, Batch evaluation, Large-scale experimentation, Relevance judgments, Retrieval evaluation

Paper type Research paper

1. Introduction

Information retrieval (IR) indicates the retrieval of unstructured records that consists mainly of free-form natural language text (Greengrass, 2000). It is a way of obtaining information that is most relevant or related to a user's query from a collection of information. Unstructured records are those documents that do not have a specific format where the information is presented. When a huge amount of information is available, retrieval of the related material is quite crucial. Ideally, the main target of an IR system should be to provide information as accurate as possible based on the user's query and that information being relevant to the user.

The two main categories of IR evaluation are user-based evaluation and system-based evaluation (also known as batch evaluation). The user-based approach focusses on the users' interaction with the IR systems and the benefit from the IR systems (Hersh *et al.*, 1995), while the system-based evaluation focusses on measuring system effectiveness in a non-interactive laboratory environment (Ravana, 2011). The system-based evaluation is

This material is based upon work supported by the University of Malaya Research Grant Program (RP028E-14AET) and Exploratory Research Grant Scheme (ER027-2013A).

Aslib Journal of Information Management
Vol. 67 No. 6, 2015
pp. 700-714

2050-3806
DOI 10.1108/AJIM-03-2015-0046

also more widely used when compared to user-based experiments and has been the leading standard in the past 30 years (Turpin *et al.*, 2009). Due to the impractical effort in judging the usefulness of the search retrieval in user-based evaluations, the system-based IR evaluation takes precedence (Mandl, 2008). The system-based evaluation advantages include easily reproducible results that make it suitable for comparative studies, lesser time in experimentation, and being less costly than user-based experiments.

Due to the lack of consistency in performing IR evaluation in real time on the web, laboratory experiments in the IR field offer regularities for evaluation. The laboratory test collections increased in size, although not as large as those found in operational systems, and contained document proxies with title and abstract or, in some, only the titles (Rasmussen, 2002). When the web boomed in the 1990s, there was a need for larger test collections when the existing test collections became insufficient (Ravana, 2011). Ad-hoc retrieval is searching for relevant documents for an earlier unknown topic using a static collection. The authors (Turpin *et al.*, 2009) stated that evaluation of this ad-hoc retrieval would need a collection of documents, a set of topics or queries that represent the need of the user, and a set of relevance judgments that indicate the relevancy of each document for each query.

Human assessors who judge the relevancy of each document per topic from pooling create the relevance judgment. However, human assessors tend to introduce errors during the judgment process (Scholer *et al.*, 2011) and have varied judgment decisions for the same document (Bailey *et al.*, 2008; Webber *et al.*, 2012). This research, on the other hand, aims to reduce these human assessors' effort and directly create the relevance judgment from pooling. We foresee the following benefits:

(1) Reduce human effort involved in IR evaluation by incorporating automated methods without human assessors to generate the relevance judgments.

(2) Avoid biases to any group of systems that happens through traditional pooling where the selection of documents for judgment is only from contributing systems while omitting documents from non-contributing systems. The proposed methods in this research include both the contributing and non-contributing systems during the pooling. This approach would avoid biasness that happens through pooling in Text REtrieval Conference (TREC). It also contributes to fair effectiveness scores for the systems.

Starting with a research background from previous works on errors introduced by human assessors and alternative methods to generate relevance judgments, the paper focusses on using the number of occurrences of documents per topic and documents' ranking. First, judging document relevancy uses exponential variation method; and second, the document ranking method. Then, the results and discussion from experimentation are included. Finally, the conclusion is drawn, and the future work is proposed.

2. Research background

Generating relevance judgments for large-scale test collections through human assessors consumes a lot of time and is prone to induce errors during judgments (Scholer *et al.*, 2011; Smucker and Jethani, 2012). Relevance judgments generated by human assessors for large-scale test collections may not be feasible and possibilities to reproduce are slim due to varying judgment decisions at various times by different assessors or the same assessor. Hence, there is a need for alternative methods to generate relevance judgments with reduced human assessors' effort.

2.1 Human assessor errors in generating relevance judgments

Human expertise is a reliable source of judgment when weighing attributes in web evaluation (Saeid *et al.*, 2011) similar to generating the relevance judgments in TREC. However, studies show the possibilities of errors introduced, level of errors, and threshold of acceptable errors when humans generate relevance judgments. Human errors have been noted during the Cranfield methodology where the relevant documents could not be identified due to human errors in indexing, searching, or in the process of preparing the catalogs. Proper indexing is crucial (Varathan *et al.*, 2014), while a study shows the retrieval system used did not appear to show a significant effect on the system performance where only one in 20 retrieval errors could be associated with the retrieval system (Cleverdon, 1991). Errors in retrieving the relevant documents due to same word with different meanings may cause a decrease in recall, and different words with the same or similar meanings may result in a retrieval of wrong or irrelevant documents, causing a decrease in precision (Carpineto and Romano, 2012).

In analyzing the human assessment error, experts are not able to assign exact weights to attributes in web evaluation (Saeid *et al.*, 2011). Similarly, assessment error is at a high level, and inconsistency exists between the various topics used during the laboratory-based evaluation (Scholer *et al.*, 2011). While judging consumes a lot of time, the authors (Scholer *et al.*, 2011) have associated the distance between two documents' matches with the amount of time between the judgments made. Inconsistency increases as the distance between the duplicate pair increases as well (Scholer *et al.*, 2011). Based on investigation, judging relevant documents needs more time compared to judging irrelevant documents (Carterette and Soboroff, 2010). As time increases while performing the assessment, the possibilities of introducing errors increase (Smucker and Jethani, 2012). In another experiment, results show that it takes a longer time for making error judgments when compared to making correct judgments (Smucker and Jethani, 2012). In either scenario, the judges are prone to induce some level of error judgments.

In an analysis where at least one of the documents judged as relevant, the fraction of inconsistently judged duplicates that were rather similar range from 15 to 24 percent (Scholer *et al.*, 2011). Multiple assessors were used to analyze the impact of the errors that could be introduced (Scholer *et al.*, 2011). Engaging different groups of assessors show a low level of agreement in judging the relevancy (Bailey *et al.*, 2008). The same documents judged by different assessors tend to cause disagreements on the relevancy where a low-ranked document judged relevant and a high-ranked document judged irrelevant cause disagreement from the other assessors (Webber *et al.*, 2012). In other studies, 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 (Carterette and Soboroff, 2010; Rasmussen, 2002).

Despite human disagreements and possible errors induced in generating relevance judgments, the experiment proves that varied relevance judgments used for evaluating same runs show high levels of correlation coefficient (Voorhees, 2000). This indicates, although different human assessor could produce different relevance judgments, that comparing the evaluation of retrieval performance is stable (Voorhees, 2000). Besides, web user's satisfaction on the retrieved ranked documents is an important aspect in addition to relevance judgments (Huffman and Hochster, 2007).

2.2 Alternative methods in generating relevance judgments

Due to possible errors from human assessors, studies to find alternate methods to generate relevance judgments without the involvement of human assessors (Nuray and Can, 2003; Soboroff *et al.*, 2001; Rajagopal *et al.*, 2014) or with minimal involvement of human assessors (Scholer *et al.*, 2011) have been conducted. For instance, a method known as exact fraction sampling of relevant document occurrences in each topic was used to populate the pseudo relevance judgments (Soboroff *et al.*, 2001). The exact fraction method draws exact numbers of relevant documents per topic based on the percentage of relevant documents calculated from the original relevance judgments. Each topic consists of different numbers of relevant documents. The experimented exact fraction method without human assessors resulted in an average correlation coefficient of between 0.385 and 0.463 for the different TREC test collections (Soboroff *et al.*, 2001). Although this method uses exact percentages per topic, the selection of relevant documents from the pool were random.

In another study, the TREC original relevance judgment was altered to suit the web resemblance scenario using a heuristics method to replicate the imperfect web environment (Nuray and Can, 2003). The experimentation used four test collections, and three of the test collections were assumed as inaccessible or not available. The original relevance judgment was modified to indicate those documents from the inaccessible test collections as not relevant. The pooling and ranking of documents were done based on the similarity scores using the vector space model. Their experiment resulted in Kendall's τ correlation of automatic method and human assessed method for average precision (AP) and precision at document cutoff value appearing to be better for a pool depth of 30 compared to a pool depth of 200 (Nuray and Can, 2003). The AP correlation for the pool depth of 30 between the automatic method and the human judged relevance judgment ranges between 0.384 and 0.405 (Nuray and Can, 2003).

Random selection of documents was performed based on the average number of relevant documents from each topic in the pool. The total percentage of relevant documents appearing in each topic is used to select and judge relevant documents for the pseudo relevance judgment. Their correlation coefficient ranges between 0.369 and 0.487 for all test collections (Soboroff *et al.*, 2001). There could be a loss of accuracy in selecting relevant documents due to the averaging of relevance documents occurrence in each pool. Alternatively, another method (Nuray and Can, 2003) had randomly selected top ten documents from some systems to form the pool and repeated the selection ten times before computing the AP correlation. The resulted correlation of 0.401 was not as strong as that proposed by Soboroff *et al.* (2001).

In summary, the involvement of human assessors during the generation of relevance judgment induces errors and inconsistent judgment for the same documents by different assessors. Previous alternate methods have attempted to eliminate human assessors or involve minimal human assessment but did not obtain strong correlation coefficient.

3. Research design

In the traditional TREC evaluation cycle, after pooling using the top X (usually 100) documents that are deemed to be most relevant from the submitted runs of participating systems, the pooled document is presented to human assessors for judgment to create the relevance judgment. Instead, this study proposes the creation of relevance judgments without human assessors. The system scores are then calculated using the chosen metrics to rank the systems.

The system ranks obtained in this study use mean average precision (MAP) metrics. Then, correlation coefficient between system ranks using original relevance judgments and pseudo relevance judgments are computed. Equations (1) and (2) show the AP and MAP equations, respectively.

If there are R relevant documents for a query and if the evaluation is being carried out to some evaluation depth k, and if $r_i = 1$ when the ith document in the ranking is relevant and $r_i = 0$ otherwise, then the AP for that query is computed as follows:

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

MAP for a set of queries, Q is the mean of the AP scores for each query, q and can be defined as:

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

A previous study (Soboroff *et al.*, 2001) uses random selection of documents from the pool to generate pseudo relevance judgments, whereas in our study, the selection of relevant documents uses calculated scores in a systematic way to generate relevance judgments. Sections 3.1 and 3.2 discuss the proposed methods, exponential variation, and document rankings in detail.

3.1 Exponential variation

The exponential variation method assumes that the possibilities of documents being relevant in a group with more occurrences are higher when compared to documents being relevant in groups with lower retrieval. The documents are grouped based on the number of occurrences but the number of relevant documents is determined using exponential 2^x, where x ranges from zero to nine. Exponent mapping decreases document relevancy exponentially down the ranked list. Possibilities of relevant documents exist at low ranks.

Pooled documents consist of retrieved top X documents from participating systems. These pooled documents are ordered in a descending manner based on the calculated percentage value (CV) computed using the following equation:

$$\text{calculated \% value}, CV = \frac{\text{number of occurrences}}{\text{total systems}} \times 100\% \tag{3}$$

The grouping of documents is based on the CV ranging from 100 to 0 percent with intervals of 10 percent for each group. Each of these group map to a particular exponent 2^x. These exponents reflect the number of documents per set per group. The steps taken to generate the pseudo relevance judgments through the exponential variation method is shown as follows:

(1) use runs from all systems after data cleaning phase;

(2) pool with depth of k (100 or 200);

(3) order the documents by topic, then by document Id;

(4) count the number of occurrences of each document per topic;

(5) order the documents in a descending manner based on number of occurrences;

(6) remove duplicate documents;

(7) calculate the percent value using Equation (3);

(8) divide the documents into groups with intervals of 10 percent based on the calculated percent value (*CV*);

(9) divide each group into sets based on the mapped exponent value;

(10) mark the first document as relevant and the remaining as irrelevant per set; and

(11) combine all judged relevant and irrelevant documents from each set to create a pseudo relevance judgment.

Group 1 consists of documents with a *CV* between 100 and $\geqslant 90$ percent. Divide Group 1 into sets of one document since the exponent matched for Group 1 is $2^0 = 1$. Judging the first document of each set as relevant means all documents in Group 1 are relevant since each set only contains one document.

Group 2 consists of documents with a *CV* between < 90 and $\geqslant 80$ percent, where each set contains two documents since it is mapped with exponent $2^1 = 2$. For each set in Group 2, judge the first document out of the two documents as relevant. Similarly, Group 3 consists of documents with *CV* between < 80 and $\geqslant 70$ percent.

Group 3 has four documents in each set based on exponent $2^2 = 4$. Judge the first document out of the four documents in each set in Group 3 as relevant.

Moving on to Group 4 with exponent $2^3 = 8$, the documents within this group have a *CV* between < 70 and $\geqslant 60$ percent. Each set has eight documents. Judge the first document on each set as relevant. As the exponent increases, the number of documents within each set of a particular group increases as well. However, the percentage of the documents judged as relevant reduces, and this is feasible because as we go down the document list, it is expected that the documents would have a slimmer chance of being relevant.

Finally, Group 10 consists of documents with a *CV* of < 10 percent, and each set within Group 10 contains 512 documents, mapped with exponent $2^9 = 512$. Similarly, judge the first document in each set in this group as relevant while the remaining 511 documents as irrelevant.

The proposed exponential variation method could allow possibilities of the document being relevant with a low-*CV* rather than focussing solely on the higher number of document occurrences. Judging the first document from each set is one systematic way in creating pseudo relevance judgment. On the other hand, other approaches, such as selecting designated numbers of documents according to the mapped exponent by random selection from each set or group, could be possible but not explored in this study to sustain the possibilities of creating similar pseudo relevance judgment.

3.2 Document rankings

The document ranking method uses two variables from the system run files, namely, the document ranks and number of occurrences of each document for each topic. The document rank provides a valuable indication about the relevancy of a document and serves as an important parameter in determining document relevancy. A document may be retrieved by many systems and ranked differently by each system. Hence, a single value is needed to indicate the average rank (sum of all document ranks divided

by number of occurrences) of the document from all the systems that retrieved it. A document retrieved by many systems but ranked lower down the list might not be better than a document ranked higher by a few systems. This method motivates to experiment judging relevant documents in the mentioned context. The calculated value (*CR*) for each of the documents is then obtained using the following equation:

$$\text{calculated value}, CR = \frac{(\text{number of occurrences})^2}{\text{sum of document ranks from all systems}} \quad (4)$$

The judging of document relevancy is based on a selection of specific percentage value from the overall top 100 pooled document. A heuristic method is used to choose the percentages as only a fraction of the documents from the pool are relevant. The three different percentages selections are 5, 10, and 20 percent. Judge the fraction of documents from the pooled documents as relevant using these percentages. Order the documents in a descending manner based on the calculated value (*CR*), and judge the relevancies starting from the document with the highest calculated value (*CR*).

For example, TREC-8 has 108,819 documents in the pool when the pooled depth is 100. Using the 5 percent selection, the top 5,441 documents is judged as relevant, the top 10,882 documents using the 10 percent selection, and top 21,764 documents will be judged as relevant for the 20 percent selection. The following are the steps taken to generate the pseudo relevance judgments:

(1) use runs from all systems after data cleaning phase;

(2) pool with depth of k ($k = 100$);

(3) calculate the value for each document in each topic (Equation (4));

(4) order the documents descending based on the calculated value (*CR*);

(5) mark the top documents' relevancy using specific percentage value (5, 10, or 20 percent); and

(6) combine all judged relevant and irrelevant documents from each percentage to create separate pseudo relevance judgments.

3.3 Correlation coefficient

For this study, Kendall's τ and Pearson correlation measures the correlation coefficient of the system rankings between the list of ranked systems using the MAP metric (refer to Equation (2)) computed using the original TREC relevance judgments and pseudo relevance judgments. The correlation coefficient measures the effectiveness of the proposed methods in generating a reliable set of relevance judgments. The correlation coefficient values closer to 1.0 is desirable, and in most of the IR evaluation related research, a correlation coefficient value of 0.9 and above represents a strong measure of linear relationship between two variables (Vorhees, 2005; Yilmaz and Aslam, 2006). Obtaining a correlation coefficient of 0.5 and above could also be sufficient to contribute to the research area as a similar study conducted previously (Soboroff *et al.*, 2001), which had obtained a correlation of approximately 0.5.

4. Results

The analysis of the results obtained through the experimentation conducted in this study using each proposed method in Sections 3.1 and 3.2 are in separate sections

starting with exponential variation method for a pool depth of 100 and 200; and document ranking method. Each results section consists of subsections as below:

(1) The correlation coefficient between the list of ranked systems using the original relevance judgments and pseudo relevance judgments to determine if the proposed method in generating relevance judgment without human assessor could generate a sufficiently close relevance judgment with that generated through human assessors (Sections 4.1.1 and 4.2.1).

(2) The ranked systems generated using the original TREC relevance judgments is divided into three groups of systems with similar performance level; good, moderate, and low performing systems. The systems are sorted in a descending manner using the MAP scores generated using the original TREC relevance judgments and divided into approximately equal number of systems in each group. Good performing systems are those that have the best MAP scores, low performing systems are those with poor MAP scores, and moderately performing systems are those that fall in between the good and low performing systems. The analysis of the correlation coefficient of these three different groups of similar system performances generated using the original TREC relevance judgments and pseudo relevance judgments is to identify if a particular group of systems perform well using the pseudo relevance judgment created through the proposed methods (Sections 4.1.2 and 4.2.2).

The experiments use two different test collections: ad-hoc track from TREC-8 and web track from TREC-9. The three groups of similar system performance are based on their original TREC system scores for each test collection (see Table I).

4.1 Exponential variation method

4.1.1 Overall correlation coefficient. Table II shows the correlation coefficient between the lists of ranked systems using system scores generated through the original TREC relevance judgments and pseudo relevance judgments for the exponential variation method for a pool depth of 100 and 200 (see Table II).

Using pool depth 100, the exponential variation method produced a moderate Kendall's τ correlation of 0.470 for TREC-8 and a strong correlation of 0.556 for TREC-9. A strong Pearson correlation, above 0.7 for TREC-8 and TREC-9, could be due to the additional documents from the non-contributing systems that were not initially in the pool for original TREC relevance judgments.

The experiment with a pool depth of 200 for the exponential variation method is conducted because it is important to know if a deeper pool depth could produce a varying outcome. There were more documents in the pooling for depth 200 but the

	TREC-8 ad-hoc track	TREC-9 web track
Total systems used	129	104
Good performing systems	43	35
Moderately performing systems	43	34
Low performing systems	43	35

Note: Groupings of similar system performance for TREC-8 and TREC-9 indicating number of systems for each grouping based on their original sorted TREC MAP scores

Table I.
Number of systems grouped based on system performance

correlation coefficient was ranging close to those pooled at depth 100. This result shows that a deeper pool depth does not necessarily produce a better correlation coefficient.

The Kendall's τ and Pearson correlation using the exponential variation method for a pool depth of 100 and 200 is almost similar for both test collections and shows improvement with an increased depth but does not give a meaningful impact to the system rankings.

4.1.2 Correlation coefficient of ranked systems with similar performance level. The Kendall's τ and Pearson correlations between the lists of ranked systems using system scores generated through the original TREC relevance judgments and pseudo relevance judgments for the exponential variation method were then computed for each group of systems with similar performances (see Table III). The systems ranked in descending order using the MAP scores were generated from the original TREC relevance judgments and divided into approximately equal number of systems in each groups or subsections.

The low performing systems have strong correlations when measured using Kendall's τ and Pearson correlation coefficient, where the Kendall's τ generated a value of 0.8 while Pearson correlation is 0.9 and above for TREC-8. Although there were additional documents from the non-contributing systems included in the pooling for pseudo relevance judgments, these documents did not influence the performance evaluation of the low performing systems. Similarly, the low performing systems for TREC-9 also produced a strong Kendall's τ value of 0.8 and Pearson correlation of 0.9 and above. Clearly, the low performing systems have performed consistently across both test collections.

Table II.
Kendall's τ and Pearson correlation values using different depth of pooling for TREC-8 and TREC-9 (metric used: MAP)

Methods	Exponential variation	
Depth k	Depth 100	Depth 200
TREC-8		
Kendall's τ	0.470	0.517
Pearson	0.735	0.751
TREC-9		
Kendall's τ	0.556	0.562
Pearson	0.789	0.790

Table III.
Correlation coefficient for three subsections of systems with similar performance level for TREC-8 and TREC-9 using exponential variation method with different depth k (metric used: MAP)

	Depth k	Good performing systems	Moderately performing systems	Low performing systems
TREC-8				
Kendall's τ	100	−0.329	0.264	*0.804*
	200	−0.303	0.359	*0.823*
Pearson	100	−0.860	0.415	*0.952*
	200	−0.862	0.544	*0.963*
TREC-9				
Kendall's τ	100	−0.176	0.057	*0.762*
	200	−0.119	0.062	*0.901*
Pearson	100	−0.411	0.041	*0.944*
	200	−0.412	0.080	*0.946*

Note: Italic values are the best values

On the other hand, the moderately performing systems for TREC-8 and TREC-9 did not produce a strong correlation. The correlation coefficient between lists of system ranks using original TREC relevance judgments and pseudo relevance judgments using exponential variation method produced a weak correlation (see Table III).

The good performing systems have negative correlations because the exponential variation method has now caused the scores of these systems to be lower when compared to the scores obtained using original TREC relevance judgments. The ranks of these systems have now decreased when compared to the original system ranks that causes the negative correlation coefficient.

The Kendall's τ and Pearson correlation for the list of ranked systems using original TREC relevance judgments and pseudo relevance judgment for a pool depth of 200 are close to that of the pool depth of 100. All the low performing systems using a pool depth of 200 have better correlation coefficient when compared to that of the pool depth of 100 for both test collections (see Table III). Although the improvement in the correlation is small with an increased pool depth to 200 using the exponential variation method, a better correlation is desirable.

Meanwhile, the moderately performing systems using the exponential variation method also has improvements with the correlation coefficient with an increased pool depth to 200. Although the improvement is not drastic, the benefit to better correlations is definitely notable. Referring to Table III, the good performing systems for TREC-8 have strong but negative Pearson correlation caused by the lower system scores obtained using pseudo relevance judgments through the exponential variation method. While the system scores from the original TREC relevance judgments were increasing, the system scores from pseudo relevance judgments were decreasing, which could be the case where fewer relevant documents appeared in these systems when using pseudo relevance judgments compared to original TREC relevance judgments.

4.2 Document ranking method

4.2.1 Overall correlation coefficient. In document ranking method, the usage of various percentages to judge document relevancy produced a correlation coefficient between lists of ranked systems using system scores generated through the original TREC relevance judgments and pseudo relevance judgments as shown in Table IV.

For TREC-8, the 10 percent selection of documents from the pooled documents after computing the calculated value (CR) using Equation (4) has shown improvement in Kendall's τ correlation when compared to the 5 percent selection of the total pooled documents. When 20 percent documents selected to judge as relevant of the total documents from the pool using TREC-8, the correlation coefficients between the lists of ranked systems using system scores generated through original TREC relevance judgments and pseudo relevance judgments starts to decrease. Similarly, TREC-9 shows improvement in Kendall's τ and Pearson correlation for the 10 percent selection of pooled documents from all participating systems when compared to the 5 percent selection of documents.

Based on the correlation coefficient, the 10 percent selection from the total pooled documents has produced better Kendall's τ and Pearson correlation compared to the 5 and 20 percent selections of pooled documents, whereby three out of four correlation coefficients values show improvement. This makes the 10 percent selection a better option for judging relevancy in generating the pseudo relevance judgments using the

document ranking method. Overall, both metrics show that TREC-9 is better at all levels of correlation coefficient.

4.2.2 Correlation coefficient of ranked systems with similar performance level. Table V indicates the Kendall's τ correlation for the three groups or subsections of the systems based on their performance rankings from the original TREC relevance judgments (see Table V).

The Kendall's τ correlation for low performing systems produced strong correlations while the good performing systems did not produce good correlation coefficients for TREC-8. It simply translates that pairs of systems for good performing systems are more discordant when compared to low performing systems having higher concordant pairs of systems. The good and moderately performing systems have high numbers of relevant documents based on the original TREC relevance judgments, but now, the additional documents from the non-contributing systems have caused lesser relevant documents for these groups of systems using the document ranking method, influencing their performance rankings.

When verified with the computed Kendall's τ correlation for TREC-9, low performing systems also show strong correlations of approximately 0.8. Meanwhile, the good performing systems did not produce a good Kendall's τ correlation where the correlation between the system ranks using the original TREC relevance judgment and pseudo relevance judgment is almost independent.

Table VI shows the Pearson correlation for the three groups or subsections of the systems based on their performance rankings from the original relevance judgments (see Table VI).

Table IV.
Correlation between TREC and pseudo relevance judgments for document rankings method

	Kendall's τ		Pearson	
% selection	TREC-8	TREC-9	TREC-8	TREC-9
5	0.536	0.565	*0.743*	0.751
10	*0.548*	*0.661*	0.734	*0.826*
20	0.512	0.633	0.706	0.807

Notes: Italic values are the best values. Kendall's τ and Pearson correlation between original TREC relevance judgment and pseudo relevance judgment for documents rankings method using pool depth of 100 (metric used: MAP)

Table V.
Kendall's τ correlation for document rankings method for three subsections of systems with similar performance level using depth 100 (metric used: MAP)

	Kendall's τ					
	TREC-8 (total systems = 129)			TREC-9 (total systems = 104)		
% selection	Good performing systems (43 systems)	Moderately performing systems (43 systems)	Low performing systems (43 systems)	Good performing systems (35 systems)	Moderately performing systems (34 systems)	Low performing systems (35 systems)
5	−0.257	0.440	*0.798*	0.042	0.241	*0.876*
10	−0.223	0.434	*0.799*	0.165	0.271	*0.862*
20	−0.215	0.312	*0.757*	0.229	0.277	*0.785*

Note: Italic values are the best values

The low performing systems, again, have strong Pearson correlations of 0.9 and above for TREC-8. The scores of the low performing systems have been increasing as those from the original system scores were increasing as well, hence, producing a very strong Pearson correlation. The similar correlation coefficient by the low performing systems for TREC-9 echoes TREC-8 in addition to having a better Pearson correlation coefficient.

Meanwhile, the moderately performing systems for TREC-8 has a Pearson correlation that is strong but only a moderate correlation for TREC-9, which is approximately below 0.4. Whereas, the good performing systems for TREC-8 have the system ranking scores decreasing when computed with the pseudo relevance judgments, while those systems were originally performing well with the original TREC relevance judgments. The good performing system scores have dipped more for TREC-8 compared to TREC-9.

It can be clearly noted that the low performing systems have strong correlation coefficients across both test collections, and the 5 percent selection of documents have the best Pearson correlation compared to 10 and 20 percent.

5. Discussion

Experiments with different pool depths have increased the numbers of relevant documents in the pseudo relevance judgments, but correlation coefficients between lists of system ranks using the original TREC relevance judgments and pseudo relevance judgments did not increase largely. As mentioned in the previous study, the results show reliable output when a sufficient pool depth of 100 is used (Zobel, 1998). Similarly, it can be reiterated that sufficiently good results are obtained through pool depth 100, although an increasing pool depth does provide improvement in the system effectiveness scores. Without pooling, there would be too many documents to judge by human assessors, which may introduce errors in the judgments. The proposed methods overcome disagreement errors introduced by human assessors while systematically generating relevance judgment. The concern of too many documents for judgment by human assessors is therefore minimized.

The proposed exponential variation method focusses on judging the relevancy of documents using exponent mapping. With the assumption that document relevancy decreases exponentially down the ranked list, this method attempts to overcome the elimination of low ranked relevant documents. It also benefits to satisfy the uncommon user needs who may find low ranked documents as relevant. In other words, this method satisfies users who find relevancy in documents that are the

Table VI. Pearson correlation for document rankings method for three subsections of systems with similar performance level using depth 100 (metric used: MAP)

	Pearson					
	TREC-8 (total systems = 129)			TREC-9 (total systems = 104)		
% selection	Good performing systems (43 systems)	Moderately performing systems (43 systems)	Low performing systems (43 systems)	Good performing systems (35 systems)	Moderately performing systems (34 systems)	Low performing systems (35 systems)
5	−0.850	0.666	*0.951*	−0.366	0.321	*0.968*
10	−0.847	0.652	*0.949*	−0.240	0.386	*0.967*
20	−0.838	0.534	*0.938*	−0.218	0.412	*0.962*

Note: Italic values are the best values

least of interest to majority of users who find relevant documents at the top ranks. The experiment conducted using the exponential variation method only uses a single variation for relevant document selection although other ways could be experimented.

On the other hand, the document ranking method takes into consideration the document ranks from various systems instead of only the number of occurrences of each document. Utilizing document ranks have produced better correlation coefficients between the lists of system ranks using original TREC relevance judgments and pseudo relevance judgments compared to the exponential variation method that did not use document ranks. This indicates that document ranks provide a useful contribution in generating pseudo relevance judgments. The implementation of the document ranking method in evaluating real web retrieval systems would require ranks from multiple retrieval systems to produce new relevant document ranks.

6. Conclusions and future work

Two main methods have been experimented in creating the relevance judgments to reduce the human efforts involved. Based on the Kendall's τ correlation, the document ranking method has higher correlations compared to the exponential variation method. In the subdivision of systems with similar performances, the low performing systems correlate positively with the original systems ranks. Pooling with non-contributing and contributing systems from TREC has a minimal impact on the system rankings of low performing systems since the methods proposed have judged the relevancy of documents in a reliable manner where the proposed methods does not re-rank the low performing systems to high ranks. However, the proposed methods with documents from contributing and non-contributing systems have affected the system rankings of the good and moderately performing systems that could be due to the additional documents from the non-contributing systems.

Experimenting with an increased pool depth of 200 did not generate sufficient improvement in the correlation coefficient but showed a slight improvement in the system rankings. The low performing systems for the pool depth 200 continues to have a strong correlation coefficient despite pooling with non-contributing and contributing systems.

The proposed document ranking method could be accepted as a reliable alternative to traditional pooling as the correlation coefficient obtained is above 0.5, better than previous study (Soboroff *et al.*, 2001). Pooling documents from contributing and non-contributing systems and generating relevance judgments without human assessors are the advantages of the proposed methods. Though alternate methods to generate human assessed relevance judgments show promising results, proceeding without humans may require sufficient studies to incorporate human behaviors in relevance assessment. Humans provide user satisfaction input, context, and error handling in relevance assessments.

There is a need for further experiments around the exponential variation method with variations in selecting relevant documents. Variation in experiments allows the identification of similar or worse correlations. It would also be interesting to experiment generating relevance judgments by including all documents from all participating systems.

Finally, as the test collections continue to evolve, methods to improve and upgrade on the creation of relevance judgments while maintaining consistency in evaluating system performance could be the possible application of the proposed methods.

References

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. and Yilmaz, E. (2008), "Relevance assessment: are judges exchangeable and does it matter?", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Singapore*, pp. 667-674.

Carpineto, C. and Romano, G. (2012), "A survey of automatic query expansion in information retrieval", *ACM Computing Surveys*, Vol. 44 No. 1, pp. 1-50.

Carterette, B. and Soboroff, I. (2010), "The effect of assessor errors on IR system evaluation", *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Geneva*, pp. 539-546.

Cleverdon, C.W. (1991), "The significance of the cranfield tests on index languages", *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Chicago, IL*, pp. 3-12.

Greengrass, E. (2000), *Information Retrieval: A Survey*, University of Maryland, Baltimore County, MD.

Hersh, W., Elliot, D., Hickam, D., Wolf, S. and Molnar, A. (1995), "Towards new measures of information retrieval evaluation", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Seattle, WA*, pp. 164-170.

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 – SIGIR, ACM, Amsterdam*, pp. 567-574.

Mandl, T. (2008), "Recent developments in the evaluation of information retrieval systems: moving towards diversity and practical relevance", *Informatica*, Vol. 32 No. 1, pp. 27-38.

Nuray, R. and Can, F. (2003), "Automatic ranking of retrieval systems in imperfect environments", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Toronto*, pp. 379-380.

Rajagopal, P., Ravana, S.D. and Ismail, M.A. (2014), "Relevance judgments exclusive of human assessors in large scale information retrieval evaluation experimentation", *Malaysian Journal of Computer Science*, Vol. 27 No. 2, pp. 80-94.

Rasmussen, E. (2002), "Evaluation in information retrieval", *The MIR/MDL Evaluation Project White Paper Collection*, 3rd ed., ACM, Toronto, pp. 45-49.

Ravana, S.D. (2011), "Experimental evaluation of information retrieval systems", unpublished manuscript, Computing and Information Systems – Theses, The University of Melbourne, Melbourne.

Saeid, M., Abd Ghani, A.A. and Selamat, H. (2011), "Rank-order weighting of web attributes for website evaluation", *The International Arab Journal of Information Technology*, Vol. 8 No. 1, pp. 30-38.

Scholer, F., Turpin, A. and Sanderson, M. (2011), "Quantifying test collection quality based on the consistency of relevance judgements", *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information – SIGIR, ACM, Beijing*, pp. 1063-1072.

Smucker, M. and Jethani, C. (2012), "Time to judge relevance as an indicator of assessor error", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Portland, OR*, pp. 1153-1154.

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 – SIGIR, ACM, New Orleans, LA*, pp. 66-73.

Turpin, A., Scholer, F., Jarvelin, K., Wu, M. and Culpepper, J. (2009), "Including summaries in system evaluation", *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, ACM, Boston, MA*, pp. 508-515.

Varathan, K.D., Sembok, T.M.T., Kadir, R.A. and Omar, N. (2014), "Semantic indexing for question answering system", *Malaysian Journal of Computer Science*, Vol. 27 No. 4, pp. 261-274.

Voorhees, E.M. (2000), "Variations in relevance judgments and the measurement of retrieval effectiveness", *Information processing & management*, Vol. 36 No. 5, pp. 697-716.

Vorhees, E.M. (2005), "Overview of TREC 2005", *NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC), November 15-18, Gaithersburg, MD.*

Webber, W., Chandar, P. and Carterette, B. (2012), "Alternative assessor disagreement and retrieval depth", *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, Maui, HI*, pp. 125-134.

Yilmaz, E. and Aslam, J. (2006), "Estimating average precision with incomplete and imperfect judgments", *Proceedings of the 15th ACM International Conference on Information and Knowledge Management – CIKM, ACM, Arlington, VA*, pp. 102-111.

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 – SIGIR, ACM, Melbourne*, pp. 307-314.

About the authors

Dr Sri Devi Ravana is a Senior Lecturer in Computer Science at the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. She holds PhD Degree in Computer Science from The University of Melbourne. Her main research area is in the field of Information Retrieval and Data Engineering. Dr Sri Devi Ravana is the corresponding author and can be contacted at: sdevi@um.edu.my

Prabha Rajagopal is a PhD Student in the University of Malaya, Malaysia and attached to the Department of Information Systems. Her main research area is in the field of Information Retrieval.

Dr Vimala Balakrishnan is a Senior Lecturer attached to the Department of Information Systems in the University of Malaya, Malaysia. Her research interests include data and knowledge engineering (social and health informatics), social networks media, and information security.