Information Research - Vol. 29 No. 3 (2024)

Improving the accuracy of the information retrieval evaluation process by considering unjudged document lists from the relevant judgment sets

Minnu Helen Joseph and Sri Devi Ravana
DOI: https://doi.org/10.47989/ir293603

Abstract

Introduction. To improve user satisfaction and loyalty to the search engines, the performance of the retrieval systems has to be better in terms of the number of relevant documents retrieved. This can be evaluated through the information retrieval evaluation process. This study identifies two methodologies that help to recover and better rank relevant information resources based on a query, while at the same time suppressing the irrelevant one.

Method. A combination of techniques was used. Documents that were relevant and not retrieved by the systems were found from the document corpus and assigned new scores based on the Manifold fusion techniques then moved into the relevant judgment sets. Documents based on judgment sets and good contributing systems have been considered in the proposed methodologies.

Analysis. Kendall Tau Correlation Coefficient, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Rank Biased Precision (Rbp) have been used to evaluate the performance of the methodologies.

Results. The proposed methodologies outperformed the baseline works and enhanced the quality of the judgment sets, achieving a better result even with lesser pool depth.

Conclusion. This research proposes two methodologies that increase the quality of the relevant documents in the judgment sets based on document similarity techniques and, thus, raise the evaluation process accuracy and reliability of the systems.

Introduction

Finding the most relevant documents from the massive data on the World Wide Web is always a challenge (Guiver et al., 2009). Whenever a user tries to retrieve information from the Web, the participating systems retrieve some relevant information based on the query given by the user. The number of relevant documents retrieved depends on the performance of the system as well as on the user's query.

Most of the research methodologies based on information retrieval systems evaluation are based on the Cranfield paradigm which utilizes a large set of test collections to evaluate the quality of different retrieval methods and techniques. A test collection in the Cranfield paradigm consists of a *document corpus* which consists of a set of documents, *topics*, user information needs, and a *relevant judgment set*, which shows the relevancy of a document over a topic. The judgment set mentioned here is a binary representation of all the documents related to all topics (Voorhees, 2002). According to Cranfield's assumption, all the relevant documents have been generated in the judgment lists, meaning that all the documents relevant to all the topics have been collected and moved to the judgment list correctly. This assumption is accurate for smaller datasets, while for larger datasets like TREC and CLEF, it may not be entirely correct, but it comes close to the Cranfield assumption (Buckley & Voorhees, 2004).

Retrieval evaluation is a process of measuring how well the participating systems meet the information required or needed by the user (Voorhees, 2002). This evaluation is done for two purposes. Firstly, to determine the performance of the systems. The performance of the retrieval systems is determined not only by its efficiency but also by its effectiveness, which is the ability to retrieve as many relevant documents as possible, rank them according to their relevancy, and at the same time suppress the irrelevant ones (Ferro, 2017). Secondly, this evaluation is done to determine why the quality of relevant judgments is important. The quality of relevant judgments increases with the number of relevant documents. If we fail to collect enough relevant documents in the judgment set, the quality of the judgment set also decreases. By increasing the accuracy of the evaluation process, we can indirectly help users rely on search engines (Rahman et al., 2020). However, evaluation based on relevance judgment sets is quite a challenge (Culpepper et al., 2014). Most of the time, the participating systems won't be able to retrieve sufficient relevant documents into the judgment list due to poor system performance., which consequently affects the effectiveness of the evaluation process.

There are two approaches to evaluating information retrieval systems: system-based evaluation and user-based evaluation. User-based evaluation measures how satisfied the users are with the systems, while system-based evaluation measures how effectively the systems retrieve relevant documents and rank them based on relevance (Voorhees, 2002). Because the main goal of information retrieval evaluation is to determine users' satisfaction with the retrieval documents, user-based evaluation is preferred. However, user-based evaluation requires a large sample of actual users, and each system being compared must be well-developed with the same user interface and compilation speeds (Mandl, 2008). Additionally, user-based evaluation is subjective and depends on the user's perspective, requirements, and judgments, which can change over time (Zuva *et al.*, 2012). Each experiment requires significant human participation, making it a costly and time-consuming process.

System-based evaluation relies entirely on a test collection that has been developed with limited resources of expert judges (Maddalena et al., 2017). A test collection comprises a document corpus, topics, and a set of relevant judgments (Mandl, 2008; Melucci & Baeza-Yates, 2011; Voorhees, 2002). Although it is expensive to produce a test collection, the advantage is that it can be reused for each experiment. Experiments can be repeated multiple times with the same test collection, which cannot be done with user-based evaluation. Experiments based on a test collection consider topics as the primary experimental unit, and systems collect

documents from the document corpus based on each topic. The evaluation of the retrieved documents is based on the relevance judgment set available in the test collection. This set displays the relevance of each topic to each document.

Various evaluation metrics can be used to measure the performance of the participating systems, either by considering the number of relevant documents or by the quality of the ranking of these retrieved documents. The metrics such as precision, P@k (Hembrooke et al., 2005), average precision, AP (Buckley et al., 2017), normalized discounted cumulated gain, NDCG (Järvelin and Kekäläinen, 2002), and rank biased precision (Rbp; Moffat & Zobel, 2008) can all be used to evaluate the performance of systems. Various methodologies such as pooling (Sparck-Jones & van Rijsherhen, 1975), human accessors contributions (Alonso and Mizzaro, 2012), considering topics (Roitero et al., 2020), and document similarities (Djenouri et al., 2018) are proposed by the researchers to increase the quality of the judgment sets by increasing the number of relevant documents.

This study's objectives are:

- To increase the accuracy of information retrieval evaluation by increasing the number of relevant documents in the relevance judgment list, by proposing a methodology considering pooling and document similarity techniques.
- To improve the effectiveness of the information retrieval participating systems by proposing an enhanced methodology by considering documents from the good contributing systems, which are considered based on the evaluation metric scores.

Literature about various existing methodologies and issues related to the quality of the judgment sets is presented. Then, the methodologies proposed to increase the quality of the judgment sets are presented. Next, the performance of the methodologies based on the retrieval systems is presented and the results are discussed. Finally, conclusions are drawn.

Related works

Large numbers of relevant documents in a judgment set always help to increase the quality of the judgment set and through that, can increase the accuracy of the evaluation process. Finding the most relevant documents in the judgment sets is always a challenge. Fewer relevant documents in the judgment sets always affects the effectiveness of the evaluation process. Retrieving insufficient relevant documents is due to factors such as topic difficulty, human accessor errors, and biases in the ranking of documents, all of which affect the quality of the judgment sets. Many studies have been done by researchers to increase the quality of the judgment sets.

The importance of methodologies in addition to relevance

Researchers have used many methodologies to increase the effectiveness of the evaluation process. Finding the greatest number of relevant documents from the large data collection is time-consuming and costly. Previous research has shown that finding the subset of documents and evaluating these documents will have almost the same effect as evaluating the whole document corpus (Sparck-Jones & van Rijshergen, 1975).

Pooling methodologies

Identifying relevant documents from the retrieved list of the participating systems has been done by humans who were experts in these areas. The TREC test collection, initiated by the NIST organizers, provides a large collection of documents for the evaluation of systems at scale. The size of each collection ranges from millions to billions of documents. However, evaluating such large collections through expert judges can take decades to complete and can be expensive (Moghadasi et al., 2013). To overcome this issue, crowdsourcing was considered as an alternative. The idea was to collect relevant documents from real users on the crowdsourcing platform (Tonon et al., 2015). However, this method also had some limitations, such as being more prone to errors.

Pooling is a technique proposed (Sparck Jones & van Rijshergen, 1975) to address information

retrieval challenges. It involves considering only a subset of documents from a merged ranked list. Specifically, it takes only the top-k documents from each system run. The technique assumes that all the documents in the pooled list are relevant, and any document not in the pool is irrelevant. The pool depth chosen and the retrieval methods used for the evaluation can affect the quality of the relevance judgment set (Buckley et al., 2007).

One of the most widely used pooling methods for information retrieval evaluation is the depth@k method, which involves selecting the top k relevant documents from each topic in the runs generated by participating systems. To avoid duplicates, all identical documents are removed from the list before being presented to human assessors for evaluation. This method helps to reduce the size of the judgment list (Sparck Jones & Rijshergen, 1975), as only a partial relevant judgment set is considered instead of the entire list. This partial relevant judgment set is then used for evaluation purposes.

The traditional method of pooling has gained popularity as it helps to maintain the accuracy of the evaluation process. However, it has a drawback - the pool depth cannot be fixed to a specific size. When pooling is done with a fixed pool depth, it may fail to produce enough relevant documents. As the document size increases, the pool depth may need to be increased to maintain the quality of the judgment sets. However, this can result in additional effort, cost, and time for the human assessors. In order to reduce costs and effort, it is necessary to reduce the number of judgments required. One alternative option is to extract the top-k documents and take a 10% sample from that list, which can then be used for evaluation purposes (Buckley et al., 2007). Another option is pooling based on evaluation measures using a methodology called active sampling. This involves using a sampling strategy to determine the runs that are most likely to contain relevant documents and ranking them based on this process. Samples are then retrieved from the better-performing runs, which are evaluated using metrics such as the Horvitz-Thompson estimator (Li & Kanoulas, 2017).

There are different methodologies for pooling in information retrieval. One such method is dynamic pooling, which selects documents from the unjudged list based on the documents already judged. This approach is different from the traditional pooling method but helps in adding more documents to the judgment sets. Dynamic pooling can be done using meta-ranking and statistical sampling techniques such as MFT, hedge, and bandit methods (Cormack & Grossman., 2018). Fair pooling is another way of pooling where a fairness score is applied to create a subset of documents that are as similar as possible for all runs. Opportunistic pooling, on the other hand, creates a subset of documents based on the number of judgments needed and a set threshold value (Tonon et al., 2015).

Rank-biased precision (RBP) is a methodology that selects relevant documents based on a fixed size N and a fixed budget. It considers documents based on document rank probability and examines them in turn, moving from one document to another. If the user prefers the ith document, the probability of moving to the next document is i+1 (Moffat & Zobel, 2007). Moffat and Zobel (2007) proposed three RBP methods: Method A: RBP Abased@N, Summing Contributions, which considers documents selected into the pool based on their overall contributions to the effective evaluation; Method B: RBP Bbased@N, Weighting by residual, which considers documents based on their overall contribution to the pool as well as the weighting of individual documents; and Method C: RBP CBased@N, Raising the power, which increases the score component by raising the power of the current score. Lipani *et al.* (2021) proposed three strategies based on common evaluation measures: Take@N, which chooses the top N documents from RBP runs; DCGBased@N, which applies a discount function to rank documents into the pool; RRF@N, which finds the system effectiveness based on document contribution score; and PPBased@N, which calculates the ratio of the number of relevant documents at rank k to the number of documents in k.

Multi-armed Bandits is a methodology for ordering documents which helps to identify relevant documents for the judgment list or pooled list. This approach was introduced by Losada et al., (2016). The k-armed bandit technique is used to adjudicate meta-search documents, making it easy to add more documents to the judgment list with minimal effort (Losada et al., 2018). To improve the quality of the pooled list, shallow pooling based on preference judgments is done by crowdsourcing. This helps to make more relevant judgments based on mean reciprocal rank and top-judged documents. The runs are re-evaluated to reproduce more documents into the pooled list (Arabzadeh et al., 2021).

Human accessors methodologies

The help of human accessors in finding relevant documents has had a significant impact on the information retrieval evaluation process. However, it may not always be feasible to get their help, especially if the test collection is large. Re-creating the judgment list with human accessors can lead to different decisions in each occurrence, no matter whether it's with the same or different accessors. Disagreement among accessors is a major issue encountered by researchers during the evaluation process (Alonso et al., 2012). Document ambiguity or topic ambiguity may cause disagreements. For instance, differences in the meaning of terms used in documents, unclear information in queries, and assessors' or users' moods or environments can lead to discrepancies. Another major issue is the high cost of utilizing human accessors for each round of the evaluation process. To reduce the cost, many researchers have studied alternatives, such as considering documents only from a pooled list instead of evaluating the entire document list retrieved (Carterette et al., 2008; Cormack, Palmer, & Clarke, 1998), or by reducing the number of topics accessed (Rajagopal & Ravana, 2019). To minimize the need for human assessors, crowdsourcing has been proposed as an alternative solution because it offers several advantages, including cost-effectiveness and flexibility. Crowdsourcing is still a better option for evaluating documents with topic-document pairs than assigning relevance labels to documents. Multiple assessors collect topic-document pairs, and the quality of judgment sets has shown to increase compared to previous ones. Relevancy depends on the distribution of documents and topic pairs among the assessors, not based on the absolute value assigned to the documents (Maddalena et al., 2017). Previous research indicates that there is not a significant disagreement between users and individual human assessors, but there is a considerable discrepancy when multiple human assessors work together. Crowdsourcing has produced better results than expert judges in some cases. During a TREC collection evaluation process, crowdsourcing has been shown to provide accurate and faster judgments at a lower cost (Alonso et al., 2012).

Crowdsourcing with large data sets can be challenging due to various disagreements and issues in the indexing, searching, and catalog creation process. This can lead to a high probability of errors in the judgment process. For instance, the same word with different meanings can affect the quality of retrieval documents. Similarly, different words with the same meaning can lead to incorrect document selection and a reduction in the number of relevant documents in the judgment set (Carpineto & Romano, 2012). To address these issues, the pseudo-relevance judgment process has been introduced. This methodology helps to reduce the effort of human accessors by generating a document ranking for the set of relevant documents. Pseudo-relevance judgments consider two important factors: the frequency of each document for each run from all the systems' runs, and the document ranking. Unlike traditional pooling, which only considers pooled documents from the contributed systems, all the documents from all the systems, including contributed and non-contributed documents, are considered in this methodology (Ravana & Rajagopal, 2015).

The magnitude estimation technique can reduce the effort required by human assessors. In this method, a scale measurement is used for estimation tasks that are assigned to a crowdsource, which results in better outcomes compared to classical binary relevance judgments. This estimation task helps to

evaluate the ranking of documents based on the frequency of terms used in each topic and check the consistency of the ranking of documents in terms of topic understandability. The results have shown overall better performance and a more robust evaluation of the relevancy of documents (Mizzaro et al., 2017). In some research, the evaluation of system effectiveness using existing methods by real users is prone to errors and has a big variation in the results compared to expert judges.

Pair-wise judgment is an alternative solution to human judgment. To rank relevant documents, we need to consider how one document is more relevant than another for a particular topic. Multiple grades of relevance can be created using the pair-wise preference judgment or the nominal graded method. Accessors are required to judge the documents for both these processes. Pair-wise preference judgment is preferred by accessors as it requires only binary marking as either relevant or irrelevant. The nominal graded method is used to assign multiple relevance grades. Pair-wise judgment helps accessors to quickly assign relevancy, making it more popular among researchers. The Elo rating system is used to combine or merge documents using the pair-wise judgment method (Bashir *et al*., 2013). Another pair-wise judgment methodology involves finding a fixed number of relevant document pairs that are purely accurate and using them to auto-generate similar document pairs. This helps to generate many preference judgments based on point-wise judgments, reducing human involvement and increasing system effectiveness (Roitero et al., 2022). Differences in ranks can also be found by considering partial preference based on the top-ranked results. This process involves taking the top-k ranks of the documents and helps to increase the quality of judgment sets (Clarke et al., 2021).

Based on studies of existing methods, a combination of different best methods helps to achieve better results than a single method. This approach is more effective when applied to machine learning algorithms. The frequencies of topic-document pairs resulting from these methodologies help evaluate the performance of the system even without relevant judgment sets (Roitero et al., 2020).

Topics methodologies

The evaluation of a system's performance heavily relies on the topics used. Certain topics produce better relevance judgments compared to others. Researchers often find it challenging to identify the best topics that can generate more relevant judgments (Breto et al., 2013).

The process of evaluating information retrieval typically involves retrieving the maximum number of relevant documents from a document corpus based on topics (Pang et al., 2019). The relevancy of a document is usually predicted based on the topic, but one of the main challenges for researchers is determining the difficulty of a given topic. Topics can be classified as hard, medium, or easy based on the number of relevant documents retrieved, and human assessors tend to choose easy topics over harder ones. This can make it difficult to include relevant documents related to harder topics in the judgment list, which can negatively impact the quality of the judgment sets. Comparing system performance based on topics can be useful in assessing system effectiveness. However, it's worth noting that different sets of topics with the same size can produce different results, and the same sets of topics with different sizes can also produce different results (Berto et al., 2013). The difficulty level of a topic depends on how well it retrieves quality documents. Generally, human accessors prefer easy topics as they provide better results with a smaller pool of documents. Harder topics, on the other hand, may have relevant documents in deeper pools that are not considered in the relevance judgment list. One way to determine the difficulty of a topic is by calculating its average precision. If the average precision is high, the topic is considered easy, while lower average precision indicates a difficult topic for a particular system (Mizzaro, 2008).

The size of a topic has a significant impact on the evaluation score of a system, especially when the topic is difficult. For researchers, it can be challenging to determine which topic pairs or combinations will work best when

dealing with large document collections that require high computational costs and time. Therefore, one alternative solution is to find the difficulty level of the topics in a given run and identify the optimal sets of topics that have contributed to the pool. This will help adjust the topic size accordingly. In most cases, it is observed that easy topics work well with judgment sets, leading to an increased effectiveness score (Pang et al., 2019). Another study suggests that considering the top-k documents from both easy and difficult topics can make the earlier work easier and yield better results even for harder topics. Moreover, the results are consistent across different evaluation metrics (Roitero et al., 2017).

As the number of documents on the Web continues to increase, evaluating information retrieval systems becomes a challenging task for researchers. The effectiveness of these systems can be measured by the quality of topics considered for evaluation and the number of relevance judgments produced (Rajagopal & Ravana, 2019). However, evaluating with a greater number of topics may come at the cost of higher computational expenses and longer processing times. Therefore, most research prefers evaluating with a smaller topic size and easy topics, because even with fewer topics, they can achieve better results and maintain good effectiveness scores (Berto et al, 2018; Carterette et al., 2008). Finding the best topics for retrieval is a challenge, but one effective method of doing so is by using earlier measures such as precision, where the k value is determined by the size of the document collection (Dincer, 2013).

Numerous studies have been conducted to improve the performance of topic modelling while reducing computational costs. Researchers have found that better results can be obtained by reducing the number of topics and increasing the depth of evaluation or increasing the number of topics and reducing the depth of evaluation. However, the number of topics to be used ultimately depends on the user's preference. Human experts tend to favour a lesser range of topics that may have more quality even by considering harder topics. At the same time, actual or real users prioritize comprehensibility and tend to favour easier topics due to their ease of understanding, regardless of their effectiveness. This can lead to relevant documents being left out of the evaluation process, which can negatively affect the system's evaluation score. Therefore, researchers should focus on low-effort or easy-to-understand topics with various evaluation depths to ensure standardized evaluation metrics. Rajagopal and Ravana (2019) highlighted that there is no correlation between system evaluation metrics and real users. Topic modelling is a useful method for evaluating large datasets. It can help to select the best subset of documents and reduce noise.

Topic modelling can be used to create topics for formal, multi-model, or multilingual datasets (Churchill et al., 2021) and can improve the results for multilingual datasets. However, different topic modelling methods and criteria produce varying accuracy, making it difficult to choose the best evaluation metrics (Rudiger et al., 2022). Therefore, it is important to understand the different topic modelling techniques, and which one is best for different content-based datasets.

The interests of real users are crucial in determining the quality of an evaluation metric. To achieve this, we need to identify the topics that real users find interesting, which can help improve the system evaluation score. Topic modelling can use a specific criterion to evaluate the quality of the topics. This methodology has been used in various applications, such as text classifiers and image classifications. By using pre-defined keywords, topic modelling can mine the best topics which retrieve as many relevant documents as possible. Furthermore, it can extract a quality metric based on topics that can predict the number of relevant judgments that real users can give about that particular topic (Nikolenko et al., 2017).

Document similarity methodologies

When dealing with large document collections, traditional techniques such as pooling, sampling, and evaluation metrics can be used to retrieve more relevant documents for evaluation. However, these methods are time-

consuming and computationally expensive. Clustering and classification are well-known techniques that can help overcome these drawbacks. Unlike traditional methods, only the documents within the cluster or class need to be considered for the evaluation process. However, the quality of the judgment sets obtained through clustering and classification is generally lower compared to traditional methods.

Extensive research has been conducted to improve the performance of evaluation through clustering and classification techniques. Clustering can be performed using supervised and unsupervised algorithms (Taha, 2023). One of the most popular methods is combining clustering with frequent itemset mining. The documents are ranked based on relevance and then clustered according to their similarity using k-means clustering. Similar documents are paired based on the frequency of terms they share. When a user query is entered, the terms are used to identify the cluster that contains the maximum number of matching document pairs. The clusters that contain the most frequent terms are then considered, and the documents in these clusters are moved to the judgment sets (Djenouri et al., 2018). To increase the effectiveness of search, clustering, and incremental relevance feedback can be combined. In this clustering, all the documents are clustered based on the initial judgment feedback, not on the ranking of the documents. The best clusters will be found based on density strategies and the documents in them will be sorted by their relevance score. Finally, top-k documents from these clusters will be considered for the evaluation process (Iwayama, 2000).

One common approach to evaluating the retrieval process is to cluster documents based on user queries. These clusters are then used to rank the documents. At the same time, all the clusters are examined to identify similar features of the documents, which in turn assists in ranking the documents. These rankings based on various features help to increase the document similarities in a vector space (Markovskiy et al., 2022). Another approach to improve retrieval effectiveness through clustering is by incorporating topic modelling. Each topic in the cluster is evaluated with a set of terms in the document collection and the frequency of each term is analysed. Topics with the same frequency are considered to represent various themes. This methodology helps to retrieve meaningful representations of clusters and also predicts the quality of the clusters (Yuan et al., 2021). Clustering documents using k-means is an effective way to group them and retrieve more relevant information (Aliwy et al., 2022; Wang, 2021). Another concept is clustering based on findability effort, which involves grouping documents as either relevant or irrelevant based on the effort required to find them. The results of system-based evaluations show that the performance of different systems varies when findability effort is combined with relevance (Rajagopal et al., 2022).

Classification can be done by simply classifying the documents based on their similarity. Active Learning Algorithm is a methodology which is based on classification technique, and which does not consider pooling and system ranking. For each topic, a topic-specific document classification is considered. This methodology selects a subset of the documents first and classifies them based on their similarity. A set of randomly chosen relevant and non-relevant documents will be selected by the judges and based on these, similar documents will be found and classified. This technique considers document selection and labelling of documents that have not been considered in the judgment sets. Comparing the subset selected and the documents that have not been included in the classified list helps to improve the relevance score of the judgment list. However, using this methodology might create biases in the evaluation process when considering the subset of the documents. As an alternative, a hybrid combination of human assessors and automated classification techniques has been considered (Rahman et al., 2020). In many evaluation processes, only the documents that are included in a pooled list are considered for evaluation, and those not on the list are deemed irrelevant. However, a more effective methodology involves training a classifier on the pooled list and using it to identify similar

documents from the irrelevant sets. Any documents found to be similar are moved into the judgment sets, improving the system's effectiveness (Büttcher et al., 2007). Additionally, calculating both frequency terms and sparse data in various dimensions can help to find similarity measures and increase the performance score of the classified documents. This involves classifying the documents based on term frequencies, finding centroids, and creating a vector space model. This methodology has resulted in increased precision, recall, and F1 scores (the mean of the precision and recall of a classification model) in many evaluation metrics (Eminagaoglu, 2022).

The issues of existing methodologies in the relevance of judgment-set performance

A large number of relevant documents in the judgment sets, also called *qrels*, helps to increase the accuracy of the evaluation process. Pooling helps to increase the quality of the judgment sets, but it considers the whole document corpus, which is time-consuming. Document similarity through clustering and classification will overcome the issue faced in pooling as it considers only one cluster or class with a high similarity score. However, the drawback of document similarity is the quality of the documents retrieved through document similarity is less compared to the pooling (Djenouri et al., 2018). Human accessors' judgment through crowdsourcing is more error prone compared to that of expert judges. It always varies depending on the user's readability effort, understandability effort, and findability effort (Rajagopal and Ravana, 2019). The methodology through topics is also quite challenging. Based on topic hardness, the retrieval efficiency also varies. Users will be able to retrieve more relevant documents which are with easy topics. For hard topics, retrieval of relevant documents for both experts and real users will be difficult due to topic hardness (Ravana et al, 2015; Roitero et al., 2020).

Many studies have been done with human accessor methodologies and topic selection methodologies to overcome the limitations of the improvements of relevance judgment sets. However, it has been noticed that more exposure is needed in pooling and document similarity methodologies. So, the baseline works considered for the experiments are based on pooling and on document similarity with clustering and classification techniques. Three methodologies from the existing works were considered for the evaluation purpose. One pooling methodology merged documents from the runs based on the Combsum rank aggregation technique. From these merged ranked lists, top-k relevant documents from each run have been considered and given for the evaluation process (Losada et al., 2018). The other two methodologies were based on document similarity, and based on classification and clustering techniques. The cluster-based methodology, named ICIR (Intelligent cluster-based Information Retrieval), combines k-means clustering with frequent itemset mining to extract the clusters of documents to find the frequent terms in the cluster. Whenever a new user query comes, the patterns are discovered in each cluster, and we can determine the most relevant clusters that match the user query. The clustered documents are considered for the evaluation process (Djenouri et al., 2021). The classification-based methodology, namely CAL (Continuous Active Learning), considers a set of documents based on the active learning algorithm, which considers documents that might be chosen by the accessors. Based on this subset, the active learning algorithm automatically classifies the unjudged documents (Rahman et al., 2020).

The baseline experiment with these methodologies has been done with the TREC-8 Adhoc Track collection. Fig. 1 shows the baseline line experiment results. It has been noticed that compared to the document similarity methodology, the pooling methodology performs better. However, in both methodologies, the average number of relevant documents retrieved based on the number of judgments is less. The less relevant documents in the judgment sets affects the quality of the judgment sets and through that, it affects the effectiveness of the system's performance and also the accuracy of the evaluation process.

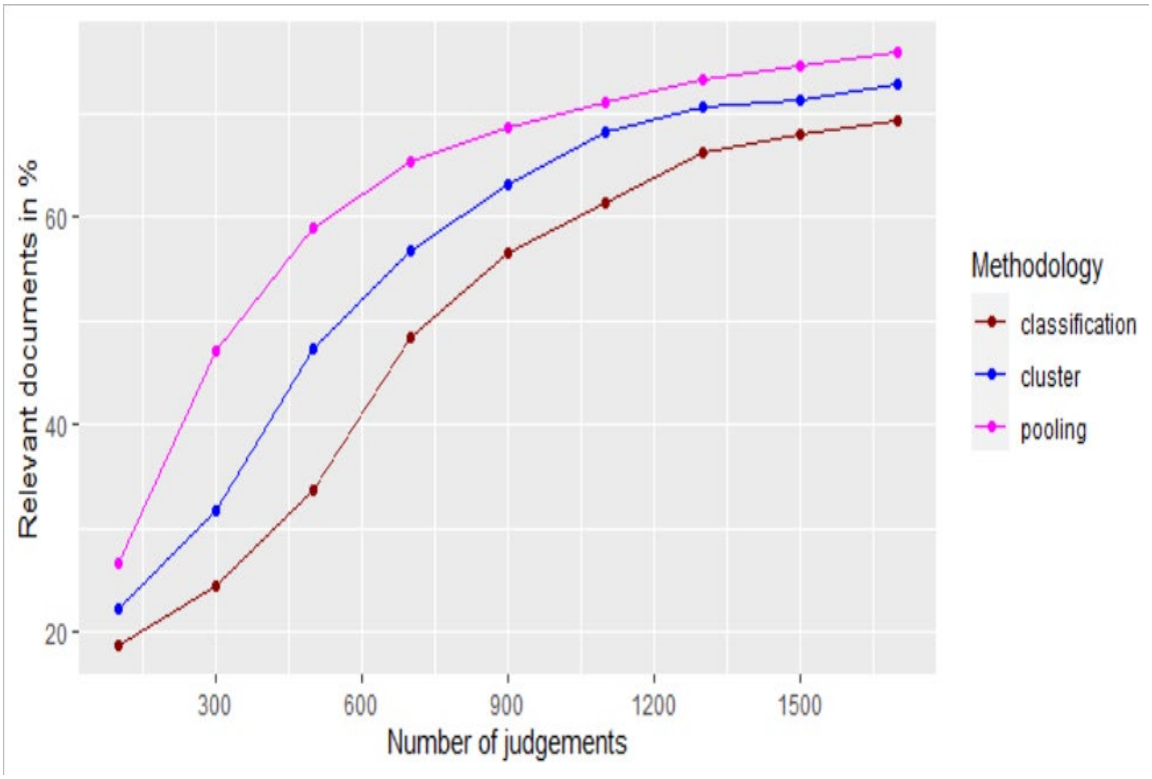

Figure 1. Baseline experiments based on pooling and document similarity

cluster- (ICIR) - (Djenouri et al., 2018), classification- (CAL)- (Rahman et al., 2020), pooling- (Losada et al., 2018)

Methodology

Increase effectiveness based on pooling and document similarity methodologies

To improve the quality of the relevant judgment sets, a new methodology has been proposed. The TREC dataset was used for this experiment which consists of document corpus, topics, and relevant judgment sets. Figure 2 shows the experimental methodology proposed. Each participating system retrieves a set of relevant documents from the document corpus based on the topics. Each of these documents sets retrieved by each participating system will be called runs. Each run will be ranked according to its relevancy. These runs will be merged using the Combsum ranking algorithm. The pooling technique applied on these runs and chosen the top relevant documents based on top-k technique. k value depends on the number of documents that need to be judged (e.g., 10, 100, and 200). These documents are called pooled documents ($p_i...p_k$). The documents that have not been considered in the pooled list are called unjudged document lists. These documents will be clustered based on their relevancy $(U_{Cx,Jy})$. This indicates yth unjudged document of J with the xth cluster of C. Document similarity has been carried out between these pooled documents and unjudged documents $(Sim_{(Pi)})$. If a similarity is found, these documents will be moved into the pooled list by assigning new scores for those documents. The evaluation has been done based on a pool depth of 100 and an evaluation depth of 1000.

New Score assigning = $\frac{sim(di,U_{Cij})}{\sum_{l=1}^{k} sim(dl,U_{Cij})}$ (Liang et al., 2018)

Once enough documents are moved into the pooled list, the pooled list will be re-ranked. Through that, the relevant judgment set quality can be increased with the number of relevant documents. The same experiment was done with the classification technique also.

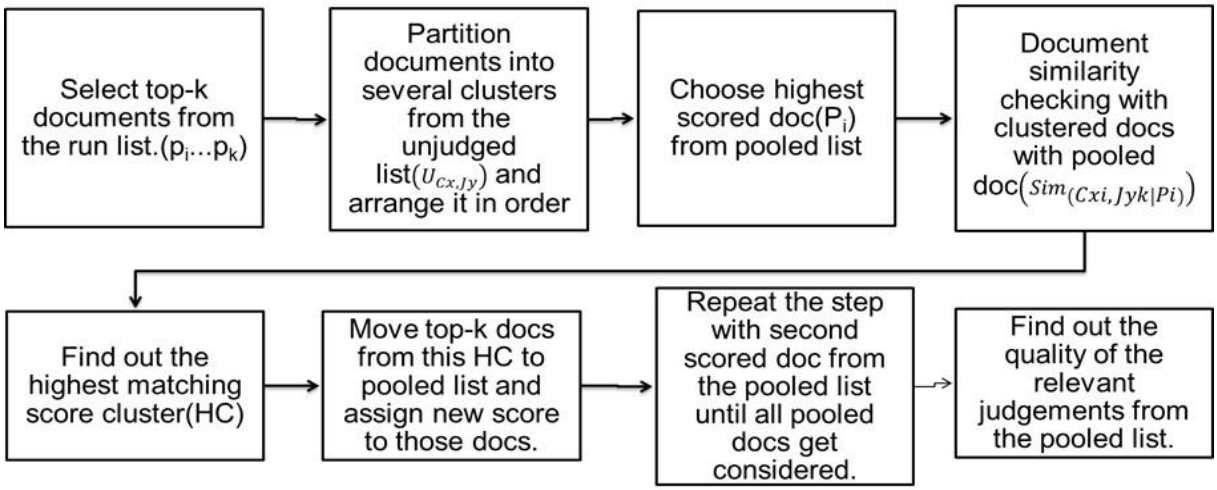

Figure 2. Experimental methodology based on pooling and document similarity

Increase effectiveness based on system evaluation scores

To improve the quality of the relevance judgment sets, a new methodology which is an enhancement of the methodology based on pooling and document similarity has been proposed. The methodology tried to retrieve more better results by considering documents from the good participating systems. Good participating systems contribute many relevant documents to the judgment sets and at the same time rank these documents efficiently according to their relevancy. With those advantages, if the pool depth size is lesser, more relevant documents can move into the pooled list and can be it for the evaluation process. The good participated systems have been found based on the evaluation measure called average precision, AP. The average precision has been calculated based on the top 100 documents, say AP@100. The evaluation has been done based on a pool depth of 100 and an evaluation depth of 1000. Figure. 3 shows the proposed experimental methodologies based on contributed systems effectiveness based on evaluation score.

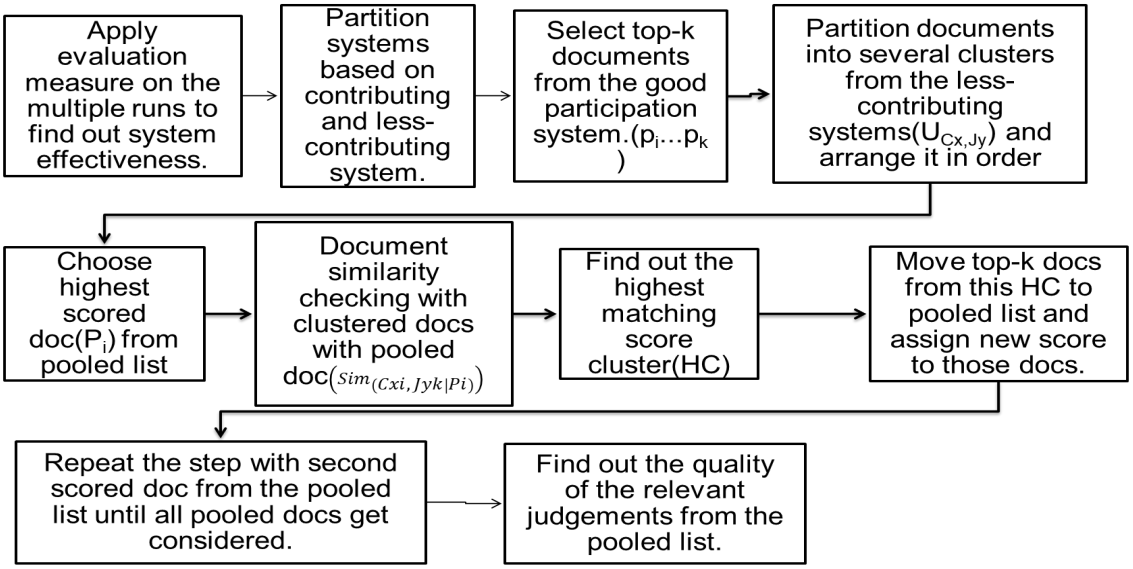

Figure 3. Experimental Methodology based on evaluation score

Experiments and results

The experiments were conducted using the TREC dataset (https://trec. nist. gov/data. html) which contains document corpus, topics, and a set of relevant judged documents. Two TREC datasets, TREC-8 and TREC-10, with 50 topics each, were used for the evaluation process (Table 1).

Dataset	Number of topics	Topics	Total systems
TREC-8	50	401-450	129
TREC-10	50	501-550	97

Table 1. TREC datasets overview

The proposed methodology based on pooling and document similarity was able to retrieve the same number of relevant documents as the baseline works. Figure 4 and Figure 5 show the results of how percentages of relevant documents were retrieved by various methodologies using the datasets. The x-axis shows the number of judgments, and the y-axis shows the relevant documents retrieved in percentages. Five methodologies were implemented: *cluster* (ICIR; Djenouri et al., 2018), *classification* (CAL; Rahman *et al.*, 2020), *pooling* (Losada et al., 2018). In addition, *pool_cluster* and *pool_classification* are the proposed methodologies.

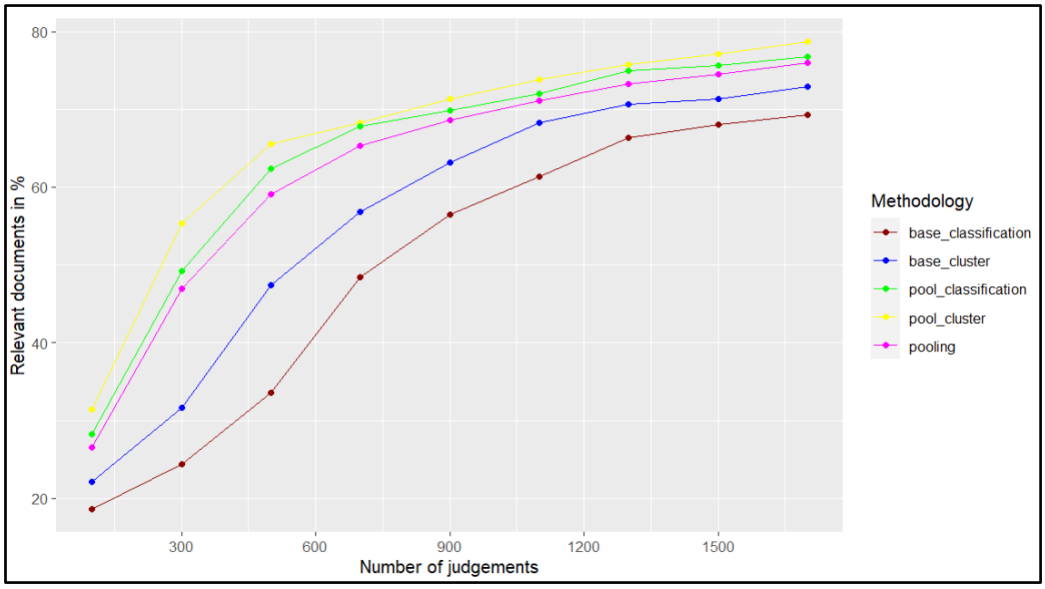

Figure 4. Relevant documents retrieved using various methodologies (in %) using TREC-8 dataset

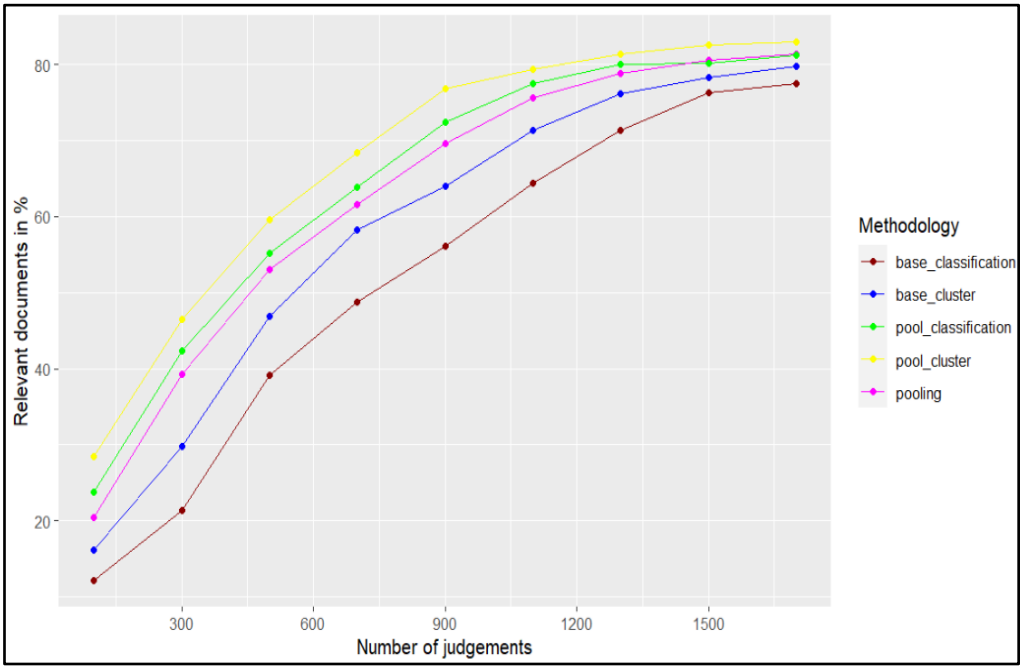

Figure 5. Relevant documents retrieved using various methodologies (in %) using the TREC-10 dataset

Based on the results, the proposed methodologies based on pooling and clustering, and pooling and classification, produced better results compared to the existing methodologies. Compared to the baseline works, *base_cluster*, it has been shown that *pool_cluster* produces more relevant documents in the judgment sets even with lesser depth. In the same way, compared to the *base_classification* methodology, the *pool_classification* produces more relevant documents in the judgment sets with lesser pool depth. For the TREC-8 test collection, as a detailed view, for the top 300 relevant documents, the *base_cluster* has retrieved only 32.1% of relevant documents. The proposed methodology based on *pool_cluster* has produced 57.2% of relevant documents with the

top 300 judgments. Compared to the *pooling* methodology, 20.8% of more relevant documents were retrieved. The same performance difference can be viewed in the TREC-10 dataset collection also.

So, it has been proven that with a lesser pool depth, the proposed methodology helped to retrieve a greater number of relevant documents. Also, it has been noticed that as long as the pool depth increases or the number of documents in the judgment sets increases, there are not many significant differences noticed compared to the baseline works. This is because, as the judgment size is increases, most of the relevant documents have been retrieved and moved to the judgment sets.

Figure 6 shows the results of the proposed methodologies with score-based evaluation scores. The score-based evaluation score is represented as *pool_classification_evaluation_score* and *pool_cluster_ evaluation_score*. The results show that the proposed methodology with evaluation scores performed better than experimental methodology performed in Figure 2. More relevant documents were able to be retrieved with the lesser pool depth itself. As long as the pool depth is increasing, the results go closer as it is due to most of the documents got participated in the judgement list. If the system can retrieve more relevant documents with lesser pool depth, this helps to increase the effectiveness of the evaluation process and also it is cost-effective.

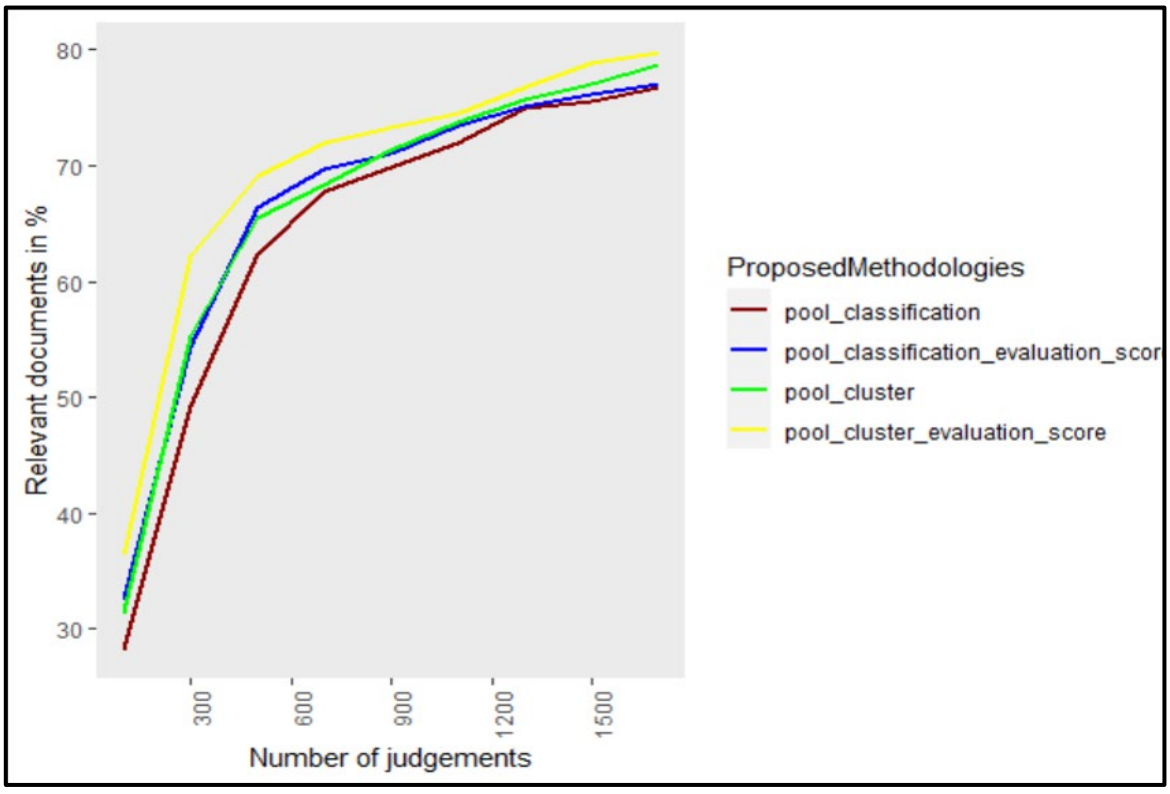

Figure 6. Comparison of the proposed methodologies (in %) using the TREC-8 dataset

Methodology	TREC-8 (Adhoc Track)	TREC-10 (Web Track)
Classification (CAL)	0.693	0.704
Clustering (ICIR)	0.748	0.726
Pooling(Combsum)	0.751	0.781
Pooling+Classification	0.772	0.784
Pooling +Clustering	**0.794**	**0.81**

Table 2. Mean average precision (MAP) results based on pooling and document similarity

Table 2 shows the mean average precision (MAP) of all the methodologies over all the systems and all topics that contributed well to the runs. The results clearly show that the proposed methodologies, pooling+ classification and pooling+ clustering retrieve better results compared to the baseline methods.

Table 3 shows the mean average precision (MAP) of all proposed methodologies. The results clearly show that the methodologies proposed based on the evaluation score produced a greater number of relevant documents as compared to the previous proposed ones.

Methodology	**TREC-8 (Ad hoc)**	**TREC-10 (Web Track)**
Pooling+Classification	0. 772	0. 784
Pooling+Clustering	0. 794	0. 81
Pooling+Classification+Evaluation_Score	0. 806	0. 801
Pooling+Cluster+Evaluation_Score	0. 827	0. 835

Table 3. Mean average precision (MAP) results based on proposed methodologies

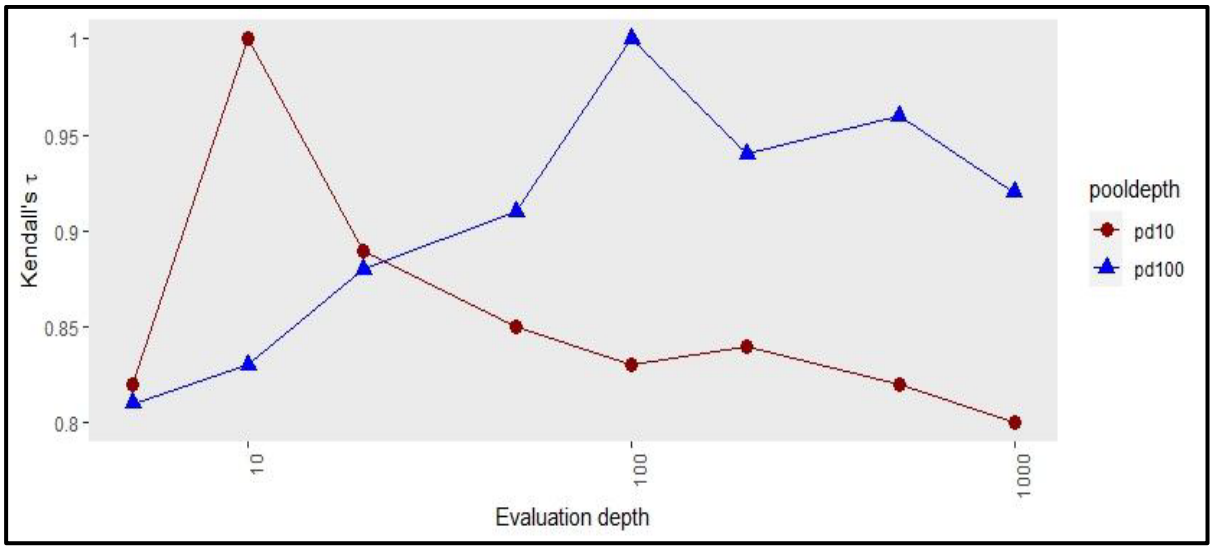

Figure 7. nDCG@d, where d is either 10 or 100, results from TREC-8

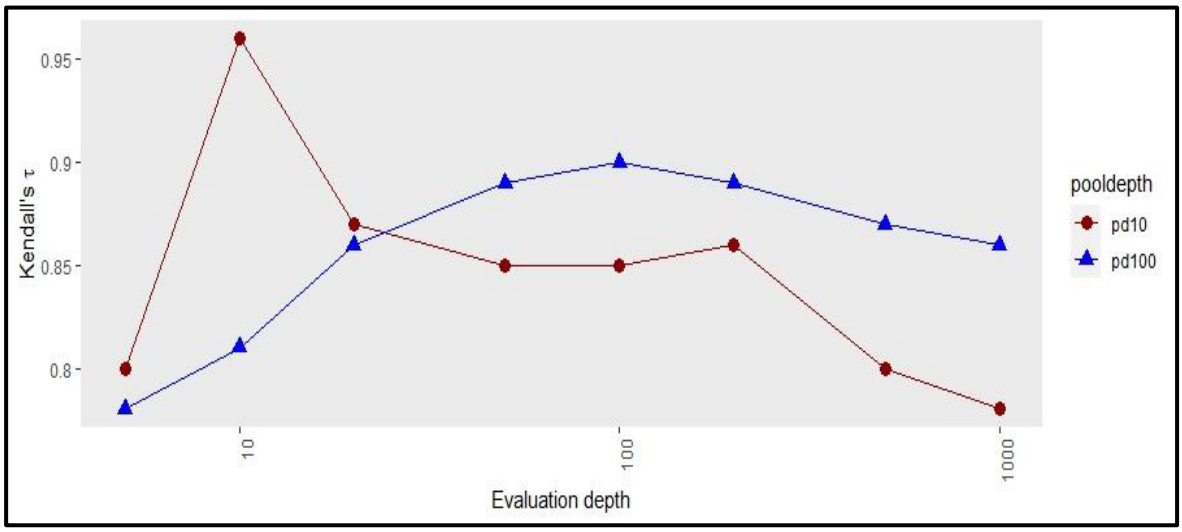

Figure 8. RBP@d, where d is either 10 or 100, results from TREC

Figure 7 and Figure 8 show the NDCG and RBP correlation values based on TREC-8. The x-axis shows the various evaluation depths and the y-axis shows Kendall's τ correlation between different pool depths and evaluation depth. With pool depth 10, the correlation has been calculated with different evaluation depths. With pool depth 100, the correlation has been calculated with and different evaluation depths. The results show that if the pool depth is greater than the evaluation depth or the pool depth is equal to the evaluation depth, the correlation values are higher. Once the pool depth value becomes less than the evaluation depth and the number of relevant documents is higher than the pool depth, there is a significant variance in the result of systems correlation. This happens because when the evaluation depth goes higher, more irrelevant documents might move into the judgment list, and this might be the reason for the variance in the results.

In order to view a more accurate result, the relevance judgments from the proposed methodologies were split into six different sets with varied sizes (1%, 20%, 40%, 60%, and 100%) and the performance of each set evaluated.

Mean average precision (MAP) has been used to evaluate the quality or the accuracy of the relevant judgment sets. Bpref measures have been used to evaluate the biasness of the number of irrelevant documents in the judgment sets and the bias in the ranking of the relevant documents in the judgment sets.

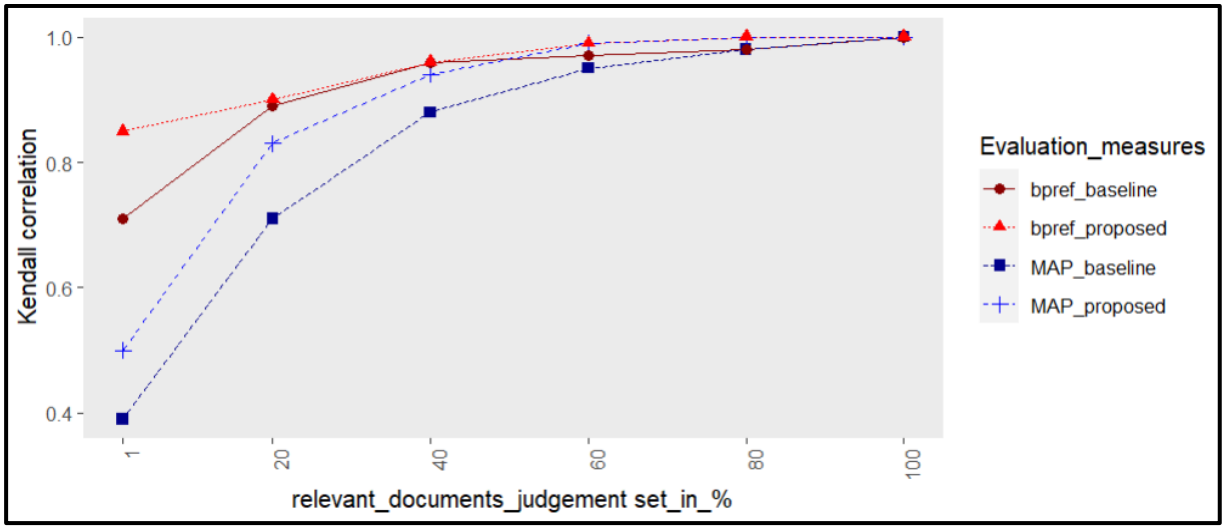

Figure 9. Changes in Kendall correlation of measures based on different judgment sets in TREC-8

Figure 9 shows Kendall τ correlations between the system ranking produced using the 100% *qrels* and the system ranking produced using the same measures but a reduced *qrels*, based on the proposed methodologies. The plot for the *bpref* measure is flatter than the plots for the other measures, indicating that the *bpref* measure continues to rank different systems in the same relative order as when using complete judgments for higher levels of incompleteness. Hence this graph shows that the proposed methodology works well with the mean average precision which shows the quality of the relevant documents retrieved and the consistency in maintaining the ranking of the documents.

Discussion

In general, in information retrieval, the user's expectation is to retrieve as many relevant documents based on the user query as possible. When a user sends a query to the participating systems, the systems collect a set of documents from the document corpus or the Web collections. These documents are then ranked according to their relevance, and the ranked documents are sent back to the user. The documents received by the end user must be relevant to the user's query. This will help the users to upholding the system's reliability and ensure the user's satisfaction. However, it has been seen that user satisfaction can only be achieved with smaller datasets. But for the large test collections, it is not easy to get enough number of relevant documents into the judgment sets. Thus, this experiment mainly concentrated on how to increase the number of relevant documents in the relevance judgment sets using different methodologies.

Pooling judgments helps to retrieve more relevant documents, which improves the quality of the judgment set. It is cost-effective and requires less time for evaluation since it only considers the top-k documents. However, a drawback of pooling is that only the top-k documents from each run are considered relevant, and documents that don't make it into the pooled list are considered irrelevant and will not be evaluated. These unjudged documents may contain relevant information, but due to system inefficiencies, they are ranked lower and not included in the pooled list (Cormack et al., 2018; Losada et al., 2018). When dealing with large amounts of data, using clustering and classification to determine

document similarity is faster than traditional methods, because it evaluates only the clusters or classes with high similarity scores. However, a drawback is that the quality of documents retrieved through clustering and classification is often lower than with traditional methods. Additionally, documents within the same class or cluster are considered to be identical and assumed to have a similar score. When a new query arises, document similarities are mainly considered based on term frequencies. Identifying the best cluster or class based on this similarity is a challenging process (Djenouri et al., 2018; Rahman et al., 2020).

Based on the research objectives, this work aimed to increase the number of relevant documents in the relevance judgment sets and through that increase the quality of the judgment sets. This will indirectly help to increase the accuracy of the evaluation process and the effectiveness of the contributed or the participated systems.

Based on the first aim of this work, the accuracy of the evaluation process can be increased by increasing the number of relevant documents in the judgment sets compared to the baseline works. Based on the second aim of this work, the effectiveness of the contributed systems can be increased by reducing the biasness in the ranking of the documents and by considering more relevant documents. To achieve this, two methodologies were proposed and these methodologies helped to achieve a better result compared to the baseline works. Figure 9 shows a clear view of the proposed methodologies' performance on increasing the quality of the judgment sets and also, the reduction and the consistency of the biasness in the ranking of documents even with different relevant judgment sets sizes.

Conclusions To improve the accuracy of the evaluation process, it is crucial to increase the quality of the judgment set. It can be achieved through increasing the number of relevant documents in the relevance judgment sets and also with reduced biasness in the ranking of these documents in the judgment sets. It has proven that the proposed methodologies achieved a better result compared to the baselines. The results also show that the proposed methodologies perform better even with a smaller pool depth and that when the pool depth is greater than or equal to the evaluation depth. However, the results also show that when the number of relevant documents exceeds the pool depth or the evaluation depth is greater than the pool depth, the systems and methodologies' performance may vary significantly.

Acknowledgments

This research work has been supported by the University of Malaya International Collaboration Grant [Grant No. : ST080-2022]

About the authors

Minnu Helen Joseph received a Master's degree in Computer Science and Engineering in 2009 from Karunya Institute of Technology, India. She is currently pursuing her Ph. D degree in Information Retrieval Evaluation from the University Malaya, Malaysia. She is currently working as a Lecturer at Asia Pacific University, Malaysia. Her research interests include Information Retrieval Evaluation, Data Analytics, Data Science, and Machine Learning.

Sri Devi Ravana received her Master's degree in Software Engineering in 2001 from the University Malaya, Malaysia, and completed her Ph. D at The University of Melbourne, Australia in 2011. Currently, she is an Associate Professor at the University Malaya, Malaysia. Her areas of expertise include Information Retrieval, Text Heuristics, IR Evaluation, Data Analytics, Data Science, and Search Engines. She can be contacted at sdevi@um.edu.my

References

Aliwy, A. H., Aljanabi, K., & Alameen, H. A. (2022). Arabic text clustering technique to improve information retrieval. *AIP Conference Proceedings*, *2386*(1). AIP Publishing. https://doi.org/10.1063/5.0066837

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, *48* (6), 1053-1066. https://doi.org/10.1016/j.ipm.2012.01.004

Arabzadeh, N., Vtyurina, A., Yan, X., & Clarke, C. L. (2021). Shallow pooling for sparse labels. *arXiv preprint arXiv:2109. 00062* https://doi.org/10.1007/s10791-022-09411-0

Bashir, M., Anderton, J., Wu, J., Golbus, P. B., Pavlu, V., & Aslam, J. A. (2013). A document rating system for preference judgements. In *SIGIR13: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 909-912). https://doi.org/10.1145/2484028.2484170

Berto, A., Mizzaro, S., & Robertson, S. (2013). On using fewer topics in information retrieval evaluations. In O. Kurland, D. Metzler, C. Lioma, B. Larsen, & P. Ingersen (Eds.), *ICTIR '13: Proceedings of the 2013 Conference on the Theory of Information Retrieval* (pp. 30-37). https://doi.org/10.1145/2499178.2499184

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 25-32). ACM. https://dl.acm.org/doi/10.1145/1008992.1009000

Buckley, C., & Voorhees, E. M. (2017). Evaluating evaluation measure stability. *ACM SIGIR Forum*, 51(2), 235-242. https://doi.org/10.1145/3130348.3130373

Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2007). Bias and the limits of pooling for large collections. *Information Retrieval*, 10, 491-508. https://doi.org/10.1007/s10791-007-9032-x

Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I. (2007). Reliable information retrieval evaluation with incomplete and biased judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 63-70). ACM. https://doi.org/10.1145/2499178.2499184

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44 (1), article 1. https://doi.org/10.1145/2071389.2071390

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & Allan, J. (2008). Evaluation over thousands of queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 651-658). https://doi.org/10.1145/1390334.1390445

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, 54 (10s), article 215. https://doi.org/10.1145/3507900

Clarke, C. L., Vtyurina, A., & Smucker, M. D. (2021). Assessing Top-Preferences. *ACM Transactions on Information Systems*, 39 (3), article 33. https://doi.org/10.1145/3451161

Cormack, G. V., & Grossman, M. R. (2018). Beyond pooling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1169-1172). https://doi.org/10.1145/3209978.3210119

Cormack, G. V., Palmer, C. R., & Clarke, C. L. (1998). Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 282-289). http://doi.org/10.1145/290941.291009

Culpepper, J. S., Mizzaro, S., Sanderson, M., & Scholer, F. (2014). TREC: Topic engineering exercise. In *Proceedings of the 37th international ACM SIGIR conference on Research & Development in Information Retrieval* (pp. 1147-1150). https://doi.org/10.1145/2600428.2609531

Dinçer, B. T. (2013). Design of information retrieval experiments: The sufficient topic set size for providing an adequate level of confidence. *Turkish Journal of Electrical Engineering and Computer Sciences*, 21 (8), 2218-2232 https://doi.org/10.3906/elk-1203-20

Djenouri, Y., Belhadi, A., Fournier-Viger, P., & Lin, J. C. W. (2018). Fast and effective cluster-based information retrieval using frequent closed itemsets. *Information Sciences*, 453, 154-167. https://doi.org/10.1016/j.ins.2018.04.008

Djenouri, Y., Belhadi, A., Djenouri, D., & Lin, J. C. W. (2021). Cluster-based information retrieval using pattern mining. *Applied Intelligence*, 51(4), 1888-1903. https://doi.org/10.1007/s10489-020-01922-x

Eminagaoglu, M. (2022). A new similarity measure for vector space models in text classification and information retrieval. *Journal of Information Science*, 48 (4), 463-476. https://doi.org/10.1177/01655515209680

Ferro, N. (2017). Reproducibility challenges in information retrieval evaluation. *Journal of Data and Information Quality*, 8 (2), article 8. https://doi.org/10.1145/3020206

Guiver, J., Mizzaro, S., & Robertson, S. (2009). A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems*, 27 (4), article 21. https://doi.org/10.1145/1629096.1629099

Hembrooke, H., Pan, B., Joachims, T., Gay, G., & Granka, L. (2005). In Google we trust: Users' decisions on rank, position and relevancy. *Journal of Computer-Mediated Communication*, 12(3), 801-823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 17-24). ACM. https://doi.org/10.1145/345508.345539

Iwayama, M. (2000). Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 10-16). ACM. https://doi.org/10.1145/345508.345538

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422-446. https://doi.org/10.1145/582415.582418

Li, D., & Kanoulas, E. (2017). Active sampling for large-scale information retrieval evaluation. In *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 49-58). ACM. https://doi.org/10.1145/3132847.3133015

Liang, S., Markov, I., Ren, Z., & de Rijke, M. (2018). Manifold learning for rank aggregation. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1735-1744). ACM. https://doi.org/10.1145/3178876.3186085

Lipani, A., Carterette, B., & Yilmaz, E. (2021). How am I doing?: Evaluating conversational search systems offline. *ACM Transactions on Information Systems*, 39(4), article 51. https://doi.org/10.1145/3451160

Losada, D. E., Parapar, J., & Barreiro, Á. (2016). Feeling lucky? Multi-armed bandits for ordering judgements in pooling-based evaluation. In *SAC 16: Proceedings of the 31st annual ACM symposium on applied computing* (pp. 1027-1034). ACM. https://doi.org/10.1145/2851613.2851692

Losada, D. E., Parapar, J., & Barreiro, A. (2018). A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion*, 39, 56-71. https://doi.org/10.1016/j.inffus.2017.04.001

Maddalena, E., Roitero, K., Demartini, G., & Mizzaro, S. (2017). Considering assessor agreement in IR evaluation. In *ICTIR '17: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 75-82). https://doi.org/10.1145/3121050.3121060

Mandl, T. (2008). Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica*, 32(1), 27–38.

Markovskiy, E., Raiber, F., Sabach, S., & Kurland, O. (2022). From cluster ranking to document ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2137-2141). ACM. https://doi.org/10.1145/3477495.3531819

Melucci, M., & Baeza-Yates, R. (2011). Chapter 4. The user in interactive information retrieval evaluation. *Advanced topics in information retrieval.* Springer. https://dl.acm.org/doi/book/10.5555/2564839

Mizzaro, S. (2008). The good, the bad, the difficult, and the easy: something wrong with information retrieval evaluation? In *ECIR 2008: Advances in Information Retrieval: 30th European Conference on IR Research* (pp. 642-646). Springer. https://doi.org/10.1007/978-3-540-78646-7_71

Mizzaro, S., & Robertson, S. (2007). Hits hits TREC: Exploring IR evaluation results with network analysis. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 479-486). ACM. https://doi.org/10.1145/1277741.1277824

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27 (1), article 2. https://doi.org/10.1145/1416950.1416952

Moffat, A., Webber, W., & Zobel, J. (2007). Strategic system comparisons via targeted relevance judgments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 375-382). ACM. https://doi.org/10.1145/1277741.1277806

Moghadasi, S. I., Ravana, S. D., & Raman, S. N. (2013). Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics*, 7(2), 301-312. https://doi.org/10.1016/j.joi.2012.12.001

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, 43(1), 88-102. https://doi.org/10.1177/0165551515617393

Pang, W. T., Rajagopal, P., Wang, M., Zhang, S., & Ravana, S. D. (2019). Exploring Topic Difficulty in Information Retrieval Systems Evaluation. *Journal of Physics: Conference Series*, 1339(1), paper 012019). https://iopscience.iop.org/article/10.1088/1742-6596/1339/1/012019

Rahman, M. M., Kutlu, M., Elsayed, T., & Lease, M. (2020). Efficient test collection construction via active learning. In *ICTIR '21: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (pp. 177-184). ACM. https://doi.org/10.1145/3409256.3409837

Rajagopal, P., & Ravana, S. D. (2019, September). Effort-based information retrieval evaluation with varied evaluation depth and topic sizes. In *Proceedings of the 3rd International Conference on Business and Information Management* (pp. 143-147). https://doi.org/10.1145/3361785.3361794

Rajagopal, P., Aghris, T., Fettah, F. E., & Ravana, S. D. (2022). Clustering of relevant documents based on findability effort in information retrieval. *International Journal of Information Retrieval Research*, 12(1), 1-18. https://doi.org/10.4018/IJIRR.315764

Ravana, S. D., Rajagopal, P., & Balakrishnan, V. (2015). Ranking retrieval systems using pseudo relevance judgments. *Aslib Journal of Information Management*, 67(6), 700-714. https://doi.org/10.1108/AJIM-03-2015-0046

Roitero, K., Checco, A., Mizzaro, S., & Demartini, G. (2022, April). Preferences on a budget: Prioritizing document pairs when crowdsourcing relevance judgments. In *Proceedings of the ACM Web Conference 2022* (pp. 319-327). ACM. https://doi.org/10.1145/3485447.3511960

Roitero, K., Culpepper, J. S., Sanderson, M., Scholer, F., & Mizzaro, S. (2020). Fewer topics? A million topics? Both?! On topics subsets in test collections. *Information Retrieval Journal*, 23(1), 49-85. https://doi.org/10.1007/s10791-019-09357-w

Roitero, K., Maddalena, E., & Mizzaro, S. (2017). Do easy topics predict effectiveness better than difficult topics? In *ECIR 2017: Advances in Information Retrieval: 39th European Conference on IR Research* (pp. 605-611). Springer International Publishing. https://doi.org/10.1007/978-3-319-56608-5_55

Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T. O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *Plos one*, (4), e0266325. https://doi.org/10.1371/journal.pone.0266325

Sparck-Jones, K., & Van Rijshergen, C. J. (1975). *Report on the need for and provision of an 'ideal' Information retrieval test collection.* University Computer Laboratory, Cambridge. Retrieved from https://sigir.org/files/museum/pub-14/pub_14.pdf

Taha, K. (2023, March). Semi-supervised and un-supervised clustering: A review and experimental evaluation. *Information Systems*, 14, 102178. https://doi.org/10.1016/j.is.2023.102178

Tonon, A., Demartini, G., & Cudré-Mauroux, P. (2015). Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval Journal*, 18, 445-472. https://doi.org/10.1007/s10791-015-9266-y

Turpin, A. H., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 225-231). ACM. https://doi.org/10.1145/383952.383992

Voorhees, E. M. (2001). The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for European languages* (pp. 355-370). Springer. https://doi.org/10.1007/3-540-45691-0_34

Wang, X., Macdonald, C., Tonellotto, N., & Ounis, I. (2021, July). Pseudo-relevance feedback for multiple representation dense retrieval. In *ICTIR '21: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 297-306). ACM. https://doi.org/10.1145/3471158.3472250

Zuva, K., & Zuva, T. (2012). Evaluation of information retrieval systems. *International journal of computer science & information technology*, 4(3), 35-43. https://doi.org/10.5121/ijcsit.2012.4304