IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 January 2024, accepted 4 March 2024, date of publication 13 March 2024, date of current version 16 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377239

TOPICAL REVIEW

Reliable Information Retrieval Systems Performance Evaluation: A Review

MINNU HELEN JOSEPH[1,2] **AND SRI DEVI RAVANA**[2]
[1]School of Computing, Asia Pacific University of Technology and Innovation (APU), Kuala Lumpur 57000, Malaysia
[2]Faculty of Computer Science and Information Technology, University of Malaya (UM), Kuala Lumpur 50603, Malaysia

Corresponding author: Sri Devi Ravana (sdevi@um.edu.my)

This work was supported in part by the Ministry of Higher Education Malaysia via Fundamental Research Grant Scheme under Grant FRGS/1.2020/ICT06/UM/02/1, and in part by the UM International Collaboration Grant under Grant ST080-2022.

ABSTRACT With the progressive and availability of various search tools, interest in the evaluation of information retrieval based on user perspective has grown tremendously among researchers. The Information Retrieval System Evaluation is done through Cranfield-paradigm in which the test collections provide the foundation of the evaluation process. The test collections consist of a document corpus, topics, and a set of relevant judgments. The relevant judgments are the documents which retrieved from the test collections based on the topics. The accuracy of the evaluation process is based on the number of relevant documents in the relevance judgment set, called qrels. This paper presents a comprehensive study, which discusses the various ways to improve the number of relevant documents in the qrels to improve the quality of qrels and through that increase the accuracy of the evaluation process. Different ways in which each methodology was performed to retrieve more relevant documents were categorized, described, and analyzed, resulting in an inclusive flow of these methodologies.

INDEX TERMS Document similarity, human accessors, information retrieval, information systems evaluation, pooling, topics.

I. INTRODUCTION

There has been a massive growth of the World Wide Web every day. Whenever a user tries to search for a particular information from the retrieval systems, a set of documents is retrieved based on the query which entered by the user. These retrieved documents are supposed to be relevant to the user query and it makes the satisfaction for the users to relay on the system again. The ranked list of these retrieved documents is ordered based on the degree of relevance to the query. Each retrieval system produces different ranked list documents based on their relevance. The only way of finding out which systems produced more relevant documents can be through an evaluation process [3].

The evaluation process can be done in two ways, User-based evaluation, and System-based evaluation [1]. User-based evaluation depends on user satisfaction or feedback and System-based evaluation is completely based on the test collection and focuses on how well the system can produce a greater number of relevant documents. User-based evaluation is comparatively high cost as it requires many users to participate in the process. In System based evaluation the same test collection can be repeated for any number of experiments [1]. The main aim of the information retrieval evaluation is to find out how well a system produces a maximum number of relevant documents and at the same time, suppress the irrelevant ones [41]. System-based evaluation is preferable compared to user-based evaluation.

Test collections are mostly used tools for the evaluation process [13]. The evaluation is completely based on a set of test collections which consists of a set of documents called document corpus, a set of queries called topics, and a set of relevance judgements. Some of the well-known test collection campaign models are the Text Retrieval Conference1, the Cross-Language Evaluation Forum2, the Community for Information Access Research project3, the NII Testbeds, and the Initiative for the Evaluation of XML Retrieval4 [2]. These test collections have a major role in increasing the

The associate editor coordinating the review of this manuscript and approving it for publication was Hai Dong.

quality of research in the evaluation-based methodologies, pooling concepts, human accessors involvements, topic selection, extraction of relevant documents, and the involvement of significance testing. The reusability of test collections is also an advantage. Much research has been done in the view of reusing these test collections. The judgments were more consistent in judging documents by the human accessors. Even though the decision-making takes time difference, the similarity in the judgment had a scope in reusing the documents [5]. Another study was based on the involvement of non-relevant documents and adjusted the score based on it. It helped to produce a greater number of relevant documents [4]. The reliability of an information system's performance based on evaluation measures by considering a large topic set has shown that an increase in the topic size when having fewer relevant documents for each topic helps to construct a good test collection [5].

Topics are queries, which retrieve relevant documents based on each topic in the system. Many studies have been done on the involvement of topics in the retrieval process. An increase in the number of topics and a reduction in the number of judgment lists shows a better result in the evaluation process and through that can reduce the human accessor effort [6].Another study by reduced the topic size and increased the evaluation depth also showed an increase in the relevant documents [7]. Another study shows that considering a huge topic or large judgment list would be a waste of budget. So, a solution, considering only a subset of topics also produces the same number of relevance judgments [8].

Relevance judgements are the documents retrieved based on topics and these documents contain the information about relevancy of each document to a particular topic. The information retrieval system performance is measured based on how many relevant documents are retrieved based on the topic [7]. The effectiveness of the system is measured based on how well a system can find the relevant documents [5].

The Information Retrieval process runs as follows. Each participating systems collect a set of relevant documents from the document corpus based on the topics. These documents were ranked based on their relevance and call it as runs. By using some rank aggregation techniques these ranked documents were merged and ready for the judgements. However, judging the whole document is practically not possible as it is costly and time-consuming. Therefore, the evaluation initiatives have proposed some methods to retrieve the most relevant documents that will be sent to the human assessors for judgment. After these judgments, we can find out which systems performed better compared to the other systems. This process can be done through any evaluation measures like precision, and mean average precision. Through these evaluation scores, we can judge which all systems performed better and can rank these documents [9].The information retrieval evaluation flow is shown in Figure 1.

The evaluation performance of the Information retrieval systems is not only by considering their efficiency but also through their effectiveness, that is their ability to produce as much of relevant documents and rank them in a better way by rejecting the irrelevant ones.

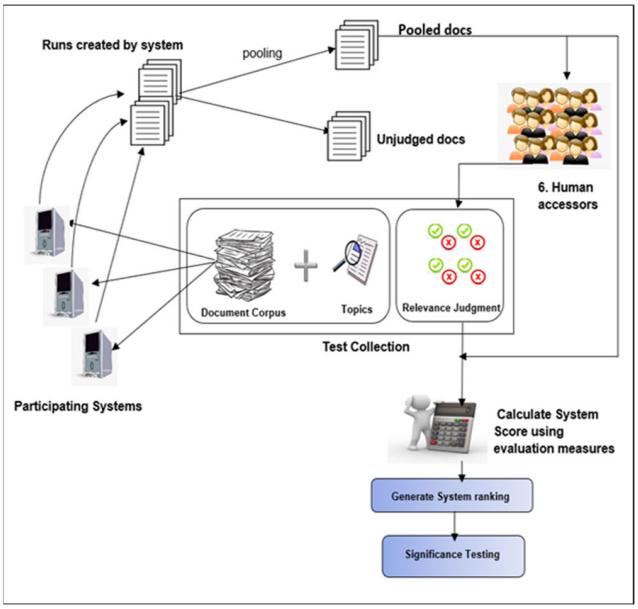

FIGURE 1. Information retrieval evaluation process.

As the number of relevant documents increases in the judgment list, the quality of the list will increase and through that, the evaluation accuracy also increases. Various methods are there to retrieve the relevant documents from the document corpus. These methodologies help to produce as much of relevant documents into the judgment list. The various methodologies categorization is shown in Figure 2. This literature review offers an depth study of various methodologies that are available in the evaluation process used by the researchers to get better results.

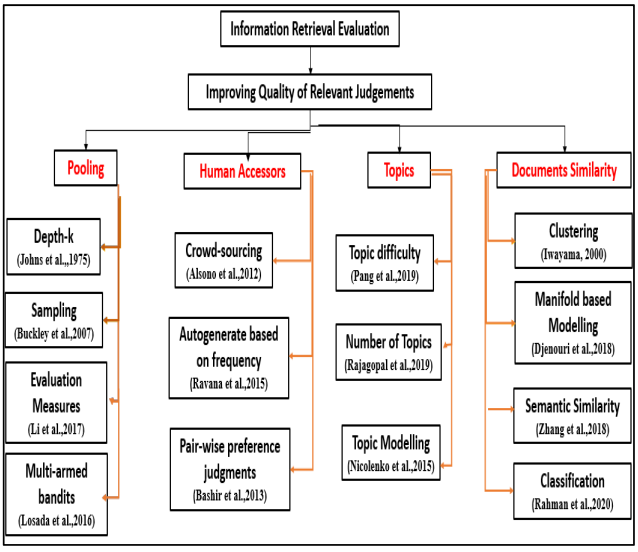

FIGURE 2. Categories of various methodologies used to generate relevant judgment set in the evaluation process.

The main contribution of this paper is

- The detailed study of various methodologies that are available in the evaluation process in order to increase the number of relevant documents in the judgment set also known as qrels.
- Comparison and discussion on the benefits and drawbacks of these methodologies during the evaluation process.

II. METHODOLOGIES

This session provides a detailed view of various methodologies used by the researchers to improve the number of relevant documents in the judgment sets and through that increase the quality of the judgment sets.

A. IMPROVING RELEVANT JUDGMENTS BASED ON POOLING

Based on the topic, finding the relevant documents from large data collection is time-consuming and expensive as it has been done by human accessors who are experts in the field of those areas. In TREC dataset, which is an initiative done by the NIST organization has provided large data collection in order to perform large-scale evaluation on retrieval systems. Each document collection is in millions and billions in numbers. Some sample TREC data sets with several documents and topics are shown in Table 1. Judging all these documents by a certain number of expert judges is impossible as it takes decades to complete it and also needs to afford high cost too [10]. Lately, instead of highly paid expert judges another alternative option was crowd-sourcing. Real users were chosen with the goal of collecting relevant judgments from the crowd-sourcing platforms [11]. However, it noticed that it was more error-prone as the real users might not be judging the documents correctly.

TABLE 1. Number of documents and topics in TREC dataset.

Experiment	No. of documents	No. of topics
TREC-3 (ad hoc track)	741,856	50
TREC-8(adhoc track	528, 155	50
TREC-8(web track)	250, 000	50
TREC-10 (web track)	1,692,096	50
TREC-2004(Robust track)	528,000	250
TREC-2004(Web Track)	25, 000, 000	50
TREC-2009(web track)	1,040,809,705	50
TREC-12 (robust retrieval)	528,155	50

A new pooling method proposed with a set of documents to be judged for a topic is constructed by taking the top k documents from the multiple runs which created by the different systems [12]. Each document in the pooled list is considered as relevant and documents not in the pooled list are assumed as irrelevant. The quality of the resulting collection depends on the retrieval methods and the pool depth [14].

The first pooling method, now referred to as Depth@ k which considers the top k relevant documents for each topic from the multiple runs created by the participated systems. All the duplicated documents were removed from the list and will be given it to the human accessors for the judgment purpose. It was the most popular and traditional retrieval approach which helped to reduce the size of the judgment list [12]. The judged list will be called a partial relevance judgment list as it considers only a part of the documents for the judgment process. The remaining documents will be considered as unjudged lists or irrelevant documents.

Traditional pooling methodology helps to extract a greater number of relevant documents, but the pool depth cannot be fixed to a particular size. Pooling done with a fixed pool depth fails to produce enough number of relevant documents as the document size increases. As the document size increases and pool depth also increases, a greater number of documents might be needed to consider for the judgment, as a result again the human accessor effort, cost, and time also gets increase. Table 2 shows the various document collections and different pool depths to achieve a certain number of relevant documents. To reduce the cost and effort, the number of judgments need to be reduced. Given a multiple ranking of documents, the extraction of documents is restricted to top-k documents. Taking a sample of 10% of the depth-k and will consider those documents as relevant [14]. By this methodology, the number of relevant documents can be reduced to a certain number, but still achieve a relevant number of documents.

TABLE 2. Variation in the number of documents, topics, and pool depth in the TREC dataset to achieve a certain number of relevant documents.

Experiment	No:of docs	No:of topics	Average pool size	% of Relevant docs
TREC-3(adhoc track)	741,856	50	2814.5	4.1
TREC-8(adhoc track)	528,155	50	2508.3	5.4
TREC-8(web track)	250,000	50	950.1	4.8
TREC-10(web track)	1,692,096	50	2787.2	4.62
TREC-2004(Robust track)	528,000	250	2466	5.3
TREC-2004(Web Track)	25,000,000	50	1189.1	18.3
TREC-2009(web track)	1,040,809,705	50	4887.23	23.34
TREC-12(robust retrieval)	528,155	50	2433.5	4.6

Another method of pooling is based on IR Evaluation measures in order to solve the large-scale retrieval evaluation

using a methodology called Active sampling. A sampling strategy is used to find out runs that provide larger probabilities of relevant documents and assigns document ranking at the beginning of the sampling process. Later document samples will be retrieved from these runs and will find out actively present good systems based on evaluation measures like the Horvitz-Thompson estimator to estimate the evaluation metric of all runs [15]. Another pooling strategy is called dynamic pooling which repeatedly selects the documents from the unjudged list based on the previously judged list until enough relevant documents are achieved. It's a different concept compared to fixed pooled depth. More number of relevant documents were able to be retrieved. Some samples of fixed pool depth are depth-k, meta-ranking, and statistical sampling. Dynamic sampling examples are MFT, hedge, and bandit methods [16].

Applying a Fairness Score, namely fair pooling, is another methodology that creates pooled documents by applying a Fairness score as similar as possible for all the participating systems runs and opportunistic pooling, based on several judgments and a threshold value which is a series of evaluations metric to create a fairness pooling [11].

Identifying relevant documents based on fixed size N and fixed budget based on top@N documents are RBP ABased@N, RBP BBased@N, and RBP CBased@N. RBP is Rank Biased Precision which considers documents based on the document rank probability, examines documents in turn, and proceeds from each to the next as like compared to i+1, users prefer i^{th} document. Three methods such as

Method A: Summing Contributions, RBP ABased@N, which considers documents to be selected based on their overall contribution to the effectiveness evaluation, rather than their peak contribution.

Method B: Weighting by residual, RBP BBased@N, which considers the overall contribution of the evaluation and also weighting of the individual documents in order to avoid leaving the individual runs.

Method C: Raising the power, RBP CBased@N, which tried to increase the score component by increasing the power of the current score [17].

Based on common evaluation measures, three strategies like Take@N, from the ranked list, and top N documents were selected from each Rp run [18]. DCG Based@N applied the discount function defined in the discounted cumulative gain to rank documents into the pool. RRF Based@N, mainly used for finding the system effectiveness, is used to find the document contribution score. PP Based@N is used to calculate the number of relevant documents at rank k to the number of documents in k.

The next one is the contribution of Mutli-armed Bandits for ordering the documents in the pooling. Using this method, were able to early identify relevant documents in the pools. The document adjudication in the pooling method has been introduced [19]. This strategy has been used to analyze the documents in the pooled list and add on more documents to the relevant judgment list with minimal effort. A k-armed bandit is an approach introduced in order to adjudicate metasearch documents [20].

Shallow pooling based on preference judgment made by the crowdsourcing makes relevant judgments between neural ranked stack based on mean reciprocal rank and top judged documents, based as top-ranked documents, and re-evaluate the runs in order to produce enough documents into the judged list it helps to score more documents [50].

B. IMPROVING RELEVANT JUDGMENTS BASED ON HUMAN ACCESSORS

Getting human assessors' help for different relevant evaluations has gained a greater impact on the Information retrieval evaluation process. But still collecting relevant documents from a large collection of datasets with the help of human assessors is not feasible and recreating or reproducing these documents for every occurrence is not practically possible due to different decisions made each time by the different accessors or the same assessors. Disagreement among the assessors has been noticed as one of the major issues from the earlier stages of the information retrieval evaluation process since the 1960s [21]. Also, the cost of judgments each time with these human assessors is too high. Many studies have been done by researchers in order to reduce this cost by considering pooled documents [6], [22] and also reducing the number of topics accessed [7].

- **Crowdsourcing:**

One of the alternative solutions to reduce the cost of human assessors is through crowdsourcing. Crowdsourcing can replace the classical human assessors. The main advantages of this approach are cost-effectiveness, flexibility, and quality also improved. The literature shows that the agreement between each worker and the TREC assessors is not high if they work individually, but it increases when they are grouped and work [21]. Even some experiments show that the results of the crowdsource are accurate or precise with the expert judges. Sometimes even with the relevant documents disagreements happen between real users and the expert judges. Even the results show that in most of the TREC dataset's judgment process, the judgments have been done faster with good results at low cost [21].

Assessors' agreement and disagreement based on a topic and how it affects the evaluation process is always a challenge among researchers. Document ambiguity or topic ambiguity can be the reasons for the disagreements. Documents might have different meanings based on the terms within it have multiple meanings, the information with the document or query might not be clear, assessor's environment and mood all matter for the disagreements among the assessors to choose a document wrongly. Compared to the previous literature which assigns relevance labels for the documents, this work suggests the evaluation metric for the topic-document pair. Crowdsourcing was the option chosen for this methodology and the judgment for the same topic-document pair has been collected from the multiple assessors and the results have

shown that judgment quality has increased compared to the previous ones. And shows that defining relevancy here is based on the distribution of documents among the accessors and not based on the absolute value assigned to the documents [27].

The agreement of crowdsourced judgments with expert judgment has been studied based on different ordinal scales and different datasets based on system effectiveness and topics. Each scale's results show a similar score of agreement with the ground truth and also show almost accurate results for each topic level based on this scale. High correlation values are shown for both systems ranking and on easy topics. Considering these scales helps to be aware of different relevance levels of judgments [51].

- **Based on frequency**

But crowdsourcing with a large set of documents always can be a challenge. There can be more chances of errors in the judgment process due to disagreements and with issues or errors in indexing, searching, and even in the process of creating catalogs. The same word with different meanings might cause the assessors to judge a document incorrectly and it leads to a reduction in the number of relevant documents retrieved. Different words with the same meanings also might lead to the choice of irrelevant documents as relevant [24]. To reduce the human accessor effort, generating a document ranking methodology called the pseudo relevance judgment process has been introduced in order to generate a reliable set of relevant documents. This methodology has considered two key factors such as the frequency of each document per topic from all the system runs and document ranking. Also compared to the traditional pooling method, only the contributed systems were considered. However, in this methodology, considered documents from both contributed systems and non-contributed systems [25].

To reduce the human assessor errors in the judgment process, the magnitude estimation technique was another alternative solution. Instead of classical binary relevance judgment an estimation task has been assigned to the crowdsource to obtain judgements at scale. An estimation task that helps to check the consistency of the rank assigned among some documents mainly for the topic understandability based on the frequency of terms in each topic. The result has shown that the magnitude estimation technique is more robust in order to check the document relevancy [27].

Evaluation of system effectiveness based on different existing methods by real-user judgments has shown that it is more error-prone and varies in results by expert judgments. A study on existing methods to improve the results has shown that instead of applying a single method, a combination of different best methods results more effective by applying machine learning algorithms and finding out the frequencies of the matching results from these methodologies help to find the performance of the system runs been without relevance judgments [32].

Judgments based on assessors or groups of assessors vary based on the quality of the topics or topic terms. Research has been done on based on the quality of these topics and if irrelevant, then remove them from the test collection. This research evaluated the system performance by considering some set of search terms and a set of documents. If the quality of these terms goes below a certain value, it is considered it ambiguous and removed from the test collection and increases the quality of the retrieved documents by the human assessors [53].

- **Pair-wise preference judgments**

Assigning ranks to the relevant documents based on their relevancy is a necessary process. It is done to consider that one document is more relevant than another document and to create multiple grades of relevance. It can be done either through pair-wise preference judgment or through the nominal graded method. Both this process requires assessors help to judge the documents. Pair-wise preference judgment is more popular and acceptable as the assessors can make a binary decision as either relevant or irrelevant instead of assigning multiple relevance grades. Assessors can quickly make pair-wise judgments instead of absolute judgments. This methodology uses use Elo rating system to combine the documents [26]

To reduce the human assessor's involvement and find the system's effectiveness, another methodology was proposed to find a fixed number of relevant document pairs that are accurate and help to auto-generate other document pairs. These methods help to simulate a large number of preference judgments based on pointwise judgments [54].

Differences between the rankers can be found based on the top-ranked results by considering partial preference judgments by taking top-k ranks of documents helps to achieve more quality of documents and through that can increase the system effectiveness [55].

C. IMPROVING RELEVANT JUDGMENTS BASED ON TOPICS

Topics have a high influence on evaluating the system's performance. A different set of topics produces different results in documents. Some set of topics predicts more accurate results compared to other documents. Finding out the best topics and which topics to consider in the judgment list is always a challenge [30].

- **Topic difficulty**

In a typical information retrieval evaluation, the relevant documents are retrieved based on the relevance of the query or topic from the document corpus collection [29]. Topic is one of the major factors to predict the relevancy of a document. The topic difficulty is a major challenge in the evaluation process. Based on accessibility, topics can be classified as hard topics, medium topics, and easy topics. Human assessors always prefer to access easy topics compared to harder topics. Due to that many relevant documents related to harder topics have not been considered in the judgment list and it reduces the system's effectiveness. Topics have a high influence on

doing comparisons with the system's performance. A different set of topics with the same size produces different results and the same set of topics with different sizes also produces different results [30].

Topics can be easy, medium, or hard based on their accessibility and it will affect the system performance score. Topics difficulty is always judged by the human assessors. Easy topics are always accessed by the human assessors and due to that some of the documents that are relevant based on a harder topic will not be considered in the judgement list. Based on the topic average precision can judge a topic's difficulty. If the average precision of a system on a particular topic is high means that the particular topic is performed well on the system and that topic is an easy topic. In the same way, if the average precision value on a system for a particular topic is fewer means, the topic is harder for that system [35].

The topic difficulty and the topic size have a great impact on the system evaluation score. It will be difficult to judge topic combinations or topic pairs with a large topic-size document list and it will lead to time-consuming and high computational costs. Assigning a Topic Difficulty Score to each topic is an effective methodology to find out the difficulty of a topic and finding a suitable set of topic lists that perform better on the particular system to adjust the topic size. Usually, easy topics perform better to effectively increase the system score [29]. As an alternative option to this, if we consider top-k documents from the run list, there is no evidence that easy topics are always performing well, even the harder topics can perform better results. Also based on the different evaluation metrics the results will be consistent [36].

The relevancy of the documents can be predicted only with the returned search results. Each query can have ambiguous meanings which might lead to selecting the irrelevant documents into the qrels files. The quality of the queries can be measured with a set of queries and a set of web snippets documents based on a selection of criteria by incorporating classification and clustering techniques. Correlation coefficient values show that some sets of queries were not performing better by retrieving negative recall values. Those queries might need to be avoided for the evaluation process [53]. With this even without human relevance judgments also evaluation can be done [52].

System effectiveness in predicting relevant documents tightly depends on the topic's hardness. The topic difficulty is not invariant and depends on the participant systems the same topics produce different outputs. Topic ordering based on relevancy needs to be carefully done to evaluate the prediction modeling [33]. The topic difficulty has been tested with different corpus sets and executed on the same systems from different datasets. The same set of topics have been used. Each system has retrieved a different set of documents in each run even though the effect is less [56].

Estimation of the difficulty of a topic has been proposed in [57]. Using NDCG measures, assign continuous hardness scores to the topics based on the system performance. The same measure can be used to select a particular set of topics help to produce high-recall documents. The topic or query formulation is an effective solution to overcome these drawbacks [33].

- **Number of topics considered**

The evaluation of information retrieval systems is always a challenge due to the increased information on the web. Systems can be measured based on main two factors, such as the quality of topics chosen, and the number of relevant judgments produced. A lot of studies happening among the researchers regarding how to increase system effectiveness either by decreasing the number of topics or with better topics [7]. More number of topics chosen might increase the number of relevant documents, but the computational cost and human effort can be higher. So, most of the research on topic-based is with fewer numbers and easy topics. With a lesser topic also it is possible to achieve the same effectiveness score. All topics might not be able to produce good system rankings [6], [30]. So, the researchers need to find out the difficulty of the topic to access and the topic size. Topic size is always challenging, as the topic size increases it increases the computational cost. So, finding out the exact number of topics is with one of the effectiveness metrics such as Precision. Precision@ 10 was the earlier standard measure. But if the document size is increased, the k value also varies between 20 and 25, and the standard one is fixed at 50. So, the evaluation measure will vary depending on the document size [34].

Accessing a large number of topics has a higher computational cost, as a result, to get better results, various studies happening on varied evaluation depth with a reduced number of topics. Based on effort-based relevance judgment, higher evaluation depth with a lesser topic size and lesser evaluation depth with a higher topic size have been considered. Achieving good evaluation outcomes based on reduced topics or higher topics is an interesting part among the researchers. Number of topics to be used is based on the user's satisfaction. In research, it has been noted that there is a non-correlation between evaluation metrics and user satisfaction. Users always prefer low-effort topics for easy access. Real users won't spend too much time like expert judges with the hard topics. Due to that many relevant documents won't get considered in the pooled list and it will affect the system evaluation score. Considering low-effort judgment or easy topics with various evaluation depths is one choice to make the evaluation metric standardized [7].

Another study was based on the topic easiness, the real user's ability to understand the document, finding out the correct topic related to the document, the way how convey the meaning of the topic title and document content all matters with the correct retrieval of documents. The effort on the understandability, findability, and readability always has an impact on the system performance evaluation score. High-effort topics and documents are always harder for real users to make a correct judgment and are neglected by the users due to their difficulty. Real users always prefer low-effort

documents for their ease of access. So, considering low-effort documents with shallow depth always gives better results in terms of evaluation metrics [7].

Partitioning the document corpus randomly into multiple parts(shard) and creating a replication of system and topic combinations helps to find out the pairs of the documents that are related to each system-topic pair list. It helps to reduce the topic size for the document list and still maintain the accuracy of the evaluation process [47]. Topic with document subparts(shards) combination also studied and it has shown that topic-shard combination produces different output with different systems and needs to be carefully chosen when the topic shard set is being considered. The topic alone and the topic-shard combination produce different system performances and through that accuracy also varies [58]. Topic sampling is crucial as it varies the results of the accuracy. Topic-aware balancing methodology helps to choose the topics and the passage efficiently and helps to maintain the result constantly even with different random ordering [59].

- **Topic Modeling**

Topic modeling is a technique that can be used to evaluate large data collection. To reduce the noise and choose the subset of the documents, topic modeling has an important role. To plot the topic modeling the best way is through matrix factorization. Topic modeling can be used for multi-lingual situations. It can used even with multi-language datasets. Generating topics based on the topic modeling concepts can be done with formal or informal datasets, multi-model, or multilingual models [60]. The accuracy of the results produced with different topic modeling techniques has been different with different datasets and different criteria. Also choosing the metrics used to evaluate the topic modeling results got quite challenging as it is difficult to predict the accuracy. Various topic modeling techniques used so far and also shown which model will be more beneficial for different content-based datasets. Also shows a clear view of the evaluation metrics used to predict the accuracy of the results [61].

To achieve a good quality evaluation metric, topic interest among real users is of great importance. Finding the correct interesting topic based on user choice will help to achieve a good system evaluation score. Researchers have proposed a model to find out the best topics based on criteria. Topic modeling has been used to evaluate the quality of the topics and it has been seen growing for various applications like text classifiers, image recognition classification, and so on. Topic modeling can be constructed with some predefined keywords. With the help of topic modeling, it can extract or mine specific relevant topics which can extract more relevant documents. And, topic modeling can extract a topic quality metric that predicts human judgment about a topic [37].

Topic coverage is an alternative option for choosing a more relevant and accurate topic for the topic list. Different topic models generate a set of topics and may not be accurate according to the expert judges. Topic coverage technique computationally matches the topics with the reference light of topics, and it helps to judge the models and also topic quality. An experiment with topic coverage in order to find out topic quality, topic categories, and also topic model evaluation [62].

D. IMPROVING RELEVANT JUDGMENTS BASED ON DOCUMENT SIMILARITIES

Traditional methods like pooling, sampling, and using evaluation metrics for large document collection are time-consuming and have higher computational costs. Clustering and classification techniques help to resolve this drawback. In traditional methodologies, the whole dataset needs to be considered for the evaluation process. However, in clustering and classification techniques, only the documents within that class or cluster need to be considered for the evaluation process. This technique helps to retrieve documents with less cost and time. However, the quality of the documents retrieved through these techniques is lesser compared to the traditional methodology. So to increase the quality of the evaluation process, much research has been done based on clustering and classification techniques.

- **Clustering**

In the last decades, several approaches have been proposed to increase the performance of cluster-based information retrieval processes. One among them is by combining both clustering and frequent itemset mining. Each document from the ranked list was clustered based on the K-means algorithm and according to its relevance, the similar documents will be grouped into each cluster. From each cluster the frequent terms in the documents were calculated and based on this frequency of terms, closed frequent pairs were collected. To find the relevant documents based on the user query is through these frequent pairs of itemsets. Based on the user query, the k most frequent patterns are used to group documents that share similar terms. Patterns or the most frequently occurred itemsets discovered in each cluster are then used to select the most relevant document clusters [38].

The search effectiveness can be achieved through both incremental relevance feedback and through document clustering. In the relevance-based feedback approach, the quality of the documents retrieved is usually lesser. Users' relevant judgments are collected from one or two systems rather than pooling and these documents are considered initial judgments these documents were returned to the systems to produce more documents and based on that feedback the relevancy is determined. The document clustering method displays the retrieved documents in a cluster form rather than just ranked ones. This helps to separate the relevant and non-relevant documents and easy to judge the relevant documents cluster documents efficiently and can reduce the focus on a single topic in a user query and ignore others. In this methodology, all the documents are clustered based on user feedback. Choosing the best cluster based on the density strategy. Within each cluster, the documents are sorted by their relevance score for the initial query. And extracted the top N documents from each cluster [39].

Clustering the documents based on a user query is a common approach for the retrieval evaluation process. The clustered documents were later considered for the document ranking. Different clusters are simultaneously used for retrieving common features of the documents based on their ranking and assigning document ranking based on these features was proposed in [64]. This approach helps to improve the document's similarity vector space.

To improve the effectiveness of the retrieval effect in clustering, considering topic modeling with clustering helps to improve the result. Each topic in the cluster is mapped with a set of terms in the document collection and find out the frequencies of each term. The similar topics are frequently considered to represent various themes. This result helps to retrieve meaningful representations of clusters and also helps as a guide to judge the quality of the clusters [65].

- **Manifold-based**

Another model is Manifold-based and, in this model, inter-document similarity will be considered to group the documents and assign new scores to these similar documents and these documents with similar degrees of relevance within that manifold are considered to be the same and will assign a similar score and will consider into the judged list [40].

Passage-based and manifold-based similarity-checking methods help to improve the ranking score of the initial pooled documents. Based on the passage-based model, term frequencies in the passage were considered and assigned based on the weightage of the documents. The manifold-based model helps to evaluate inter-document term frequencies using term modalities. This method helps to refine the scores in the initial pool of documents and re-rank those documents based on their scores [66].

To improve the efficiency of the clustering task, a rank-based manifold method has been proposed in [67]. This method helps to create different clusters based on the similarity measure. The documents within the cluster have gone through an unsupervised similarity learning technique to compute the effective measure in a data collection manifold.

- **Classification**

Even without pooling and system ranking, automatically classifying the unjudged documents based on their similarity is another method. Usually, unjudged documents are not considered in the pooled list. This approach considers a topic-specific document classification model for each search topic. This approach uses using Active learning algorithm for selecting and classifying documents. First, with the help of an active learning algorithm, select the documents that accessors should judge. Next, these documents are used to classify all the unjudged documents into separate classes based on their similarity. Active Learning algorithms are used for both these document selections and the automatic labeling of unjudged documents. Comparisons and similarity checking with the pooled documents and the unjudged documents based on the classifications improve the relevant scores. However, predicting relevance judgment via a classifier introduces bias in the evaluation of Information Retrieval systems when considering the document ranking. Also have introduced a hybrid combination of human and automatic judgments [41].

In many evaluation processes, only the pooled documents were selected into the pooled list. The documents from the unjudged list were not considered for the evaluation process and due to that the system's effectiveness is not accurate. Another methodology was proposed to increase the system's effectiveness by training a classifier on the pooled documents and considering it for predicting the relevancy of the documents from the unjudged list by considering the document similarity. Once the similarity was found, those documents were moved to the judged list or the pooled list and tried to increase the effectiveness of the systems [42].

A similarity measure by incorporating both term frequencies and sparse data in various dimensions helps to increase the performance score of the classified documents. This method classifies the documents based on term frequencies and uses centroids and vector space models for classifying the documents [63]. Precision values, recall values, and f1 score values were improved with these similarity measures. Classification accuracy improved in terms of text-based documents.

III. SUMMARIZE FINDINGS AND DISCUSSIONS

This study aimed to examine the various methodologies to improve the quality of documents in the judgment list, in inorder to increase the effectiveness of the system evaluation process. Pooling, Human accessors, Topics, and Document Similarity are the main methodologies used for these evaluation processes.

Retrieving relevant documents is one of the main concerns during the evaluation process. Based on how many relevant documents retrieved by each system, ranks are allocated and through that system, performance is being evaluated. To get better ranks, more relevant documents need to be on the judgment list. For that, various methodologies researchers have used and helped to increase the number of relevant documents in the judgment list.

- **Pooling**

Pooling is one of the traditional and most effective methods to retrieve a greater number of relevant documents into the judgment list. In pooling the whole document list is involved in the evaluation process and a subset is chosen for the judgment list. This process helps the pooling methodologies to retrieve a greater number of relevant documents and through that, the quality of the documents retrieved is also better. This helps to increase the system's effectiveness but as this process considers the whole document, it is time-consuming [38].

- **Document Similarity**

Finding document similarity either by clustering or classification technique considers only a part of documents from the class or cluster for the evaluation process. So, the quality of the documents retrieved from these judgment lists is less compared to the traditional methodology. Much

research [38], [41], [40] is happening now to increase the quality of the relevant document list through classification and clustering techniques.

- **Human Accessors**

Judgment through human accessors is more error-prone. For a topic, judgments made by real users and expert judges are different. Depending on the readability effort, understandability effort, and findability effort, it varies [7] and the environment influences while judging. As it is more error-prone, we need to reduce the human accessors' effort by automating the frequency of documents on each topic or automating document selection as similar to the human choices

- **Topics**

Topics hardness also matters in the judgment process. Easy topics can be judged by human accessors easily and can retrieve many relevant documents. For the hard topic, judgments made by the real users and expert judges are different due to the topic hardness. Topic hardness is always a quite challenge for everyone to judge a set of documents based on a topic [29]. So, the system's effectiveness also varies depending on the number of relevant documents. Either needs to reduce the topic pool depth or topics need to auto-generate topics based on the pre-defined keywords.

Various evaluation metrics are used to measure the system's effectiveness. These measures are used to find out how well the participating systems retrieve several relevant documents. The various measures are Precision, Recall, Average Precision, f-measure, RBP and NDCG, etc. Precision is based on the number of documents retrieved and Recall is based on several relevant documents in the judged list [5], [43]. RBP is rank-biased Precision which is an alternative to Recall which

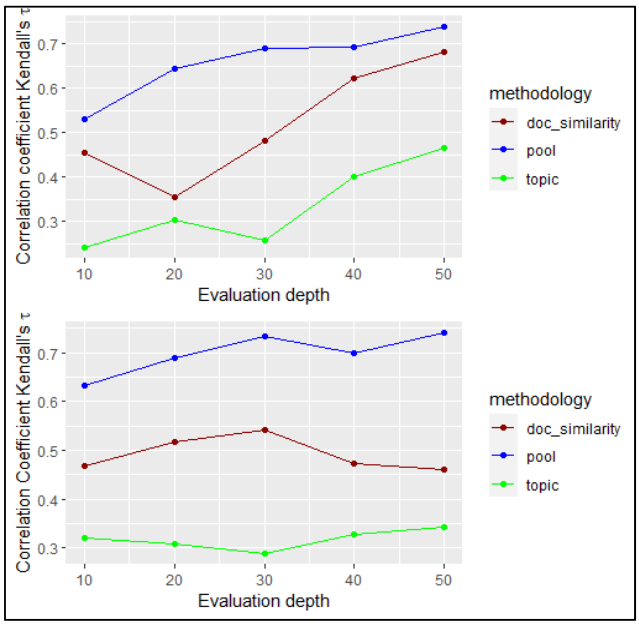

FIGURE 3. Kendall's correlation values of methodologies for various predictions of AP@10 based on TREC-8 and TREC-9.

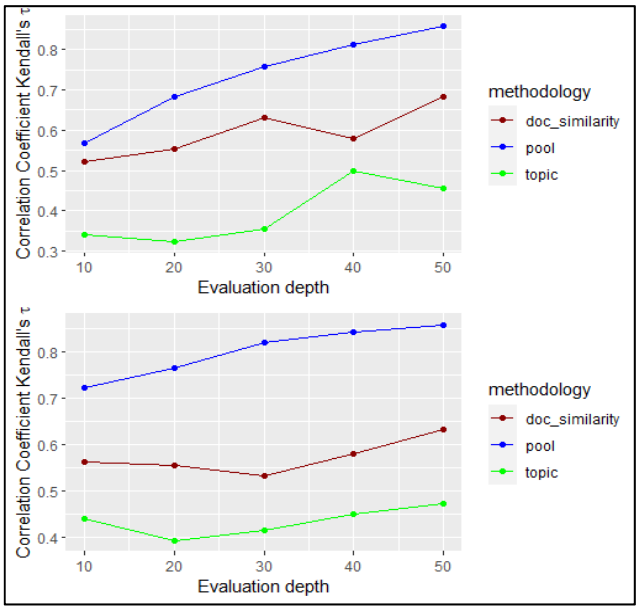

FIGURE 4. Kendall's correlation values of methodologies for various predictions of AP@100 based based on TREC-8 and TREC-9.

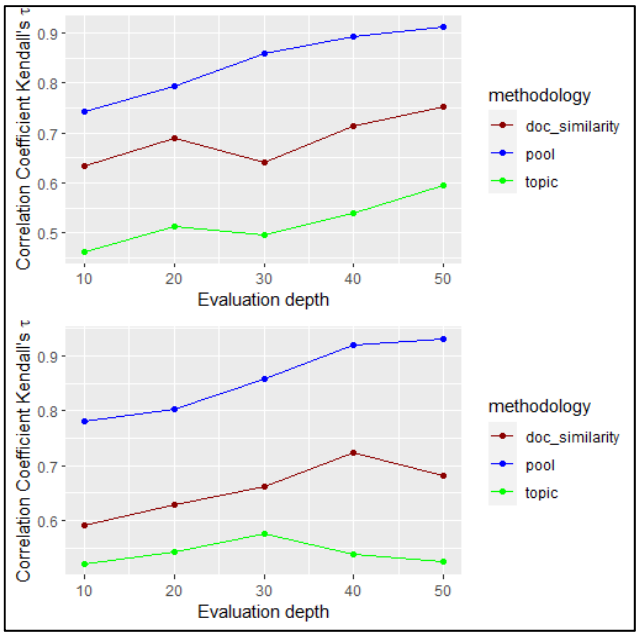

FIGURE 5. Kendall's correlation values of methodologies for various predictions of AP@1000 based on TREC-8 and TREC-9.

assigns relevance weights based on geometrical distribution based on documents that appeared in the ranking [44], [45], [46]. NDCG is for ranking the quality of the retrieved document set by the participating systems [48], [49]. Sometimes these evaluation metrics overestimate the effectiveness of the systems, especially when some of the top-ranked documents are unjudged. The correlation of metrics shows that precision evaluation metrics perform better [31].

Figure 3, Figure 4, and Figure 5 show the Kendall Tau correlation of methodologies such as pooling, by considering

topics and document similarity. Based on the TREC 8 dataset, the total runs are 129, and the selected runs are 108. Total topics are 50. TREC 9 has been used. Pooling has been done with Combsum and a pool depth of 100. The topic is with the frequency of terms with topic size 40. Document similarity done with K-means clustering and TF-IDF algorithm used for similarity checking and Cosine similarity for distance measuring. Figure 3, Figure 4, and Figure 5 presents Kendall's correlations of all methodologies for various prediction of precision@10,precision@100 and precision@ 1000. It has been computed with a cut-off depth of 10 to 50. Here it shows that for both trec datatsets, pooling performs better in finding out the relevant judgements compared to document similarity and topic based. Topic-based methodology did not retrieve better results due to the topic's hardness.

A brief issue and research direction on the various methodologies are listed in Table 3.

TABLE 3. Summarize findings from literature review.

Methodologies	Advantages	Gaps
Pooling	Traditional methodology and dominated by studies. The quality of documents is always high	As the whole document set is considered for the judgment process, time-consuming and high-cost.
Human accessors	more error-prone and different human agreements	need to reduce the human accessors' effort by automating the frequency of documents on each topic
Topics	difficulty in calculating topic hardness and quality	reduce topic pool depth and consider pre-defined keywords
Document similarity	Only a part of the document list is considered. Less cost and reduced time	The quality of the document retrieved is less.

IV. CONCLUSION

Increasing the number of relevant documents in the judgment set also known as qrels always helps to improve the quality of the judgment list and through that can increase the accuracy of the evaluation process. In this work, we have summarized various methodologies like pooling, through human accessors, topic selection, and document similarity checking, used to improve the quality of relevant documents. Much research has been done in each methodology in order to provide a greater number of relevant documents into the judgment set. Pooling is one of the traditional approaches and is more popular, it provides a tremendous result in choosing relevant documents, but it time time-consuming as it is considering the complete dataset for the evaluation process. Document similarity checking is another methodology that solves the drawbacks of the pooling method as it considers only one cluster or a class of documents for the evaluation process. As it is considered only a part of the judgment list, the quality of the document is quite low. Human accessor evaluation like crowdsourcing is highly expensive as it requires expert judges or real users for the evaluation process, and it is more error-prone. Topic selection is a quite challenging process too and judgment will vary depending on topic hardness and topic difficulty. Various research can be done by considering the drawbacks of each methodology to improve the number of relevant documents in the judgment list.

REFERENCES

[1] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Proc. Workshop Cross-Language Eval. Forum Eur. Lang.* Berlin, Germany: Springer, Sep. 2001, pp. 355–370, doi: 10.1007/3-540-45691-0_34.

[2] F. Scholer, D. Kelly, and B. Carterette, "Information retrieval evaluation using test collections," *Inf. Retr. J.*, vol. 19, no. 3, pp. 225–229, Jun. 2016, doi: 10.1007/s10791-016-9281-7.

[3] P. Clough and M. Sanderson, "Evaluating the performance of information retrieval systems using test collections," *Inf. Res. Int. Electron. J.*, vol. 18, no. 2, Jun. 2013. [Online]. Available: https://researchmgt.monash.edu/ws/portalfiles/portal/398528720/375342801_oa.IEEE

[4] W. E. Webber, "Measurement in information retrieval evaluation," Ph.D. dissertation, Dept. Comput. Sci. Softw. Eng., Univ. Melbourne, Melbourne, VIC, Australia, 2010. [Online]. Available: http://hdl.handle.net/11343/35779

[5] M. Sanderson and J. Zobel, "Information retrieval system evaluation: Effort, sensitivity, and reliability," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2005, pp. 162–169, doi: 10.1145/1076034.1076064.

[6] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2008, pp. 651–658, doi: 10.1145/1390334.1390445.

[7] P. Rajagopal and S. D. Ravana, "Effort-based information retrieval evaluation with varied evaluation depth and topic sizes," in *Proc. 3rd Int. Conf. Bus. Inf. Manage.*, Sep. 2019, pp. 143–147, doi: 10.1145/3361785.3361794.

[8] M. Hosseini, I. J. Cox, N. Milic-Frayling, V. Vinay, and T. Sweeting, "Selecting a subset of queries for the acquisition of further relevant judgments," in *Proc. Conf. Theory Inf. Retr.* Berlin, Germany: Springer, Sep. 2011, pp. 113–124, doi: 10.1007/978-3-642-23318-0_12.

[9] D. E. Losada, J. Parapar, and A. Barreiro, "A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation," *Inf. Fusion*, vol. 39, pp. 56–71, Jan. 2018, doi: 10.1016/j.inffus.2017.04.001.

[10] S. I. Moghadasi, S. D. Ravana, and S. N. Raman, "Low-cost evaluation techniques for information retrieval systems: A review," *J. Informetrics*, vol. 7, no. 2, pp. 301–312, Apr. 2013, doi: 10.1016/j.joi.2012.12.001.

[11] A. Tonon, G. Demartini, and P. Cudré-Mauroux, "Pooling-based continuous evaluation of information retrieval systems," *Inf. Retr. J.*, vol. 18, no. 5, pp. 445–472, Oct. 2015, doi: 10.1007/s10791-015-9266-y.

[12] K. Sparck Jones and C. J. Van Rijsbergen, "Information retrieval test collections," *J. Documentation*, vol. 32, no. 1, pp. 59–75, Jan. 1976, doi: 10.1108/eb026616.

[13] K. Spark-Jones, "Report on the need for and provision of an 'ideal' information retrieval test collection," Comput. Lab., Tech. Rep., 1975, doi: 10.1007/1-4020-3467-9_7.

[14] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling for large collections," *Inf. Retr.*, vol. 10, no. 6, pp. 491–508, Dec. 2007.

[15] D. Li and E. Kanoulas, "Active sampling for large-scale information retrieval evaluation," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 49–58, doi: 10.1145/3132847.3133015.

[16] G. V. Cormack and M. R. Grossman, "Beyond pooling," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 1169–1172, doi: 10.1145/3209978.3210119.

[17] A. Moffat, W. Webber, and J. Zobel, "Strategic system comparisons via targeted relevance judgments," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 375–382, doi: 10.1145/1277741.1277806.

[18] A. Lipani, D. E. Losada, G. Zuccon, and M. Lupu, "Fixed-cost pooling strategies," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1503–1522, Apr. 2021, doi: 10.1109/TKDE.2019.2947049. https://doi.org/10.1109/TKDE.2019.2947049.

[19] D. E. Losada, J. Parapar, and Á. Barreiro, "Feeling lucky? Multi-armed bandits for ordering judgements in pooling-based evaluation," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, Apr. 2016, pp. 1027–1034, doi: 10.1145/2851613.2851692.

[20] D. E. Losada, J. Parapar, and A. Barreiro, "Cost-effective construction of information retrieval test collections," in *Proc. 5th Spanish Conf. Inf. Retr.*, Jun. 2018, pp. 1–2, doi: 10.1145/3230599.3230612.

[21] O. Alonso and S. Mizzaro, "Using crowdsourcing for TREC relevance assessment," *Inf. Process. Manage.*, vol. 48, no. 6, pp. 1053–1066, Nov. 2012, doi: 10.1016/j.ipm.2012.01.004.

[22] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1998, pp. 282–289, doi: 10.1145/290941.291009.

[23] P. Rajagopal, S. D. Ravana, and M. A. Ismail, "Relevance judgments exclusive of human assessors in large scale information retrieval evaluation experimentation," *Malaysian J. Comput. Sci.*, vol. 27, no. 2, pp. 80–94, 2014. [Online]. Available: https://ejournal.um.edu.my/index.php/MJCS/article/view/6795

[24] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1–50, Jan. 2012, doi: 10.1145/2071389.2071390.

[25] S. D. Ravana, P. Rajagopal, and V. Balakrishnan, "Ranking retrieval systems using pseudo relevance judgments," *Aslib J. Inf. Manage.*, vol. 67, no. 6, pp. 700–714, Nov. 2015.

[26] M. Bashir, J. Anderton, J. Wu, P. B. Golbus, V. Pavlu, and J. A. Aslam, "A document rating system for preference judgements," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 909–912, doi: 10.1145/2484028.2484170.

[27] E. Maddalena, K. Roitero, G. Demartini, and S. Mizzaro, "Considering assessor agreement in IR evaluation," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Oct. 2017, pp. 75–82, doi: 10.1145/3121050.3121060.

[28] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin, "On crowdsourcing relevance magnitudes for information retrieval evaluation," *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 1–32, Jul. 2017, doi: 10.1145/3002172.

[29] W. T. Pang, P. Rajagopal, M. Wang, S. Zhang, and S. D. Ravana, "Exploring topic difficulty in information retrieval systems evaluation," *J. Phys., Conf. Ser.*, vol. 1339, no. 1, Dec. 2019, Art. no. 012019, doi: 10.1088/1742-6596/1339/1/012019.

[30] A. Berto, S. Mizzaro, and S. Robertson, "On using fewer topics in information retrieval evaluations," in *Proc. Conf. Theory Inf. Retr.*, Sep. 2013, pp. 30–37, doi: 10.1145/2499178.2499184.

[31] S. Muwanei, S. D. Ravana, W. L. Hoo, D. Kunda, P. Rajagopal, and P. S. Sodhi, "Correlation and prediction of high-cost information retrieval evaluation metrics using deep learning," *Inf. Res.*, vol. 27, no. 1, Mar. 2022. [Online]. Available: https://researchmgt.monash.edu/ws/portalfiles/portal/398528720/375342801_oa.pdf, doi: 10.47989/irpaper920.

[32] K. Roitero, J. S. Culpepper, M. Sanderson, F. Scholer, and S. Mizzaro, "Fewer topics? A million topics? Both?! On topics subsets in test collections," *Inf. Retr. J.*, vol. 23, no. 1, pp. 49–85, Feb. 2020, doi: 10.1007/s10791-019-09357-w.

[33] J. S. Culpepper, G. Faggioli, N. Ferro, and O. Kurland, "Topic difficulty: Collection and query formulation effects," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, pp. 1–36, Jan. 2022, doi: 10.1145/3470563.

[34] B. T. Dinçer, "Design of information retrieval experiments: The sufficient topic set size for providing an adequate level of confidence," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 21, pp. 2218–2232, Jan. 2013, doi: 10.3906/elk-1203-20.

[35] S. Mizzaro, "The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation?" in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, Mar. 2008, pp. 642–646.

[36] K. Roitero, E. Maddalena, and S. Mizzaro, "Do easy topics predict effectiveness better than difficult topics?" in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, Apr. 2017, pp. 605–611, doi: 10.1007/978-3-319-56608-5_55.

[37] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *J. Inf. Sci.*, vol. 43, no. 1, pp. 88–102, Feb. 2017, doi: 10.1177/0165551515617393.

[38] Y. Djenouri, A. Belhadi, P. Fournier-Viger, and J. C.-W. Lin, "Fast and effective cluster-based information retrieval using frequent closed itemsets," *Inf. Sci.*, vol. 453, pp. 154–167, Jul. 2018, doi: 10.1016/j.ins.2018.04.008.

[39] M. Iwayama, "Relevance feedback with a small number of relevance judgements: Incremental relevance feedback vs. document clustering," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2000, pp. 10–16, doi: 10.1145/345508.345538.

[40] S. Liang, I. Markov, Z. Ren, and M. de Rijke, "Manifold learning for rank aggregation," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1735–1744, doi: 10.1145/3178876.3186085.

[41] M. M. Rahman, M. Kutlu, T. Elsayed, and M. Lease, "Efficient test collection construction via active learning," in *Proc. ACM SIGIR Int. Conf. Theory Inf. Retr.*, Sep. 2020, pp. 177–184, doi: 10.1145/3409256.3409837.

[42] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff, "Reliable information retrieval evaluation with incomplete and biased judgements," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 63–70, doi: 10.1145/1277741.1277755.

[43] M. Arora, U. Kanjilal, and D. Varshney, "Evaluation of information retrieval: Precision and recall," *Int. J. Indian Culture Bus. Manage.*, vol. 12, no. 2, pp. 224–236, 2016.

[44] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Trans. Inf. Syst.*, vol. 27, no. 1, pp. 1–27, Dec. 2008, doi: 10.1145/1416950.1416952.

[45] Y. Zhang, L. A. Park, and A. Moffat, "Parameter sensitivity in rank-biased precision," in *Proc. ADCS*, vol. 13, 2008, pp. 46–69.

[46] N. Ferro and G. Silvello, "Rank-biased precision reloaded: Reproducibility and generalization," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, Mar. 2015, pp. 768–780, doi: 10.1007/978-3-319-16354-3_83.

[47] E. M. Voorhees, D. Samarov, and I. Soboroff, "Using replicates in information retrieval evaluation," *ACM Trans. Inf. Syst.*, vol. 36, no. 2, pp. 1–21, Apr. 2018, doi: 10.1145/3086701.

[48] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells, "On the robustness and discriminative power of information retrieval metrics for top-N recommendation," in *Proc. 12th ACM Conf. Recommender Syst.*, Sep. 2018, pp. 260–268, doi: 10.1145/3240323.3240347.

[49] S. Muwanei, S. D. Ravana, W. L. Hoo, and D. Kunda, "The prediction of the high-cost non-cumulative discounted gain and precision performance metrics in information retrieval evaluation," in *Proc. 5th Int. Conf. Inf. Retr. Knowl. Manage. (CAMP)*, Jun. 2021, pp. 25–30, doi: 10.1109/CAMP51653.2021.9497989.

[50] N. Arabzadeh, A. Vtyurina, X. Yan, and C. L. A. Clarke, "Shallow pooling for sparse labels," 2021, *arXiv:2109.00062*.

[51] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer, "On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation," *Inf. Process. Manage.*, vol. 58, no. 6, Nov. 2021, Art. no. 102688, doi: 10.1016/j.ipm.2021.102688.

[52] K. Roitero, A. Brunello, G. Serra, and S. Mizzaro, "Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102149, doi: 10.1016/j.ipm.2019.102149.

[53] D. Zhu, S. L. Nimmagadda, K. W. Wong, and T. Reiners, "Relevance judgment convergence degree—A measure of inconsistency among assessors for information retrieval," 2022, *arXiv:2208.04057*.

[54] K. Roitero, A. Checco, S. Mizzaro, and G. Demartini, "Preferences on a budget: Prioritizing document pairs when crowdsourcing relevance judgments," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 319–327, doi: 10.1145/3485447.3511960.

[55] C. L. A. Clarke, A. Vtyurina, and M. D. Smucker, "Assessing top- preferences," *ACM Trans. Inf. Syst.*, vol. 39, no. 3, pp. 1–21, Jul. 2021, doi: 10.1145/3451161.

[56] F. Zampieri, K. Roitero, J. S. Culpepper, O. Kurland, and S. Mizzaro, "On topic difficulty in IR evaluation: The effect of systems, corpora, and system components," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 909–912, doi: 10.1145/3331184.3331279.

[57] L. Gienapp, B. Stein, M. Hagen, and M. Potthast, "Estimating topic difficulty using normalized discounted cumulated gain," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2033–2036, doi: 10.1145/3340531.3412109.

[58] N. Ferro, Y. Kim, and M. Sanderson, "Using collection shards to study retrieval performance effect sizes," *ACM Trans. Inf. Syst.*, vol. 37, no. 3, pp. 1–40, Jul. 2019, doi: 10.1145/3310364.

[59] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury, "Efficiently teaching an effective dense retriever with balanced topic aware sampling," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 113–122, doi: 10.1145/3404835.3462891.

[60] R. Churchill and L. Singh, "The evolution of topic modeling," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–35, Jan. 2022, doi: 10.1145/3507900.

[61] M. Rüdiger, D. Antons, A. M. Joshi, and T.-O. Salge, "Topic modeling revisited: New evidence on algorithm performance and quality metrics," *PLoS ONE*, vol. 17, no. 4, Apr. 2022, Art. no. e0266325, doi: 10.1371/journal.pone.0266325.

[62] D. Korencic, S. Ristov, J. Repar, and J. Šnajder, "A topic coverage approach to evaluation of topic models," *IEEE Access*, vol. 9, pp. 123280–123312, 2021, doi: 10.1109/ACCESS.2021.3109425. https://doi.org/10.1109/ACCESS.2021.3109425

[63] M. Eminagaoglu, "A new similarity measure for vector space models in text classification and information retrieval," *J. Inf. Sci.*, vol. 48, no. 4, pp. 463–476, Aug. 2022, doi: 10.1177/0165551520968055.

[64] E. Markovskiy, F. Raiber, S. Sabach, and O. Kurland, "From cluster ranking to document ranking," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 2137–2141, doi: 10.1145/3477495.3531819.

[65] M. Yuan, P. Lin, and J. Zobel, "Document clustering vs topic models: A case study," in *Proc. Australas. Document Comput. Symp.*, Dec. 2021, pp. 1–8, doi: 10.1145/3503516.3503527.

[66] R. Shraga, H. Roitman, G. Feigenblat, and M. Canim, "Ad hoc table retrieval using intrinsic and extrinsic similarities," in *Proc. Web Conf.*, Apr. 2020, pp. 2479–2485, doi: 10.1145/3366423.3379995.

[67] B. Rozin, V. H. Pereira-Ferrero, L. T. Lopes, and D. C. G. Pedronette, "A rank-based framework through manifold learning for improved clustering tasks," *Inf. Sci.*, vol. 580, pp. 202–220, Nov. 2021, doi: 10.1016/j.ins.2021.08.080.

MINNU HELEN JOSEPH received the master's degree in computer science and engineering from the Karunya Institute of Technology, India, in 2009. She is currently pursuing the Ph.D. degree in information retrieval evaluation with the University of Malaya, Malaysia. She is also a Lecturer with Asia Pacific University of Technology and Innovation, Malaysia. Her research interests include information retrieval evaluation, data analytics, data science, and machine learning.

SRI DEVI RAVANA received the master's degree in software engineering from the University of Malaya, Malaysia, in 2001, and the Ph.D. degree from The University of Melbourne, Australia, in 2011. Currently, she is an Associate Professor with the University of Malaya. Her research interests include information retrieval, text heuristics, IR evaluation, data analytics, data science, and search engines.