INCREASING THE ACCURACY OF INFORMATION RETRIEVAL SYSTEMS EVALUATION BY IMPROVING THE QUALITY OF THE RELEVANT JUDGEMENTS

MINNU HELEN JOSEPH

FACULTY OF COMPUTER SCIENCE AND INFORMATION
TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2024

INCREASING THE ACCURACY OF INFORMATION RETRIEVAL SYSTEMS EVALUATION BY IMPROVING THE QUALITY OF THE RELEVANT JUDGEMENTS

MINNU HELEN JOSEPH

THESIS SUBMITTED IN FULFILMENT OF THE

REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR**

2024

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Minnu Helen Joseph

Matric No: 17198581/1

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

Increasi~~ng the accuracy of the information retrieval~~ systems evaluation by improving the quality of the relevance judgments

Field of Study: Information Storage Retrieval (Computer Science)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work.
(2) This Work is original.
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work.
(4) I do not have any actual knowledge, nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work.
(5) I hereby assign all and every rights in the copyright to this Work to the Universiti Malaya (“UM”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate’s Signature Date: 19-06-24

Subscribed and solemnly declared before,

Witness’s Signature Date: 19 June 2024

Name:

Designation:

INCREASING THE ACCURACY OF INFORMATION RETRIEVAL SYSTEMS EVALUATION BY IMPROVING THE QUALITY OF THE RELEVANT JUDGEMENTS

ABSTRACT

Information retrieval evaluation is a process of measuring how well the participating systems can meet the information needed by the user. The system's performance is evaluated based on the relevance judgment set quality. The quality of the judgment set is measured based on the ability of the participating systems to retrieve as many relevant documents based on topics into the judgment sets and rank them in a better way and also, at the same time suppress the irrelevant ones. However, it has been noticed that for smaller test collections, this assumption might be correct. But for large test collections like TREC(Text Retrieval Conference), this assumption might not always be true. It has been noticed that the quality of the judgment sets is not up to the level or incomplete according to the Cranfield paradigm methodology, especially through document similarity techniques. The main aim of this thesis is to increase the quality of the relevance judgment sets during the evaluation process. The quality of the judgment sets can be increased by augmenting the number of relevant documents in the judgment sets. It will indirectly help to increase the accuracy of the evaluation process. This thesis's main contribution is to increase the quality of the judgment sets by proposing some methodologies. The first experiment explored the issues of partial relevance judgments on existing methodologies. The methodologies' inability to retrieve all the relevant documents into the relevance judgment sets is considered. By considering the limitations of the existing methodologies, a methodology has been proposed to increase the relevant documents in the judgment sets. The proposed methodology combines the pooling and document similarity using clustering and classification techniques. Documents similarity has been done between pooled and clustered or classified unjudged documents. If a similarity is found, a new score will be assigned to those documents and moved that document into the pooled list. The evaluation continues until all the documents from the pooled list are considered for the similarity-checking process. The results show that the proposed methodology can achieve a greater number of relevant documents in the judgment sets and also helps to achieve a better result with lesser pool depth. The second experiment explored how to further improve or maintain the quality of the judgment set by considering the test collection. For this experiment, topics and participating systems from test collections were considered. Based on the results, it has been proven that a smaller number of the most effective topics, or easy topics, can maintain the quality of the judgment sets. Also, based on the system contributions, an enhanced methodology has been proposed and the results show that it helps to achieve better quality judgment set and also can achieve better results with lesser pool depth. Both, by considering only the most effective topics and good contributing systems documents helps to reduce the computational cost of the evaluation process. Lastly, it has been proven that the proposed methodology helped to reduce the incompleteness of the judgment sets, and biasness in the ranking of the judgment sets.

Keywords: Information retrieval evaluation, pooling, document similarity, incomplete judgments, rank biasness

MENINGKATKAN KETEPATAN PENILAIAN SISTEM PENCAPAIAN MAKLUMAT DENGAN MENINGKATKAN KUALITI PENGHAKIMAN YANG BERKAITAN

ABSTRAK

Penilaian perolehan maklumat adalah satu proses untuk mengukur sejauh mana sistem yang mengambil bahagian dapat memenuhi maklumat yang diperlukan oleh pengguna. Prestasi sistem dinilai berdasarkan kualiti set pertimbangan perkaitan. Kualiti set penghakiman diukur berdasarkan keupayaan sistem yang mengambil bahagian untuk mendapatkan sebanyak mungkin dokumen yang berkaitan berdasarkan topik ke dalam set penghakiman dan menyusunnya dengan cara yang lebih baik dan juga, pada masa yang sama menyekat dokumen yang tidak berkaitan. Walau bagaimanapun, telah diperhatikan bahawa untuk koleksi ujian yang lebih kecil, andaian ini mungkin betul. Tetapi untuk koleksi ujian besar seperti TREC, andaian ini mungkin tidak benar selalu. Telah diperhatikan bahawa kualiti set penghakiman tidak mencapai tahap atau tidak lengkap mengikut metodologi paradigma Cranfield, terutamanya melalui teknik persamaan dokumen. Matlamat utama tesis ini adalah untuk meningkatkan kualiti set pertimbangan yang relevan semasa proses penilaian. Kualiti set penghakiman boleh ditingkatkan dengan menambah bilangan dokumen yang berkaitan dalam set penghakiman. Ia secara tidak langsung akan membantu meningkatkan ketepatan proses penilaian. Sumbangan utama tesis ini adalah untuk meningkatkan kualiti set penghakiman dengan mencadangkan beberapa metodologi. Percubaan pertama meneroka isu pertimbangan perkaitan separa pada metodologi sedia ada. Ketidakupayaan metodologi untuk mendapatkan semula semua dokumen yang berkaitan ke dalam set penghakiman relevan dipertimbangkan. Dengan mempertimbangkan batasan metodologi sedia ada, satu metodologi telah dicadangkan untuk menambah dokumen yang berkaitan dalam set penghakiman. Metodologi yang dicadangkan menggabungkan pengumpulan dan persamaan dokumen menggunakan teknik pengelompokan dan pengelasan. Persamaan dokumen telah dilakukan antara dokumen terkumpul dan berkelompok atau dokumen terperingkat yang tidak dinilai. Jika persamaan ditemui, skor baharu akan diberikan kepada dokumen tersebut dan memindahkan dokumen tersebut ke dalam senarai terkumpul. Penilaian diteruskan sehingga semua dokumen daripada senarai terkumpul dipertimbangkan untuk proses semakan persamaan. Keputusan menunjukkan bahawa metodologi yang dicadangkan boleh mencapai lebih banyak dokumen berkaitan dalam set penghakiman dan juga membantu untuk mencapai hasil yang lebih baik dengan kedalaman kumpulan yang lebih rendah. Percubaan kedua meneroka cara untuk menambah baik atau mengekalkan kualiti set penghakiman dengan mempertimbangkan pengumpulan ujian. Untuk eksperimen ini, topik dan sistem yang mengambil bahagian daripada koleksi ujian telah dipertimbangkan. Berdasarkan keputusan, telah terbukti bahawa sebilangan kecil topik yang paling berkesan, atau topik mudah, dapat mengekalkan kualiti set penghakiman. Selain itu, berdasarkan sumbangan sistem, metodologi yang dipertingkatkan telah dicadangkan dan hasilnya menunjukkan bahawa ia membantu untuk mencapai set pertimbangan kualiti yang lebih baik dan juga boleh mencapai keputusan yang lebih baik dengan kedalaman kumpulan yang lebih rendah. Kedua-duanya, dengan mempertimbangkan hanya topik yang paling berkesan dan dokumen sistem penyumbang yang baik membantu mengurangkan kos pengiraan proses penilaian. Akhir sekali, telah terbukti bahawa metodologi yang dicadangkan membantu mengurangkan ketidaklengkapan set penghakiman, dan juga, berat sebelah dalam kedudukan set penghakiman.

Katakunci: Penilaian perolehan maklumat, pengumpulan, persamaan dokumen, pertimbangan tidak lengkap, berat sebelah pangkat

ACKNOWLEDGEMENTS

First and foremost, all praises and thanks to God, the Almighty, and Mother Mary, for their showers of blessings throughout my studies and in the completion of this work.

I would like to express my deep and sincere gratitude to my research supervisor, Assoc.Prof.Ts. Dr.Sridevi A/P Ravana, for allowing me to do my research under her invaluable guidance. Her dynamism and motivation have deeply inspired me and invaluable help in the form of constructive comments and suggestions throughout the study really contributed to the success of this research.

I am extremely grateful to my parents, Mr. N. J Joseph and Mrs. Mary Joseph for their love, prayers, and encouragement. My life partner, Mr. Jose Gregorious Perumalil, thank you so much for your support mentally and financially. My kids, Annmariam Jose Perumalil and Georgy Ouseph Jose Perumalil, thank you so much for your love, understanding, and continuing support. Despite all the difficulties, you all stand by me without fail.

Lastly, my appreciation goes to my colleagues and friends for their kindness and moral support. Your kindness means a lot to all those who indirectly contributed to this study. Thank you very much.

TABLE OF CONTENTS

UNIVERSITi MALAYA

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBRIVATIONS

IR	:	Information Retrieval
IRE	:	Information Retrieval Evaluation
TREC	:	Text REtrieval Conference
NIST	:	National Institute of Standards and Technology
CLEF	:	Cross Language Evaluation Forum
ANOVA	:	Analysis of Variance
VLC	:	Very Large Corpus
FIRE	:	Forum of Information Retrieval Evaluation
LSR	:	Learned Sparse Representations
LSA	:	Latent Semantic Analysis
LDA	:	Latent Dirichlet Allocation
AP	:	Average Precision
MAP	:	Mean Average Precision
GMAP	:	Geometric Mean Average Precision
TF-IDF	:	Term Frequency-Inverse Document Frequency
NDCG	:	Normalized Discounted Cumulative Gain
RBP	:	Rank Biased Precision
ICIR	:	Intelligent Cluster-based Information Retrieval
CAL	:	Continuous Active Learning
AAP	:	Average of Average Precision
TD	:	Topic Difficulty

CHAPTER 1: INTRODUCTION

Nowadays, there is a vast amount of information available both on paper and online media. The online-based repositories are constantly adding new data and it keeps growing. Data can be stored in any format, including text, audio, and video. People primarily use online media to find answers to their questions or clarifications. Finding the most relevant information from these huge repositories is not so easy. If the search engines are unable to find a sufficient number of relevant documents, this will have an impact on users' reliability and their loyalty to the search engines. The real users will benefit by getting the correct or reliable information that they need through some mechanisms or methodologies that are proposed by the researchers.

Information retrieval Evaluation is a process of measuring how well the systems can retrieve as much of relevant documents based on user queries. Evaluation can be done either system-based or user-based. System-based evaluation completely relies on a test collection and user-based evaluation depends on real user feedback. User-based evaluation measures the user’s satisfaction and requires a large sample of users for evaluation purposes. User-based evaluation is expensive and difficult to do it correctly. The system-based evaluation depends on a set of test collections which consists of a document corpus, topics or user queries, and a set of relevant judgment sets. These relevant document sets are already judged by the experts based on the topics or user query and document corpus (Jones & Willett,1997), (Voorhees, 2001)(Rajagopal et al., 2022).

The information retrieval evaluation process involves participating systems or search engines that retrieve a set of relevant documents from the document corpus based on the topics or queries given by the users. The number of relevant documents retrieved by the participating systems is evaluated through the already judged relevant judgment sets available in the test collection. The number of relevant documents retrieved will depend on the efficiency of the participating systems. The performance of the participating systems is evaluated through various evaluation metrics. Evaluation of the information retrieval process based on system-based is shown in Figure 1.1.

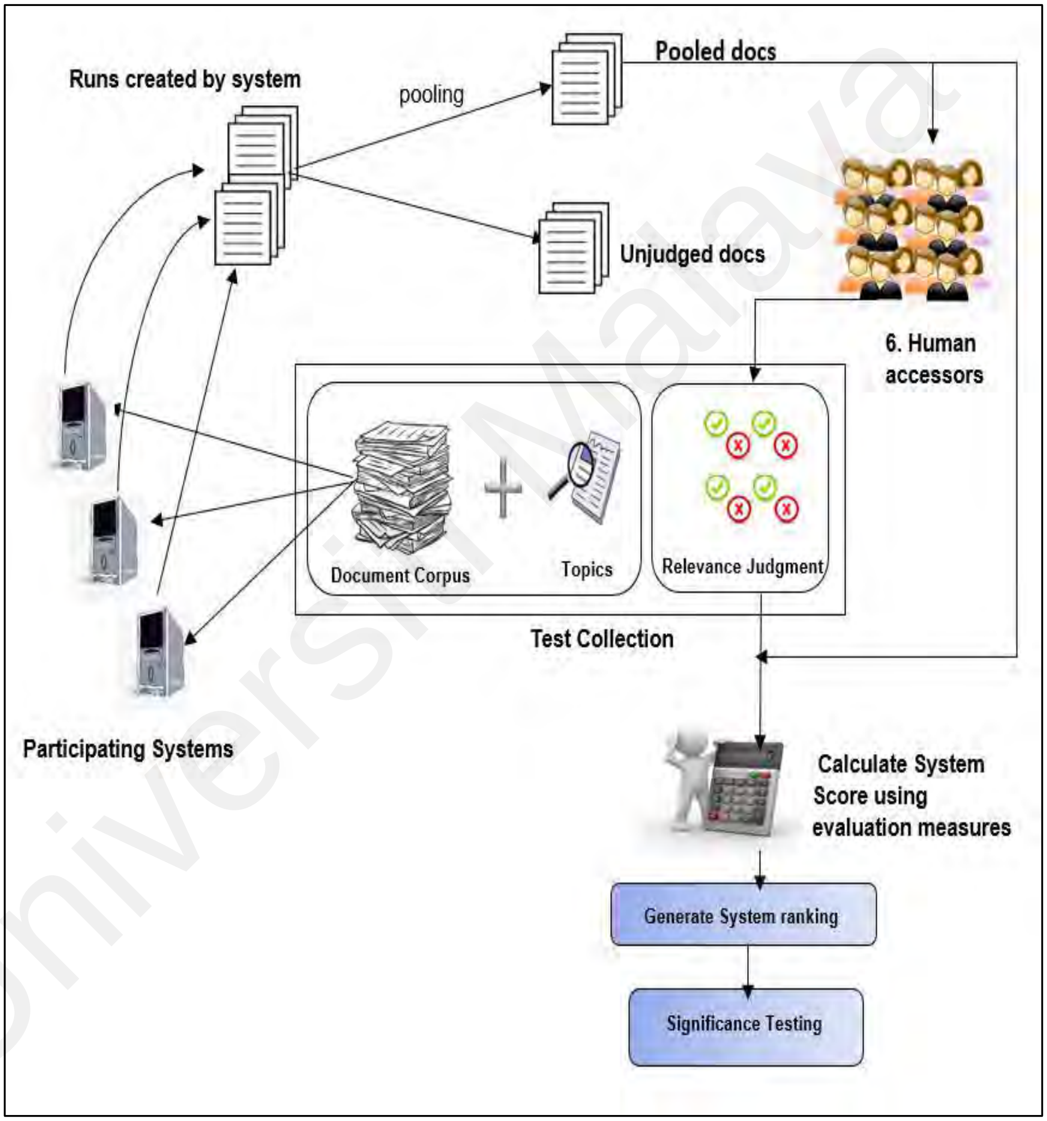


Figure 1.1: System based information retrieval evaluation process

1.1 The importance of the Quality of the Judgment sets

Information Retrieval is a process of retrieving relevant documents from the document corpus or web collection based on a user query by the participating systems. The main aim of the information retrieval evaluation process is to measure how good are the participating systems in the matter on its performance. The performance of the participated systems is determined not only by their efficiency in terms of speed, time, storage space, etc. (Wu, 2016), (Mhawi et al., 2022), but also by their effectiveness, that is the ability to retrieve as many relevant documents as possible into the relevance judgment sets with a better ranking based on its relevancy and also at the same time, suppressing the irrelevant ones (Nichola Ferro,2017), (Lin et al., 2021). The quality of the relevance judgment set matters to the performance of the participating systems. If the quality of the relevance judgment set is lesser, the accuracy of the evaluation process also gets lesser. If a greater number of relevant documents are added to the relevance judgment set, it increases the quality of the judgment set and through that, it can increase the accuracy of the evaluation process (Rahman et al., 2020). Indirectly it will increase the user's reliability and loyalty to the search engines or the participating systems.

1.2 Motivation of the Study

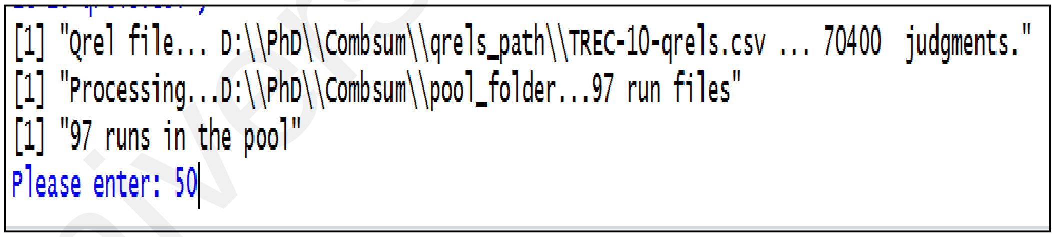
Based on the user's query, the participating systems have to retrieve as many relevant documents as possible and return them to the users with a better ranking based on the relevancy of the documents. For that, in the information retrieval evaluation process, the researcher aims to build an unbiased set of relevance judgment sets, which helps to increase the accuracy of the evaluation process. Through that, it helps to increase the user's reliability to the participating systems. Building an unbiased set of relevant judgments from the multiple ranked lists generated by the participated systems will lead to higher information retrieval evaluation accuracy, especially if the judgment list is rich in relevant documents. This research aims to improve the quality of relevant documents in the relevant judgment list by augmenting an unbiased set of relevant documents from the unjudged list using proposed methodologies.

In the evaluation process, each participating system collects a set of relevant documents from the document corpus. This relevant document list is called runs and these runs will be ranked according to the relevancy. These runs will be merged using some rank aggregation techniques such as Doc_ID (Voorhees,2002), Rbp (Moffat et al., 2007), (Moffat &Zobel, 2008), Borda (Aslam & Montague,2001), Combsum and CombMNZ (Shaw & Fox,1994). These runs can be given for the evaluation process. However, judging the whole document is not feasible, costly, and time-consuming (Roitero et al., 2022). An alternative approach called pooling (Aslam et al., 2006), (Carterette, 2007), (Voorhees, 2001) can be used to take a sample of documents. As these documents are ranked according to their relevancy, the documents on the top are considered relevant, and by taking the top *p* documents (most probably 50<p<100) from each ranking will be considered as to-be-judged documents. This set of documents is called a pool of judged documents or pooled documents and this technique is referred to as pooling (Buttcher et al., 2007), (Clarke et al., 2008), (Valcarce et al., 2020), (Sparck Jones & Van Rijsbergen,1976). Pooling assumes that the sample of relevant documents found in the pooled list is unbiased (Buckley et al., 2007).

Pooling is a good technique that can be used to evaluate the quality of the first *p* documents by the search systems (Buttcher et al., 2007). The documents that were not included in the pooled list were called unjudged document lists and assumed it as irrelevant. If doing shallow pooling from a small test collection, pooling is a better technique. But shallow pooling with large test collections like TREC, ClueWeb, and all, results might show all potentially relevant documents unjudged. All the relevant documents were not included in the pooled list (Shani & Gunawardena, 2011)(Rajagopal et al;,2022). The quality of all the documents retrieved by the search system cannot be evaluated using this pooling technique. One of the limitations or biasness of the Cranfield paradigm is the incompleteness of the judgments (Valcarce et al., 2018) (Voorhees,2011). The issue with the incompleteness of the judgment list is shown in Figure 1.2 and Figure 1.3.

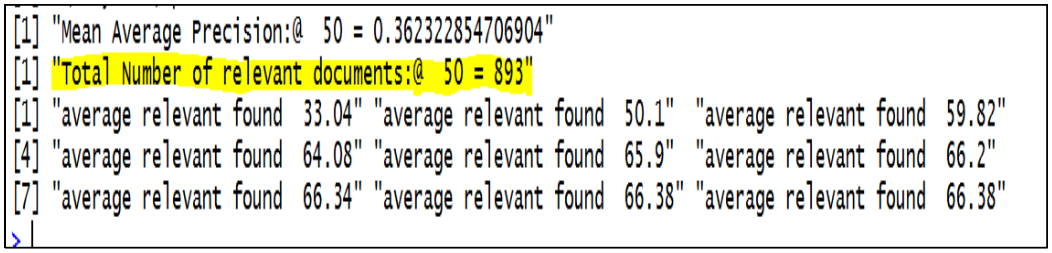
The experiment was done with TREC-10 collection in which the document corpus size is about 10GB. A total of 97 participated systems were involved in this experiment and a total of 50 topics were considered.70400 judgment set included in the relevant judgment set.

First, the experiment was run with pool depth, *p*=50.



```
[1] "Qrel file... D:\\PhD\\Combsum\\qrels_path\\TREC-10-qrels.csv ... 70400  judgments."
[1] "Processing...D:\\PhD\\Combsum\\pool_folder...97 run files"
[1] "97 runs in the pool"
Please enter: 50
```

Figure 1.2: Pool depth 50 assigned to TREC-10



```
[1] "Mean Average Precision:@  50 = 0.362322854706904"
[1] "Total Number of relevant documents:@  50 = 893"
[1] "average relevant found  33.04" "average relevant found  50.1"  "average relevant found  59.82"
[4] "average relevant found  64.08" "average relevant found  65.9"  "average relevant found  66.2"
[7] "average relevant found  66.34" "average relevant found  66.38" "average relevant found  66.38"
> |
```

Figure 1.3: Mean Average precision and total number of relevant documents retrieved from TREC-10 with pool depth 50

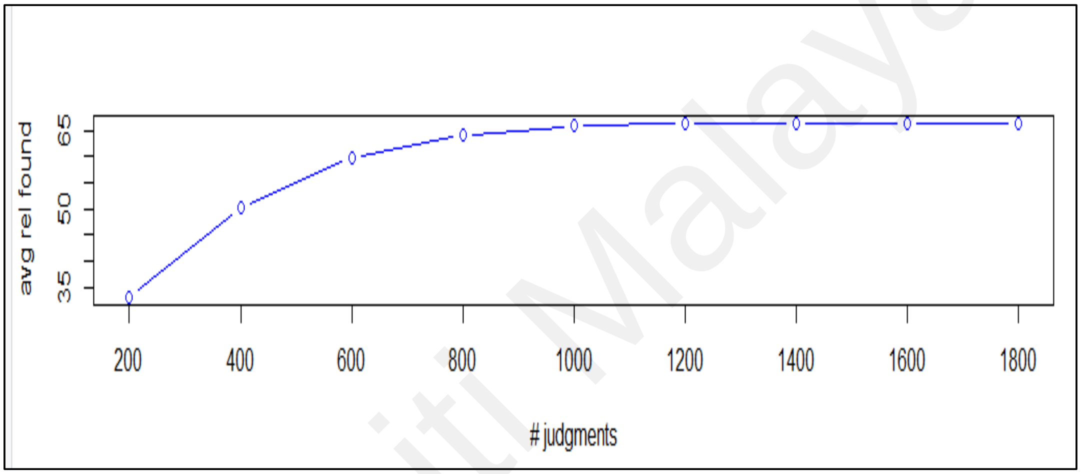
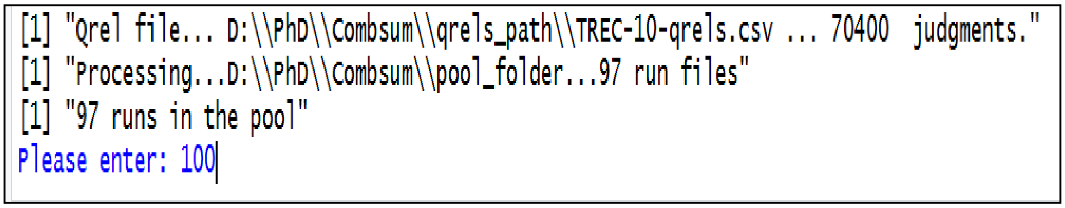


Figure 1.4: Average number of relevant documents retrieved for several judgments from TREC-10 with pool depth 50

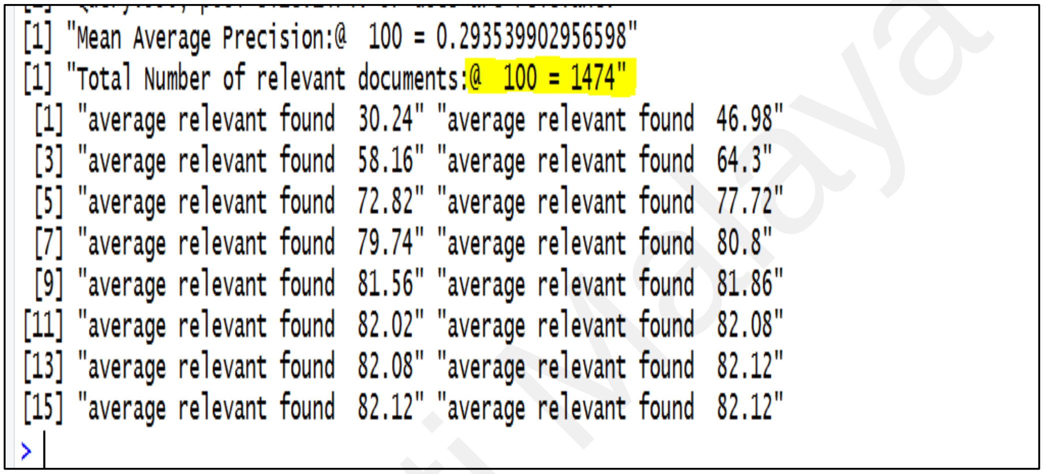
Figure 1.2, Figure 1.3, and Figure 1.4 show the results of pool depth 50. As per the results from the top 1800 documents, only 893 documents are relevant.

Now tried to increase the pool depth to 100, to find the more relevant documents. As per the assumption, more relevant documents might be in the pooled list by assigning lower ranks. In standard TREC settings, a pool depth of 100 is a standard pool depth suggested by the researchers as it has always been shown to be an effective way to evaluate the retrieval systems' effectiveness (Zobel, 1998), (Yilmaz & Aslam, 2006).



```
[1] "Qrel file... D:\\PhD\\Combsum\\qrels_path\\TREC-10-qrels.csv ... 70400  judgments."
[1] "Processing...D:\\PhD\\Combsum\\pool_folder...97 run files"
[1] "97 runs in the pool"
Please enter: 100
```

Figure 1.5: Pool depth 100 assigned to TREC-10



```
[1] "Mean Average Precision:@  100 = 0.293539902956598"
[1] "Total Number of relevant documents:@  100 = 1474"
 [1] "average relevant found  30.24" "average relevant found  46.98"
 [3] "average relevant found  58.16" "average relevant found  64.3"
 [5] "average relevant found  72.82" "average relevant found  77.72"
 [7] "average relevant found  79.74" "average relevant found  80.8"
 [9] "average relevant found  81.56" "average relevant found  81.86"
[11] "average relevant found  82.02" "average relevant found  82.08"
[13] "average relevant found  82.08" "average relevant found  82.12"
[15] "average relevant found  82.12" "average relevant found  82.12"
>
```

Figure 1.6: Mean Average precision and Total number of relevant documents retrieved from TREC-10 with pool depth 100

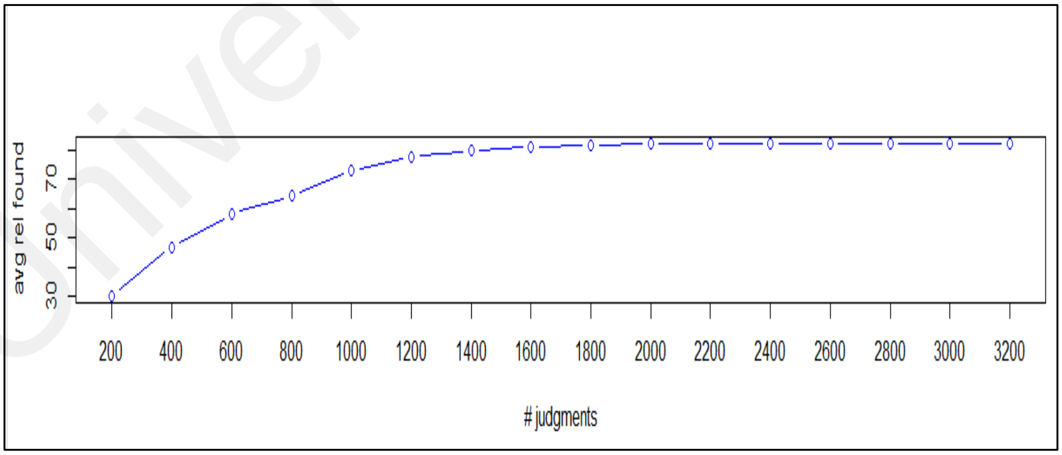
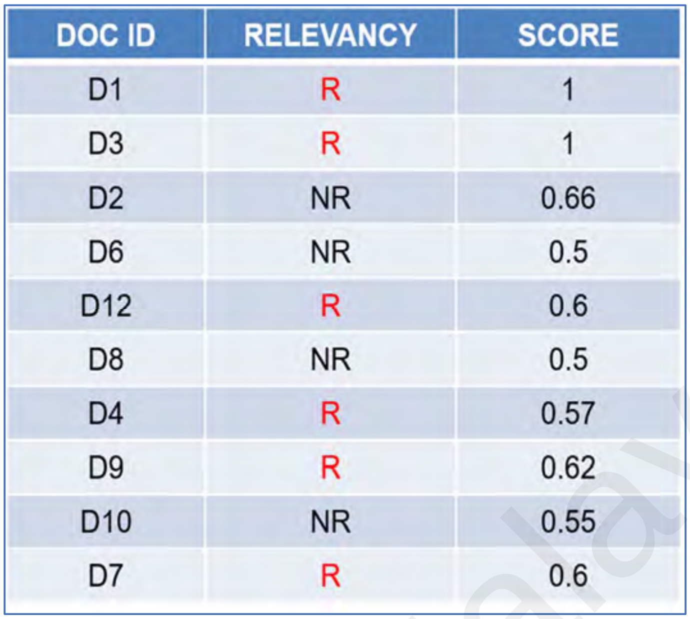


Figure 1.7: Average number of relevant documents retrieved for several judgments from TREC-10 with a pool depth of 100.

As per the results with pool depth 100, Figure 1.5, Figure 1.6, and Figure 1.7 show that even with pool depth 100, from the top 3200 documents only 1474 relevant documents are there. This shows the results of the incompleteness. Many relevant documents have not moved into the pooled list due to the performance of the system. This incompleteness is calculated using Mean Average Precision(MAP) of each topic with pool depth 100.Also, sometimes when the document collection is dynamic, in web collection, documents are kept on getting added over time, and during the judgment time, the pooled list will become a smaller subset of the entire document collection. Also, some documents might get deleted from the collection such as broken links. These all can make the relevance judgment set *imperfect* (Yilmaz & Aslam, 2006)(Kirnap et al., 2021).

Secondly, ranking the documents based on their relevancy shows the system's effectiveness. The ranking of documents has a high impact on the mean average precision score. Many search systems are sometimes defined as non-probabilistic models, which are not capable of handling uncertainty about document relevance (Diaz et al., 2020). If the document is relevant and not assigned a good ranking over non-relevant documents, affects the precision score and accuracy of the evaluation process. Two types of ranking problems were described. Ranking creation and Ranking aggregation. Ranking creation is to create the ranking list of documents using the similarity of the document-topic pairs and Ranking aggregation is to create a ranking list of the documents using multiple ranked lists of runs that have the ranked documents (Li,2022). If ranking is not done correctly, the accuracy of the evaluation process also goes down. One of the main issues of bias of ranking is that participating systems are assigning higher ranks to the documents that are not relevant than the highly relevant documents.

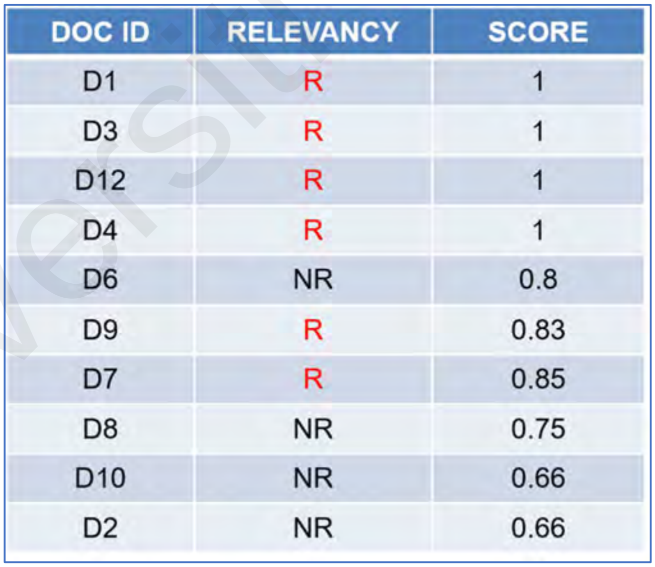
System A



DOC ID	RELEVANCY	SCORE
D1	R	1
D3	R	1
D2	NR	0.66
D6	NR	0.5
D12	R	0.6
D8	NR	0.5
D4	R	0.57
D9	R	0.62
D10	NR	0.55
D7	R	0.6

Figure 1.8: An example of a run list from System A

System B



DOC ID	RELEVANCY	SCORE
D1	R	1
D3	R	1
D12	R	1
D4	R	1
D6	NR	0.8
D9	R	0.83
D7	R	0.85
D8	NR	0.75
D10	NR	0.66
D2	NR	0.66

Figure 1.9: An example of a run list from System B

Figure 1.8 and Figure 1.9 shows an example of biasness in the ranking. Consider two systems, System A and System B. Both systems retrieved the same set of documents with different rankings. If we calculate the Average precision, System A generates Average Precision, AP @10=0.731

and System B generates Average Precision, AP @ 10=0.941

The same set of documents retrieved by the two systems with different rankings shows that if the systems rank the relevant documents without biasness will increase the precision value and through that can increase the effectiveness of the evaluation process.

Figure 1.10 shows the absolute differences in the mean average precision of the pooled documents evaluated with and without relevant document sets from the TREC-8 adhoc test collection. In this graph, systems have been randomly chosen from the systems that contributed to the pooled list. The mean average precision of runs with a value, of 3.85 has the absolute difference with 0.0001.

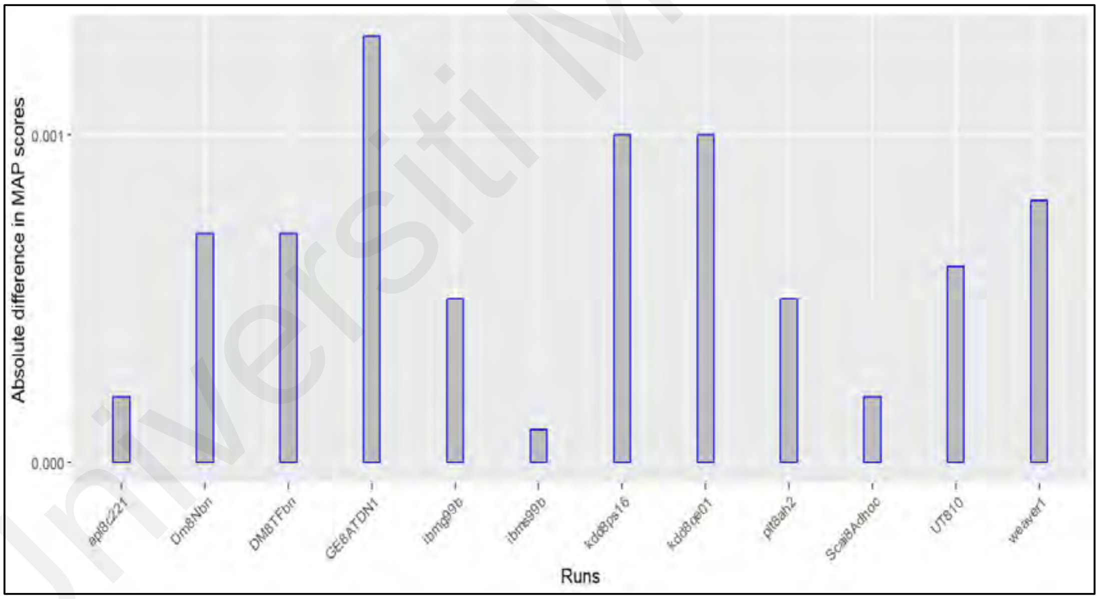


Figure 1.10: Absolute difference in the mean average precision of random samples of runs evaluated with and without relevant documents in TREC-8

The quality of the judgment sets helps to increase the user satisfaction and reliability of the participated or contributed systems. In order to improve the quality of the judgment sets, more number of relevant documents have to be there in the judgment sets, and also need to reduce the biases in the ranking of the relevant documents in the judgment sets. So this research mainly focused on how to improve the quality of the relevance judgement sets and through that increase the accuracy of the evaluation process. The experiments related to this research run on a test collection called TREC, which was developed by the NIST organization, a large dataset mainly to support research within the information retrieval community. This large-scale test collection contains document corpus, topics or queries, and also relevant judgment sets.

1.3 Problem Statements

The main aim of this research is to improve the quality of the judgment sets. However, as discussed in Section 1.2, it has been noticed that many relevant documents were not retrieved into the judgment sets during the pooling process. It leads to the incompleteness of the judgment process. Also, it has been noticed that the ranking of the documents in the judgment sets is uneven. These all will affect the quality of the judgment process. These issues are highlighted in detail below.

Partial relevance judgments- One of the main issues addressed by this research is the partial relevance judgments. During the pooling process, researchers assume that all the top relevant documents from each run have moved into the pooled list and these documents can be given for evaluation purposes. However, as per the experiments, all relevant documents have not moved into the pooled list due to some system's performance in ranking of the documents. The relevant documents in the unjudged list are considered irrelevant and not considered for the evaluation process. It affects the quality of the relevance judgment sets and affects the accuracy of the evaluation process.

Biasness- Using some rank aggregation techniques, the runs will be merged. From these runs, pooled documents were selected. Even in the pooled list, documents were not ranked according to their relevancy. Due to that average precision value of the systems goes down. The documents that are not relevant are getting higher ranks than the documents that are relevant. It creates the biasness in the system's effectiveness during the evaluation process.

This research entitled these two problems and tried to increase the number of relevant documents in the relevance judgment sets and through that increase the quality of the judgment sets through a cost-effective method and through that increase the accuracy of the information retrieval evaluation process. Also, this research focused on reducing the biasness in the system rankings of the relevant documents and through that increasing the system effectiveness in the evaluation process.

1.4 Research Questions

Based on the problems identified and considered for this research, the research questions proposed are highlighted below.

- RQ 1. How the document similarity and pooling methodologies within the document manifold will increase the number of relevant documents in the pooled list based on their relevancy?

  RQ1.1: How do the document similarity and pooling methodologies increase the number of relevant documents in the pooled list/judgment list based on their relevancy?

RQ1.2: Is there variation in the system rankings when using different evaluation depths and pool depths for the evaluation process?

RQ1.3: Can the consideration of unjudged documents help to retrieve more relevant documents compared to the baseline methodologies proposed earlier by the researchers?

RQ1.4: Which clustering techniques help the proposed methodology to perform better?

- RQ 2. How do the global similarities between the documents and considering system evaluation scores increase the quality of the relevant judgments?

RQ2.1: Can the participating systems retrieve a greater number of relevant documents with reduced topic size?

RQ2.2: Can the participating systems retrieve a greater number of relevant documents by considering documents from good contributing systems?

RQ2.3: Does considering good contributed systems documents specifically have any benefit over baseline methodologies?

- RQ 3. How to overcome the baseline methods limitations in terms of biasness in the ranking and incompleteness in the judgment sets?

RQ3.1: Can the consideration of document similarity and unjudged documents from pooling techniques help to reduce the biasness of the system rankings during the evaluation process?

RQ3.2: How to evaluate the effectiveness of the proposed ideas compared to the baseline methodologies using various evaluation metrics?

1.5 Research Objectives

- RO 1. To propose an experimental methodology to improve the accuracy of Information Retrieval Evaluation by increasing the number of relevant documents in the judgment list as compared to the baseline methodologies

  RO1.1 To propose an experimental methodology to improve the accuracy of the information retrieval evaluation by considering documents from unjudged clustered or classified documents set.

  RO1.2 To measure the system performance using various evaluation metrics with different evaluation depths and pool depths.

  RO1.3 To measure the effectiveness of the proposed methodology in terms of the number of relevant documents.

  RO1.4 To explore the effectiveness of the proposed methodology using different clustering techniques.

- RO 2. To increase the quality of the relevant judgment list by considering the topics and participating systems compared to the baseline methodologies.

  RO2.1 To measure the quality of the relevant judgment list by considering the reduced topic size.

  RO2.2 To increase the quality of the relevant judgment list by considering participating systems evaluation scores.

  RO2.3 To measure the effectiveness of the proposed methodologies' cost-effectiveness based on topics and participated systems.

- RO 3. To measure the effectiveness of the proposed evaluation methodology in terms of biasness in ranking and incompleteness in ranking.

RO3.1 To measure the effectiveness of the proposed methodologies by considering biasness in ranking documents.

RO3.2 To measure the overall effectiveness of the proposed methodologies to improve the accuracy of the evaluation process.

1.6 Scope

Information Retrieval is a process of retrieving relevant documents from the document corpus or web collection based on a user query. How much relevant information is retrieved from the users is evaluated based on the information retrieval evaluation process. Document corpus consists of various types of data including text, audio, and video. This thesis mainly concentrated on test collections which consist of text-based documents only. For image-based and audio-based test collections, methodologies and techniques considered might be different. In addition, in this research the primary focus is on how to increase the number of relevant documents in the judgment list, and through that it helps to increase the quality of the judgment sets and through that can increase the accuracy of the evaluation process. And this research is not to focus on optimizing or improving the quality of any test collection.

1.7 Contributions of the research

Notable contributions have been given in this thesis in the area of IR research. These contributions can result in the answers to the research questions mentioned above.

To start with, this research's main aim is to increase the quality of the relevant judgment sets during the IR evaluation and through that increase the accuracy of the evaluation process. Based on the literature review, it has been noticed that existing methodologies or techniques have limitations in producing enough relevant documents in the relevance judgment sets. Some existing methodologies have tried to increase the number of documents in the judgment sets, but the quality of these documents based on their relevancy is lesser compared to the traditional approaches. It might be due to either not having enough relevant documents in the relevance judgment sets or the ranking of these documents might not be accurate.

First, An experimental methodology has been proposed to increase the number of relevant documents in the judgment sets. The results have shown that the proposed one was able to increase the quality of the judgment sets compared to the baseline methodologies. Also, compared to the baseline works, the proposed methodologies were able to retrieve a greater number of relevant documents with lesser pool depth. It helps indirectly to improve the accuracy of the evaluation process by improving the system effectiveness score. Also, through the evaluation measures, it has been noticed that the systems perform better based on different pool depths and evaluation depths. The system performs better when the pool depth is greater than the evaluation depth and also when the pool depth is equal to the evaluation depth. Also, when the number of relevant documents in the relevance judgment sets is greater than the pool depth and when the evaluation depth becomes greater than the pool depth, the performance of the system varies significantly. Also, various clustering techniques have been compared to the proposed methodology and shown which one performs better.

Secondly, To further improve the quality of the judgment sets by cost-effectively considering the test collections. Test collections have a greater impact on the quality of the judgment sets. This research has considered topics and participated systems components. Here cost effectiveness is considered based on the topic sizes and participated systems efficiency. The results show that if consider only easy topics or most effective topics, can maintain the quality of the judgment sets. Even hard topics or less effective topics also can be considered for judgment purposes, but the topic size needs to increase. However, with a smaller number of effective topics, easy topics can maintain the reliability of the effectiveness measurement of information retrieval systems.

Also, With the enhancement of the proposed methodologies by considering good participated systems documents, the results show that a greater number of relevant documents were able to achieve the relevance judgment set. Also, it has shown that with lesser depth itself, systems were able to achieve more relevant documents. Also, it has shown that as the judgment document size increases, the MAP value gets closer to all the methodologies. These results show that computational cost gets lesser with reduced topic size and good contributed systems documents by using proposed methodologies.

Thirdly, consider the problem of incompleteness or biasness in the judgment list. This means the judgments that are biased against the systems or systems that are not contributing enough relevant documents to the pooled document list. Also considered the issues on systems that are assigning higher ranks to the documents that are irrelevant than the documents that are relevant. Whenever the relevance judgment set size increases, the incompleteness or the biasness also increases. The consistency is not maintained well. The results show that with the help of score adjustments and document selection for the relevance judgments, incompleteness has been reduced. Also, the proposed methodologies were able to adjust the order in which the documents are added to the judgment set. With the help of bpref measure, it has been proven that incompleteness has reduced, and it is almost flatter as the relevant judgment set size increases. Also, it continuously ranks all systems in the same preference order as when using the complete judgments for a higher level of incompleteness. The overall contribution of this research is shown in Figure 1.11.

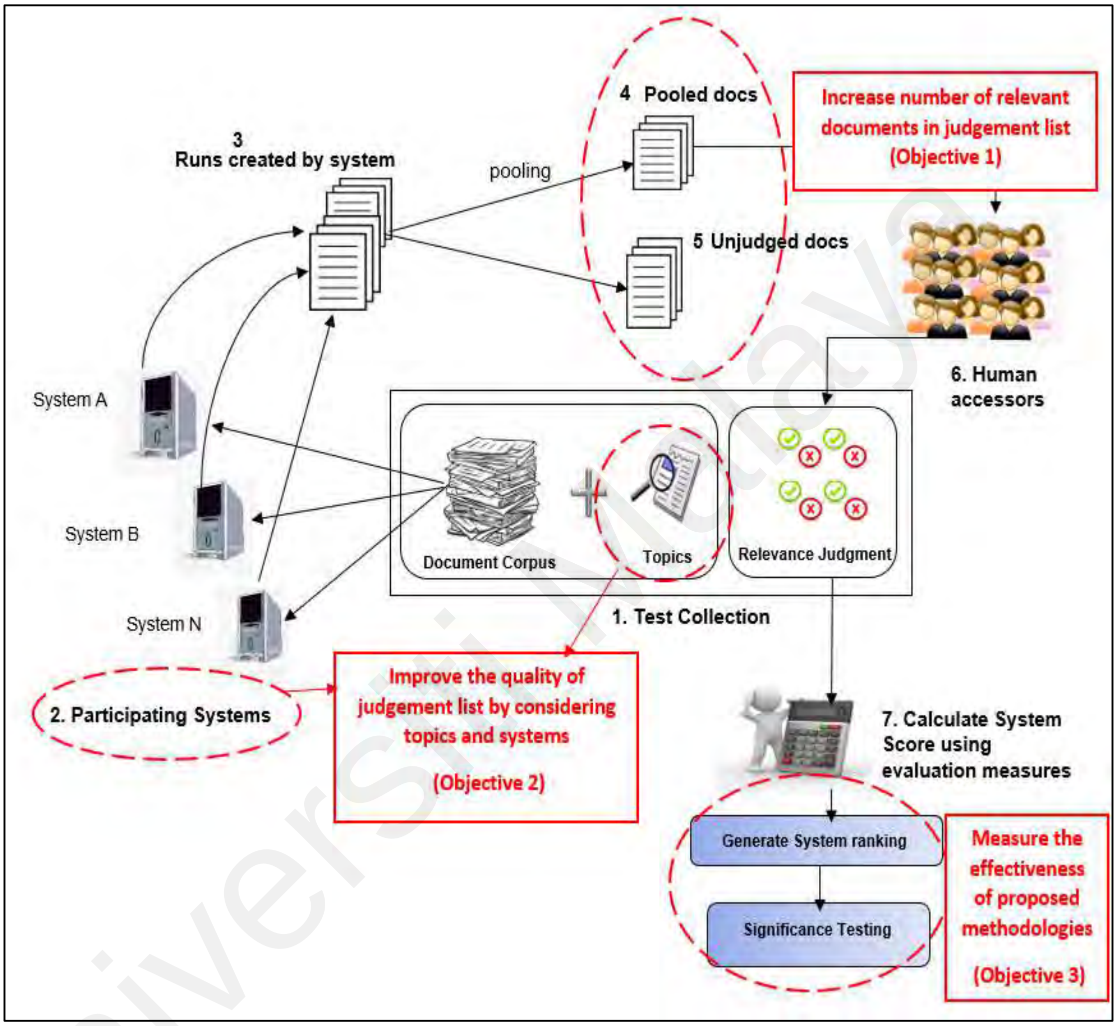

4 Pooled docs
Increase number of relevant documents in judgement list (Objective 1)
3 Runs created by system
pooling
5 Unjudged docs
6. Human accessors
System A
System B
System N
Document Corpus
Topics
Relevance Judgment
1. Test Collection
2. Participating Systems
Improve the quality of judgement list by considering topics and systems (Objective 2)
7. Calculate System Score using evaluation measures
Generate System ranking
Significance Testing
Measure the effectiveness of proposed methodologies (Objective 3)

Figure 1.11: System-based IR evaluation with contributions

1.8 Thesis Structure

The overall structure of the thesis is described below. Chapter 2 gives a detailed overview of information retrieval, information retrieval evaluation, various ways of evaluating information retrieval, the importance of retrieval evaluation, the various metrics used in the information retrieval evaluation, The TREC dataset collections, and the other test collections like CLEF, NTCIR, etc have outlined in detail in this session. Section 2.1 gives a detailed view of the various existing methodologies used to increase the number of relevant documents into the judgment list and also the effectiveness of these methodologies was described in detail. Later, it described the issues of the existing methodologies' performances in the matter of retrieving the number of relevant documents. Section 2.2 gives a detailed description of test collections with topics and participating systems. It also explained how these components have an impact on the existing methodologies in the matter of the quality of relevance judgment sets. In Section 2.3, the issues of incompleteness or biasness in the judgment sets were considered. Also, the impact on the ranking of documents in the judgment sets is considered.

Chapter 3 covers the methodology proposed by this research. Section 3.1 describes the research approach and Section 3.2 describes the overall research framework and experimental methodology that has been proposed in order to retrieve a more effective number of relevant documents from the runs. Also, the effectiveness of this methodology over different clustering techniques was explored well.

Chapter 4, describes how to maintain the quality of relevant judgment sets by using the reduced topic size and through that reduce the computational cost in the matter of lesser test collection. Also, this section proposes an improved version of the proposed methodology mentioned in Section 3.2, which considers the documents only from a set of retrieval systems that have been chosen based on some evaluation score and through that also reduces the computational cost in the matter of lesser test collection. The performance of the proposed methodologies by ranking the documents in the relevant judgment sets based on their relevancy and improving the quality of the judgment sets have been explored using the standard evaluation metrics.

Chapter 5 covers the results and discussions of the methodologies proposed in Chapter 3 and Chapter 4. Section 5.1 covers the performance of the methodology proposed in Section 3.2 has been evaluated by comparing it with the baseline methodologies. The effectiveness of this methodology has been evaluated using various evaluation metrics by considering different pool depths and evaluation depths. and found out which clustering technique might perform better results in the matter of increasing the quality of the judgment sets. Section 5.2 covers the effectiveness of the improved proposed methodologies from Sections 4.1 and 4.2, based on reduced topic size, and considered documents from systems based on evaluation scores are explored in this session. Section 5.3 covers the effectiveness of the proposed methodologies to reduce the biasness in ranking documents compared to the baselines and at the same time how effectively improved the quality of the judgment sets are shown.

Chapter 6 concludes the thesis by emphasizing the thesis contributions to the proposed methodologies and their effectiveness in improving the quality of the judgment sets and through that increasing the accuracy of the evaluation process. It also includes the limitations and future works that can enhance the performance of these methodologies.

CHAPTER 2: LITERATURE REVIEW

The overall view of information retrieval, information retrieval evaluation, various evaluation metrics, and the various test collections available for the evaluation process are described in detail in this session.

2.1 Information Retrieval

The field of Information Retrieval(IR) was born in the 1950s and the term "Information Retrieval" was popularized among the IR community researchers in 1961. Information retrieval is a process of retrieving information from the raw data, which is in a large database collection (Sagayam et al., 2012). This process includes filtering, searching, matching, and ranking operations (Roshdi & Roohparvar, 2015). These processes have three main components including the contents of the document, the user's information need, and a comparison of these two. The contents of the document needed an indexing process which helps to index the document for the matching process. This indexing process happens offline, in which the end-users are not involved. User-information need is the information searched by the user undergoes a query formulation process and the result of this process is query. Comparison of these two results is called matching and the end of this result is the retrieved documents. Once the retrieved documents are produced, based on the feedback, new user information need or query will be generated (Djoerd,2009).

Whenever a user sends a query to the participating systems, the systems collect a set of documents from the document corpus or the web of collections, these documents will be ranked according to their relevancy and these ranked documents are sent back to the users.

The documents received by the end users will be relevant to the users' query. The flow of the information retrieval process is shown in Figure 2.1.

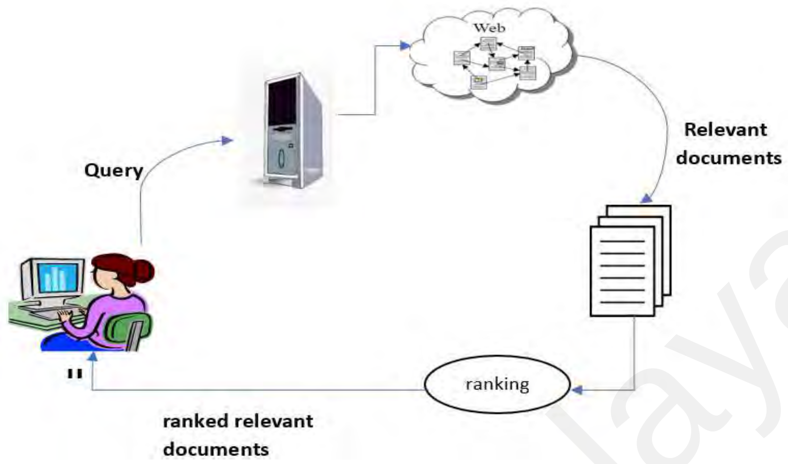


Figure 2.1: Information retrieval process

The quality of the documents retrieved varies based on the participating systems' performance. Many studies have been done on the growth of the internet and the technologies used to retrieve data from various data sources such as web pages, media, and hosts(Martinez-Rodriguez et al.,2020)(Nowrozi et al.,2022). It has shown that 80% of the users are depending on the search engines and search services to collect their information. At the same time, they claim that users are completely not satisfied with the information retrieved by the search engines due to lower retrieval speed, communication delay, noise, and broken links (Kobayashi & Takeda,2000).

Various models of document retrieval have been proposed by the researchers to retrieve the documents effectively. These models were proven effective with small datasets. Later large datasets proposed by the US Government under the organization of NIST(National Institute of Standards and Technology) have changed the view of information retrieval. With the updated models and techniques, the systems were able to retrieve documents more effectively based on their relevance. The performance of these models can be evaluated based on learning-to-rank models consisting of a model for learning and a learned model to re-rank the documents. Based on top-k documents a sample will be created and a query-dependent document retrieval task to see the performance of the systems. By using a loss function, a new query, and ranked top-k documents, the learned model predicts a relavance score to know the quality of the documents retrieved (Aydin et al., 2024)

2.2 Information Retrieval Evaluation

The research methodologies related to the information retrieval systems are based on the Cranfield paradigm, which consists of a set of test collections that is quite large and used for evaluating the quality of the different retrieval methods and techniques. A test collection 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. This judgment set is a binary representation of all the documents to all topics (Voorhees,2002). Cranfield assumes that all the relevant documents have been generated in the judgment lists, which means all the documents that are relevant to all the topics have been collected correctly and moved to the judgment list. For smaller datasets, this assumption is correct, and large datasets like TREC(Text REtrieval Conference), and CLEF(Cross Language Evaluation Forum) might be closely accurate to the Cranfield assumption (Buckley & Voorhees, 2004).

Information Retrieval Evaluation is a process of measuring how well the participating systems meet the information required or needed by the user (Voorhees,2002). The evaluation of information retrieval systems is done for two purposes. First, to know the performance of the systems. Based on the user information needs, the resources must be ranked according to their relevancy. The performance of the retrieval systems is determined not only by their efficiency but also by their effectiveness, that is the ability to retrieve as many relevant documents, rank them according to their relevancy, and at the same time suppress the irrelevant ones (Ferro, 2017). Second, to know why the quality of the relevant judgments is important. The quality of the relevant judgments increases based on the number of relevant documents increases. If we fail to collect enough relevant documents in the judgment set, the quality of the set also decreases, and through that increases the accuracy of the evaluation process, and indirectly it will help the users to rely on the search engines (Rahman et al.,2020).

To evaluate the information retrieval systems, two approaches can be adopted. These are system-based and user-based evaluations. User-based evaluation measures the satisfaction of users with the systems and System-based evaluation measures how well the systems retrieve the relevant documents effectively and at the same time rank them according to their relevancy (Voorhees,2002). The main aim of the information retrieval evaluation is to find out the user's satisfaction with the retrieval documents, so user-based evaluation is preferable to the system-based evaluation. However, User-base evaluation requires a large sample of actual users for evaluation purposes. Each of the systems to be compared must be well developed, same user interface, and with same compilation speeds (Mandl,2008). Also, user-based evaluation is subjective, it varies based on the user's perspective, user requirements, and user's judgments, and is dynamic based on time to time (Zuva et al., 2012). It varies based on the user's readability effort, understandability effort, and also findability effort (Rajagopal & Ravana, 2019). Each experiment requires lots of human participation and thus, it is costly and time-consuming.

On the other hand, system-based evaluation is completely dependent on a test collection that has developed with limited resources of expert judges (Maddalena et al.,2017). The test collection consists of a document corpus, topics, and a set of relevant judgments (Voorhees,2002), (Mandl,2008), (Melucci & Baeza-Yates,2011). Even though it's costly to generate the test collection, the advantage is that it can be reusable for each experiment. The experiments based on test collection consider topics as the main experimental unit and based on each topic, the systems collected documents from the document corpus. The evaluation of the retrieved documents will be based on the relevance judgment set available in the test collection. This set will show the relevancy of each topic to each document (Moghadasi et al.,2013), (Carterette et al,2010).

This thesis is a complete reply to the system-based evaluation, in which the test collection used for this experiment is based on the TREC dataset. TREC is one of the well-known test collections which have developed by the U.S National Institute of Standards and Technologies (NIST) based on a large set of IR evaluation series. This test collection has been used mainly for the evaluation of large-scale text retrieval methodologies. The detailed view of TREC test collection is described in Section 2.3.

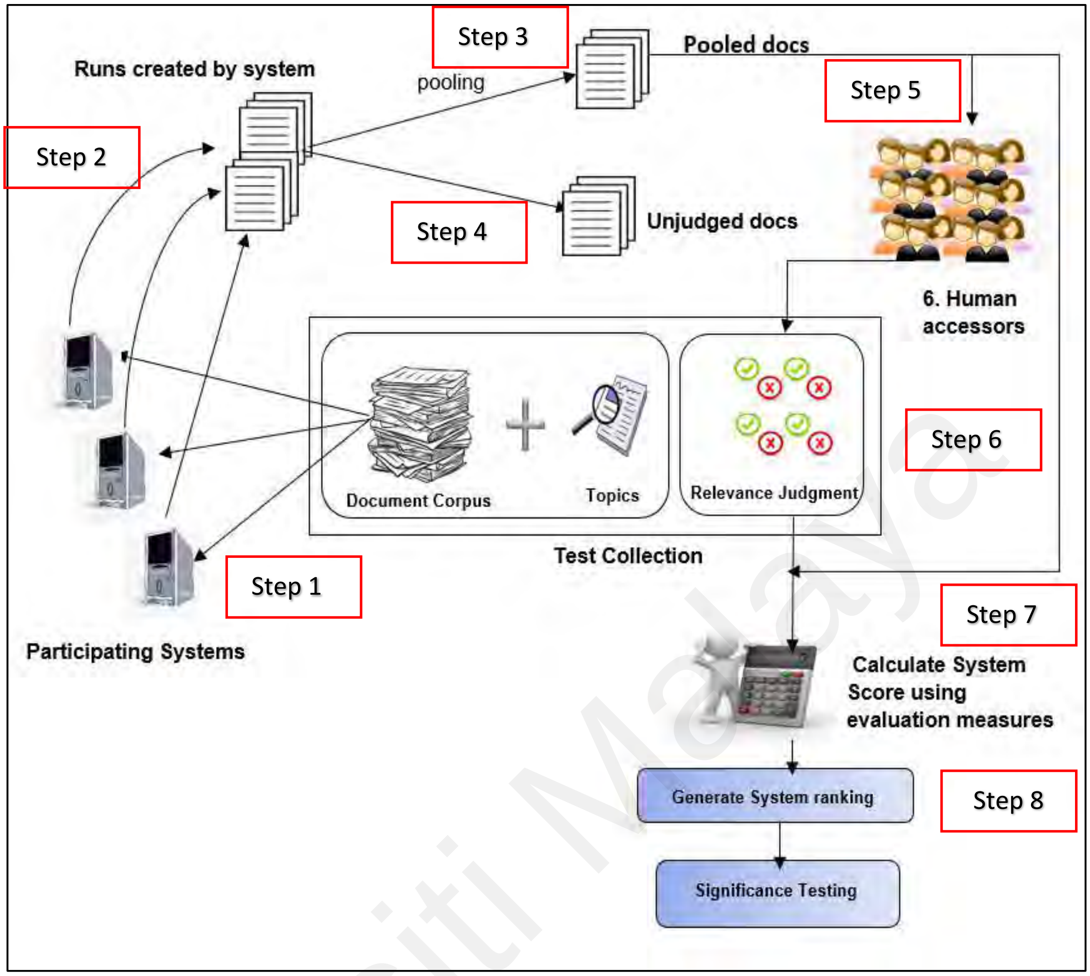


Figure 2.2: Information retrieval evaluation process flow with step-by-step process

The TREC evaluation process works as follows. The TREC data collection consists of a document corpus, topics, and a set of relevant judgments. Each participated system collects a set of relevant documents from the document corpus based on the topics. These documents will be called runs and they will be ranked according to their relevancy. These runs will be merged using any rank aggregation technique and called multiple ranked lists. These ranked lists can be given for the evaluation process. However, judging the whole document is time-consuming and costlier. So, the evaluation initiatives have proposed some techniques to retrieve some sets of documents which considered highly relevant from these runs. Some of these techniques are pooling, sampling, etc. The pooled documents will consider only a subset of documents from the runs which considered highly relevant and sent these pooled documents for the evaluation process (Losada et al.,2018). Evaluation is conducted mainly to find out the performance of the participating systems in how many relevant documents have been retrieved by these systems. Evaluation of these systems is generally done with some evaluation metrics such as Precision, Average Precision, etc, which will be discussed in Section 2.5. The overall flow of the Information retrieval evaluation process is shown in Figure 2.2.

2.3 TREC collection

The U.S. government's National Institute of Standards and Technologies (NIST)has run a large set of yearly workshop series called Test REtrieval Conference (TREC) to provide the infrastructure necessary for the large-scale evaluation of text retrieval methodologies. To improve the research in this area, TREC has provided large full-text documents and standardized the evaluation methodologies. TREC started in the year 1992, and since then the impact on the research in this area has significantly improved, and the effectiveness of the retrieval has almost doubled. With the development of TREC collections, the problems faced with the existing systems' capabilities and measurement techniques in evaluating operational systems were solved. The issue of developing a large dataset and evaluating the methodologies over the large dataset to evaluate these methodologies were big concerns for the researchers. This test collection provides gigabytes of test data, search statements, and expected results of the search results which helps the researchers to overcome those issues.

TREC collections are mainly used for evaluating the methodologies and models used in the evaluation process. Also used to select the best contributing systems, monitor, and evaluate system performance and effectiveness, evaluate query generation, and find out easy and hard topics in the matter of retrieving a greater number of relevant documents. It also helps to provide the inputs to the cost-effective analysis of information systems. Also, through the evaluation metrics score can determine the changes that need to be made to an information system for effective retrieval.

2.3.1 Overview of TREC versions

The Cranfield collection, created in 1960 contained around 1400 documents and 225 queries which made the researchers difficult to evaluate the retrieval systems effectively at later times. The main issue faced during those evaluation periods was the same set of documents with the same evaluation techniques which made the researchers difficult to compare the system's efficiency and also the techniques' efficiency due to time constraints. Text Retrieval Conference (TREC) overcame these issues and helped the researchers to do the research in retrieval systems using large dataset collection. 25 participated systems were there in TREC-1 with around one giga-bytes TIPSTER collection of topics and documents including training sets and test sets. The TREC-1 result can be considered as a baseline for future research with a large test collection (Harman,1993).

The TREC-2 conference occurred in 10 months less than the first conference. Many of the TREC-1 groups were managed to complete the system re-building and tuning by this time which helped to show better results compared to what happened in the TREC-1 evaluation time. In TREC-2, 150 people were involved with 31 participating systems. Large variation in results was shown by including methods like term weighting, natural language processing, and pattern matching. Considered TREC-2 result as a baseline for more complex experimentation (Harman, 1995). TREC-3 goal was to allow participating groups to freely devise their experiments within the TREC task. This includes manual or automated topic expansion, manually modifying the expanded topics, topic weighting, and passage retrieval. Also, at this conference extension of the English language to other languages especially Spanish. This helped the users with the query modification and achieved better recall scores (Harman, 1995).

In TREC-4, more tracks were introduced other than adhoc and routing, with different data and evaluation techniques, based on some specialized tasks. Five tracks were introduced as a Multilingual track mainly for non-English test collection, a Filtering track for evaluating routing systems, an interactive track, a database merging task, and a confusion track (Herman, 1996). In TREC-6, even though VLC tracks were introduced with 7.5 million text corpus, researchers started to investigate the possibility of collecting a test collection that reflects the aspects of Web searching. Also finding the accurate topic difficulty requires a set of relevant document lists which indirectly help to increase the retrieval effectiveness (Voorhees & Harman, 2000). Even TREC-7 also had a VLC track, officially web track was introduced in TREC-8. Web Track concentrated on two web tasks. Small and Large Tasks. The smaller one is made with 2 gigabytes with 250,000 document corpora distributed as a WT2g collection. The larger one featured later with 100 gigabytes with 18.5 million web pages. The purpose of Web Track was to find out how WT2g performs well for Adhoc Track and also monitor the effectiveness of the ranking of the search engines (Hawking et al., 1999).

Adhoc tracks were the main track over the eight previous tracks. In TREC-9, researchers realize that enough infrastructures exist to support the retrieval task, so from TREC-9 onwards, the ad-hoc track has been removed and included seven tracks like web retrieval, cross-language retrieval, spoken document retrieval, query analysis, question answering, interactive retrieval, and filtering (Hawking, 2000). TREC continues to grow, and many changes and updates occurred over the tracks based on the infrastructure availability. Some track goals were achieved and removed from the track list like the spoken document track. Most of the remaining tracks continued, but with some changes like Web Track included navigation topics, cross-language tracks have added documents with Arabic and French languages, and so on (Voorhees, 2000).

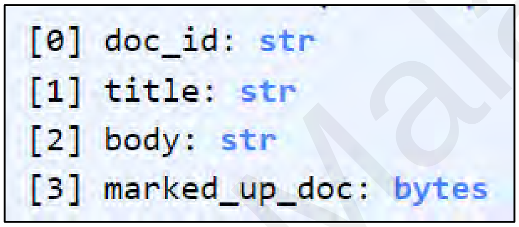
2.3.2 Test Collections

For a long time, text retrieval has been done with the help of test collections which are mainly used for retrieval experimentations. TREC continues this tradition, and it has a test collection that helps to evaluate retrieval systems and techniques efficiencies with large data collection. Test collection is an abstraction of the retrieval environment which consists of mainly three parts such as document corpus, topics, or queries, which means a set of information needs and relevant judgments, which shows an indication of which documents are relevant and must be retrieved based on the topics or queries.

2.3.2.1 Document Corpus

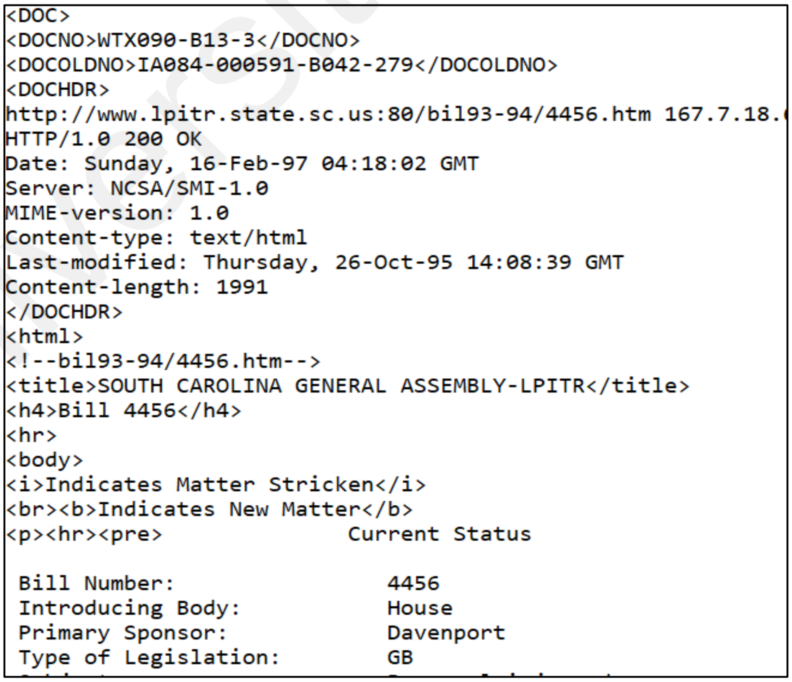
Document corpora consist of a large number of documents which consist of samples of texts that reflect the variety based on subjects, document formats, word choices, etc. These documents are used for the retrieval performance. Usually, these document sets are quite large. The earlier TREC datasets contain documents that are based on newspapers or newswire articles, some government documents like patents, and computer science abstracts. It was about 2 gigabytes of data. The document set used in various tracks varies with smaller and is larger depending on the requirements of those tracks.

As the tracks are getting increased, the document corpus sizes also increase. For example, the TREC-8 dataset, stored the document corpus in 5 CD-ROMs in which each disk contains compressed 1GB of data. NIST organization has tried to keep the originality of the data as it is without updating the contents. Each document consists of a document id DOCID, title of the content, body of the content, and dome marked-up details regarding the documents as shown in the figures, Figure 2.3 and Figure 2.4.



```
[0] doc_id: str
[1] title: str
[2] body: str
[3] marked_up_doc: bytes
```

Figure 2.3: Document type structure



```
<DOC>
<DOCNO>WTX090-B13-3</DOCNO>
<DOCOLDNO>IA084-000591-B042-279</DOCOLDNO>
<DOCHDR>
http://www.lpitr.state.sc.us:80/bil93-94/4456.htm 167.7.18.
HTTP/1.0 200 OK
Date: Sunday, 16-Feb-97 04:18:02 GMT
Server: NCSA/SMI-1.0
MIME-version: 1.0
Content-type: text/html
Last-modified: Thursday, 26-Oct-95 14:08:39 GMT
Content-length: 1991
</DOCHDR>
<html>
<!--bil93-94/4456.htm-->
<title>SOUTH CAROLINA GENERAL ASSEMBLY-LPITR</title>
<h4>Bill 4456</h4>
<hr>
<body>
<i>Indicates Matter Stricken</i>
<br><b>Indicates New Matter</b>
<p><hr><pre>                 Current Status

 Bill Number:                   4456
 Introducing Body:              House
 Primary Sponsor:               Davenport
 Type of Legislation:           GB
```

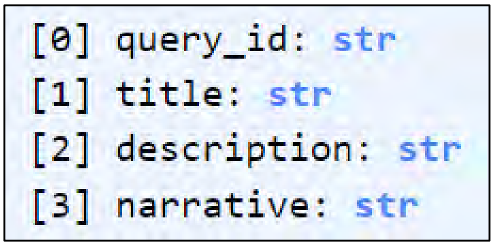
Figure 2.4: A sample of document from TREC-8 ad-hoc retrieval tasks

2.3.2.2 Topics

Topics are the statements of the information needed by the user and queries are the formatted data structure given to the retrieval systems. The TREC test collection provides a large number of topics that help to construct queries through various methods and also it provides some information regarding why this document is considered relevant. There is a traditional format followed by the TREC organization. Earlier versions of the topics were very detailed with multiple fields and concepts related to the topic subjects. From TREC-3 onwards, the concept-based contents were removed. But still, the accessors felt that the descriptions were too long. So other fields' contents also were reduced. From TREC-4 onwards, the description was made into one sentence of the information needs (Hawking & Voorhees, 1999). In later versions, only query identifier, title, description, and narrative only provided as shown in Figure 2.5.

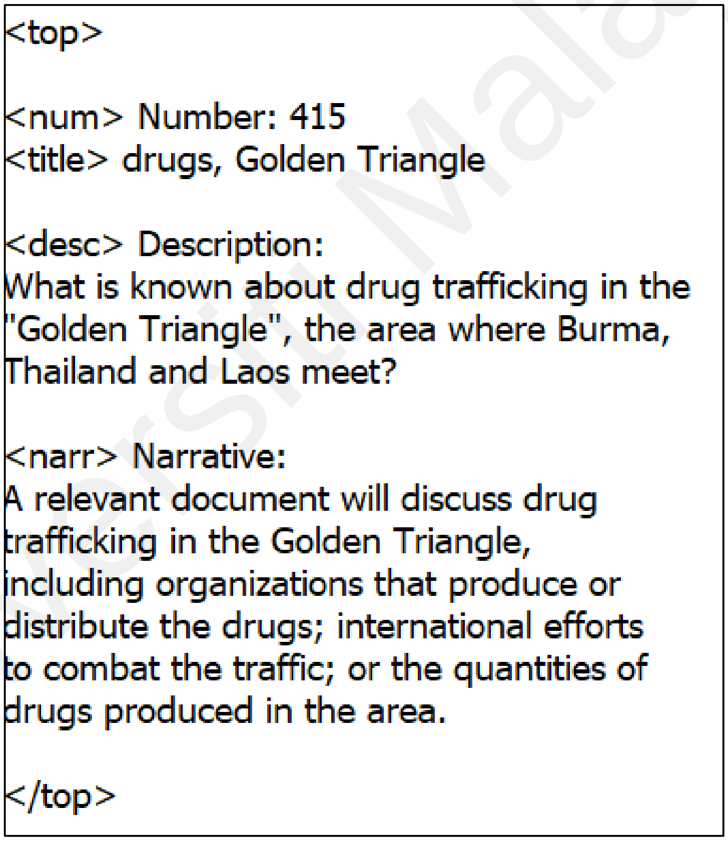
The query identifier is used as an identifier. The titles are designed to allow running the experiments with short queries. The title consists of three words, that neatly shortly describe the topic. The description is provided in one sentence which helps to describe the information needed by the user of the topic area. The description sentence contains all the words provided in the topic title. The narrative provides a small description of why the document is relevant to the topic. The sample structure of a topic is shown in Fig 2.6.

Participants were allowed to use any method of query creation either through automatic or manual methods. Automatic methods create queries from the topic statements without any human intervention. Manual methods consider manual query construction in the initial stage and reformulate the query based on the document set retrieved. Manual query construction is a very broad area. All these topics were constructed based on the accessor's interests and they are the same person who did the relevance assessment on what all documents are relevant to those topics. The NIST TREC team selects some topics from these lists based on the number of relevant documents retrieved and also considers the accessors' load balancing (Harman,1995).



```
[0] query_id: str
[1] title: str
[2] description: str
[3] narrative: str
```

Figure 2.5: Topic structure



```
<top>

<num> Number: 415
<title> drugs, Golden Triangle

<desc> Description:
What is known about drug trafficking in the
"Golden Triangle", the area where Burma,
Thailand and Laos meet?

<narr> Narrative:
A relevant document will discuss drug
trafficking in the Golden Triangle,
including organizations that produce or
distribute the drugs; international efforts
to combat the traffic; or the quantities of
drugs produced in the area.

</top>
```

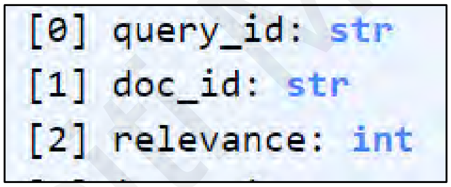
Figure 2.6: A sample of the topic from TREC-8 ad-hoc retrieval tasks

2.3.2.3 Relevance judgments

Relevance judgments are one of the main components of the test collection. Based on the topics and documents' relevancy, relevance judgments have been created. The retrieval task aims to generate a set of relevance judgments that can retrieve all the relevant documents and at the same time, suppress the irrelevant ones. TREC follows binary relevance judgment. 1, indicates the document is relevant to the topic and 0, indicates the document is irrelevant to the topic. The relevance judgment is created by the accessors, in which they are asked to create a report based on the topic they have chosen. If the document has any of the content which has given in the report, then this document will be considered relevant to the topic. The document relevancy is not chosen based on how many other documents have the same content (Harman, 1992).

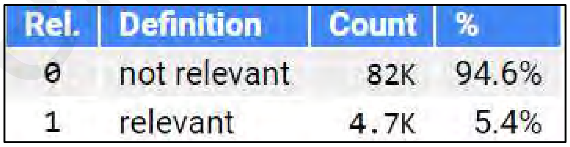
Judging a document based on topic content is subjective. It varies based on accessors to accessors. And also, the judgment varies for a document by the same accessor at different times (Linda, 1994). The relevance judgment of earlier test collections was complete due to the smaller size, and it was feasible. From TREC-3 onwards, as the dataset size is increasing, making judgments throughout the whole document corpus is a tedious process as the test collection has around 800,000 documents and relevancy checking of all these documents for one topic takes around 6500 hours. So, the researchers have used a technique called pooling (SparckJones & Rijsbergen, 1975) to collect a subset of documents for the judgment of the topic. The documents that have not been considered for the pooled list are considered irrelevant documents (Buttcher et al.,2007).

The pooled list of relevance judgments is created as follows. When the participants submit the runs to the NIST, they rank these runs in the order in which they prefer to judge the documents. NIST will consider how many runs need to be considered for the judgment process, and these documents have been sent for the evaluation process. Usually, the top 100 documents were considered for the evaluation process. So around 1/3 of the documents have only been considered and it became feasible for the accessors to make the relevance judgment set. The quality of the relevance judgment sets varies depending on the pool depth and the size of the topics considered for the evaluation process (Zobel, 1998). The relevance judgments are beneficial when the same test collections are used for future purposes. Also, the cost of recreating the relevant judgments can be avoided (Carterette et al., 2010). The relevance judgment list consists of topic id, document id, and relevancy identifier as shown in Figure 2.7, Figure 2.8, and Figure 2.9. Fig 2.7 shows the structure of the relevance judgment sets, Figure 2.8 shows the average list of relevant and non-relevant documents in the TREC-8 track.



```
[0] query_id: str
[1] doc_id: str
[2] relevance: int
```

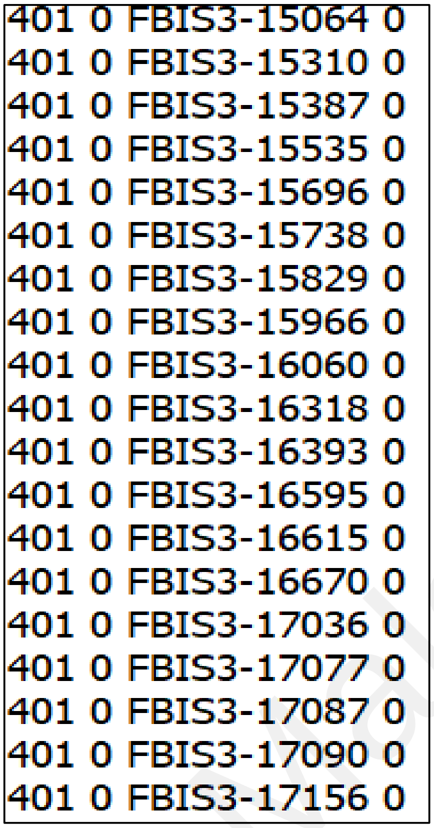
Figure 2.7: Relevance judgment structure



Rel.	Definition	Count	%
0	not relevant	82K	94.6%
1	relevant	4.7K	5.4%

Figure 2.8: Relevance judgment list contents based on TREC-8

0, indicates non-relevant around 94.6% of documents and 5.4 % of relevant documents based on pooled documents for each topic for all documents in the pooled list. Figure 2.9 shows the sample qrels sets which consist of topic id with all documents and show the relevancy.



```
401 0 FBIS3-15064 0
401 0 FBIS3-15310 0
401 0 FBIS3-15387 0
401 0 FBIS3-15535 0
401 0 FBIS3-15696 0
401 0 FBIS3-15738 0
401 0 FBIS3-15829 0
401 0 FBIS3-15966 0
401 0 FBIS3-16060 0
401 0 FBIS3-16318 0
401 0 FBIS3-16393 0
401 0 FBIS3-16595 0
401 0 FBIS3-16615 0
401 0 FBIS3-16670 0
401 0 FBIS3-17036 0
401 0 FBIS3-17077 0
401 0 FBIS3-17087 0
401 0 FBIS3-17090 0
401 0 FBIS3-17156 0
```

Figure 2.9: A sample of relevant judgment sets from TREC-8 ad-hoc retrieval tasks

2.4 Other Test Collections

For most of the information retrieval evaluation text collections are required regardless of its size, whether is smaller or larger. So many evaluation series have been run by the researchers to make the text retrieval evaluation easier. Most of the test collections are based on adhoc retrieval systems evaluations.

Cranfield test collections: In earlier times, Cranfield test collections were the pioneering ones, which helped to retrieve precise quantitative measures of information retrieval effectiveness. Comparing the effectiveness of the retrieval systems using different languages on a single document set with a set of topics, helps the researchers to develop the Cranfield paradigm (Cleverdon, 1967). This collection was developed in the United Kingdom in the late 1950s and it contains around 1500 documents with 225 topics and also a set of relevant judgment sets with topic-document pairs (Voorhees, 2019).

CLEF: Conference and Labs of the Evaluation form, (formally known as Cross-Language Education and Functions) has bought a substantial increase in the participating groups immediately after the TREC-8 series. CLEF has run mainly as a successor of the TREC-6-8 cross-language (CLIR) track (Braschler,2000). Multilanguage retrieval was the main aim of the CLEF, and it has had a greater impact on the researchers. As the CLEF series continues, CLEF 2006 tried to enhance the development of monolingual and cross-language textual retrieval systems. Monolingual tasks offered querying and finding documents in one language, Bilingual tasks offered querying in one language and finding documents in another language. Cross-language offers finding documents for difficult queries and it tried to conduct using expert participants (Nunzio (2007).

NTCIR: NTCIR is a series of evaluation workshops happening every and a half years and the aim of this workshop is to provide a large test collection that can enhance the information access technologies like information retrieval, cross-lingual information retrieval, text summarization based on both automatic and manual way, question answering and text mining. NTCIR workshop started in late 1997 with the aim of cross-lingual information retrieval including text stemming, and indexing. Earlier CLEF was only with English and its language became difficult for international information transfer in the Asian countries. Like performing CLIR between languages that have different data structures like English with Japanese, Chinese, etc (Kando,2004). This concern has been overcome in NTCIR, and it has attracted lots of international participants to this workshop. NTCIR has started with Adhoc, CLIR, and Term extraction tasks. The following workshops included Chinese, Japanese (Monolingual), and Text summarization tasks. Question answering, Patent, and Web Retrieval came in NTCIR -3 workshops (Kando,2007).

FIRE: Forum for information Retrieval Evaluation is a forum started based on TREC, CLEF and NTCIR. The main goal of this forum was to create a cross-lingual information retrieval mainly for the Indian Languages. This effort has developed based on a nationally funded project called Cross-Lingual Information Access (CLIA). The goal of this forum was to develop resources for a cross-lingual information access system between English and 6 other Indian languages. This forum started with two tasks as Adhoc Mono-lingual retrieval and Adhoc cross-lingual retrieval (Majumder et al., 2008).

INEX: INEX focuses on structured documents which can provide large text collections with structured documents, uniform evaluation measures, and also a forum for the organizers to compare their results. Research tracks have been included in this collection such as the Social Book Search track, Linked Data Track, Tweet contextualization, and Snippet Retrieval Track. Most of these tracks have used an XML version of the Wikipedia corpus and this forum was mainly concentrated on NLP researchers (Bellot et al., 2013).

2.5 Improving the accuracy of the Information Retrieval Evaluation process

The evaluation of information retrieval systems' performance is not only based on their efficiency but also their effectiveness. Effectiveness is calculated based on several relevant documents retrieved by the participating systems. The ability of the systems to retrieve as many relevant documents and at the same time suppress the irrelevant ones (Ferro, 2017). The main aim of information retrieval evaluation is to increase the accuracy of the information retrieval evaluation by increasing the quality of the relevant judgment sets and it can be achieved by increasing the number of relevant documents in the judgment sets.

2.5.1 Improving accuracy by considering the number of relevant documents in the judgment sets

Various studies and experiments have been done by the researchers to improve the quality of the judgment sets. These experiments helped to increase the accuracy of the evaluation process. Figure 2.10 shows the various methodologies proposed earlier by the researchers and the categorization of these methodologies. This literature review shows the depth of these categories, and each category of these methodologies was described in detail.

Information retrieval evaluation is a vast area. One way of evaluating these systems is to find the quality of the judgment sets and to find out ways to improve the quality of the judgment sets. Many ways were proposed by the researchers to improve the quality of the judgment sets such as pooling, human accessors, based on topics, and document similarity. Each of these methodologies was described in detail and the advantages and disadvantages of methodologies are described here.

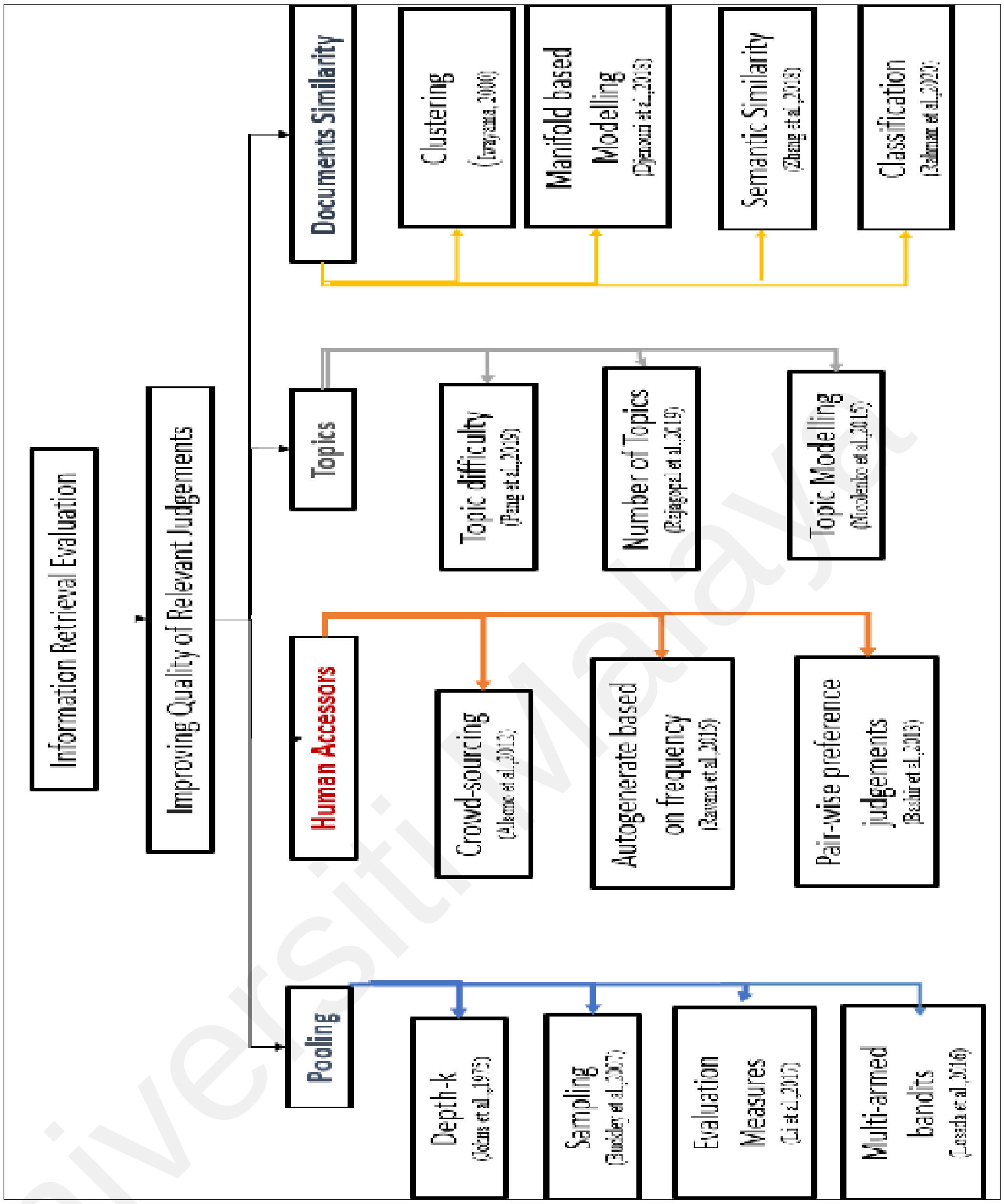
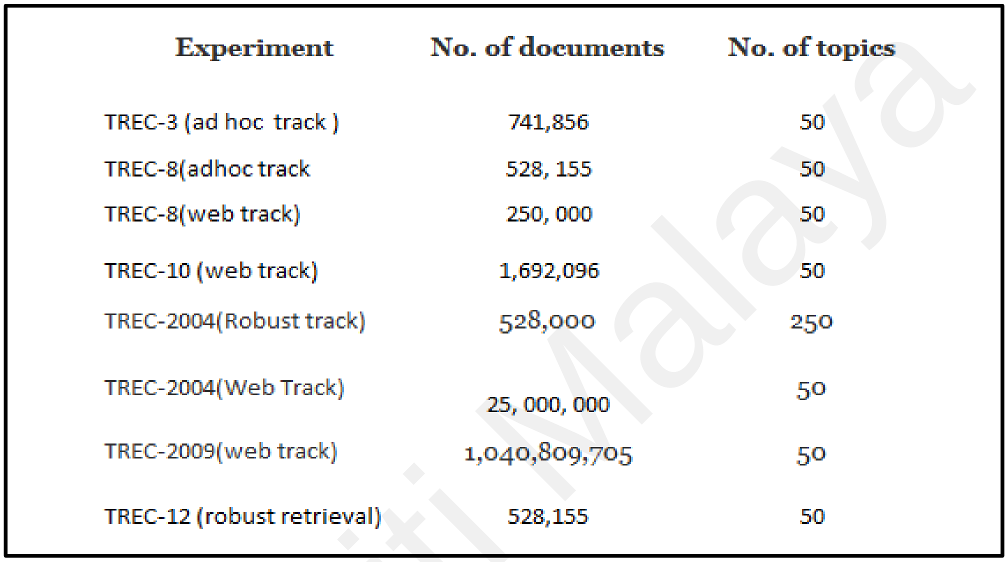


Figure 2.10: Categories of various methodologies are used to generate relevant judgment sets in the evaluation process.

2.5.1.1 Increasing relevant documents based on pooling

Finding relevant documents from the merged run list is costly and time-consuming. This process was done by the human accessors earlier who were experts in those areas. TREC test collection which is an initiative from the NIST organizers has provided a large collection of documents to do the large-scale evaluation of systems. Each TREC collection size is in millions and billions. An example of TREC collections with several documents and several topics is shown in Table 2.1.

Table 2.1: TREC collection samples



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

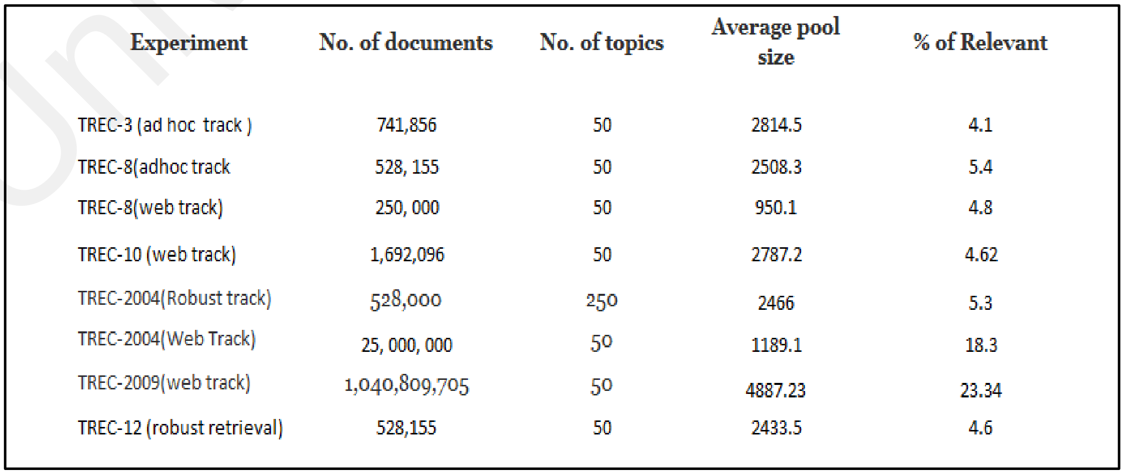
Judging such big collections through expert judges takes decades to complete the task and it's costly to afford too (Moghadasi et al., 2013). An alternative solution for this issue was the crowdsourcing. Intending to collect relevant documents from real users in the crowd-sourcing platform was the next choice (Tonon et al., 2015). But this method also had some limitations such as it was more error prone.

As a solution, (Spark Jones & Rijsbergen,1975) have proposed another technique called pooling, which considers only a subset of documents from the merged ranked list. It takes only top-k documents from each run created by the systems. As per the assumptions, all the documents in the pooled list are relevant, and documents that have not moved into the pooled list are considered irrelevant. The quality of the relevance judgment set is based on the pool depth chosen and also the retrieval methods used for the evaluation (Buckley et al., 2007).

The traditional and still popular first pooling method is depth@k, which considers the top k relevant documents from each topic from the runs created by the participated systems. All the duplicated documents were removed from the list and given to the human accessors for evaluation purposes. This method helps to reduce the size of the judgment list (Spark Jones & Rijsbergen,1975). The judged list will be called a partial relevant judgment set as it is considered only a part of the whole judgment list.

The traditional pooling method became very popular and helps to maintain the accuracy of the evaluation process, but the pool depth cannot be fixed to any chosen size. Pooling done with a fixed pool depth might fail to produce enough relevant documents. As the document size increases, pool depth also might need to be in-depth, which helps to maintain the quality of the judgment sets. But these results again affect the human accessors' effort, cost, and time. Table 2.2 shows some TREC document collections and various pool depths to achieve a certain percentage of relevant documents.

Table 2.2: 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 documents	No. of topics	Average pool size	% of Relevant
TREC-3 (ad hoc 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

To reduce those cost and effort, the number of judgments need to be reduced. An alternative option was the extraction of documents that have to be top-k and taking a 10% sample of the documents from the top-k list. These samples are given for the evaluation process (Buckley et al., 2007). Pooling based on evaluation measures was the alternative option to solve the large-scale evaluation and these used a methodology called Active sampling. A sampling strategy is used to find out the runs that hold the higher probability of the relevant documents and the ranking of documents done based on the sampling process. Later samples were retrieved from these runs which performed better which found out based on the evaluation measures such as the Horvitz-Thompson estimator. This estimator is used to evaluate the evaluation metric of all the runs (Li & Kanoulas, 2017).

Another pooling methodology is called the dynamic pooling method which repeatedly chooses documents from the unjudged list based on the documents from the judged list. This concept of choosing sampling was different from the pooling method but helped to retrieve more documents into the judgment sets. These samplings have been done based on meta-ranking and statistical sampling techniques. MFT, hedge, and bandits methods are some of the examples of these techniques (Cormack et al., 2018). Fair pooling is another way of doing pooling which is done by applying a fairness score, which creates a subset of pooled documents as similar as possible for all runs. In the same way, another pooling is called opportunistic pooling, which creates a subset of documents based on several judgments needed and based on a threshold value (Tonon et al., 2015).

Another methodology was based on rank-biased precision, Rbp, which identifies relevant documents based on fixed size N and fixed budgets. Rbp is a rank-biased precision that considers documents based on document rank probability and examines documents in turn, which move from one document to another. If the user prefers the i^{th} document, the probability of moving to the next document is i+1 (Moffat & Zobel,2007). Three methods are proposed by (Moffat & Zobel,2007) are Method A: RBP Abased@N, Summing Contributions, which considers documents to be 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 overall contribution to the pool and also weighting of the individual documents. Method C: RBP CBased@N, Raising the power, which tries to increase the score component by increasing the power of the current score. Based on common evaluation measures, three strategies were proposed by (Lipani et al., 2021). Those are Take@N, from Rbp runs, choosing top N documents. DCGBased@N discounted cumulative gain, which applied a discount function to rank documents into the pool. RRF@N, based on document contribution score finding the system effectiveness. PPBased@N calculates the number of relevant documents at rank k to the number of documents in k.

Another methodology includes the contribution of ordering the documents into the pool using a concept called Multi-armed Bandits. This method helps to identify most of the relevant documents into the judgment list or pooled list. This method was introduced by (Losada et al., 2016). This method helps to add more documents to the judgment list with minimal effort using the technique called k-armed bandit, which is an approach used to adjudicate meta-search documents (Losada,2018). Shallow pooling based on preference judgments, which is done by crowdsourcing helps to make more relevant judgments based on mean reciprocal rank and top-judged documents and re-evaluate these runs to reproduce more documents into the pooled list to increase the quality of the pooled list (Arabzadeh et al., 2021).

2.5.1.2 Increasing relevant documents based on human accessors impact

Human accessors' help in finding relevant documents has gained a greater impact on the information retrieval evaluation process. But every time it won't be feasible to get the human accessors help for the evaluation process especially if the test collection is quite large. Every time recreating the judgment list with the human accessors makes decisions differently in each occurrence with the same or different accessors. Disagreement among the accessors is one of the major issues noticed among the researchers during the evaluation process (Alonso et al., 2012). The next major issue was the high cost of utilizing these human accessors every time for each round of the evaluation process. Many studies have been done by researchers to reduce the cost by considering the documents only from a pooled list instead of evaluating the whole document list retrieved (Carterette et al., 2008), (Cormack., 1998). Also, another alternative solution was reducing the number of topics accessed (Prabha & Sridevi, 2019).

Crowdsourcing

To reduce the involvement of human accessors, the alternative solution was crowdsourcing. Crowdsourcing has a lot of advantages which mainly include replacing human accessors' help and through that can be cost-effective and flexible. The previous research shows that the disagreement between the users and human accessors is not so high if they work individually, but it is quite large if they work as a group. Sometimes it shows that crowdsourcing has produced better results compared to the expert judges. Results show that during the TREC collection evaluation process, the judgments were done faster with good results at low cost Alonso et al., 2012).

In literature, a big challenge among the researchers that have noticed is the agreement and disagreement between the accessors based on a topic. Document ambiguity or topic ambiguity might be the reasons for the disagreements. Some of the reasons can be that the terms in the documents might have different means, information in the query might not be clear, and accessors' or users' moods or environments all matter for the disagreement among them. Still crowdsourcing was a better option for the evaluation of documents with topic-document pairs compared to the assigning of relevance labels to the documents. Topic-document pairs have been collected from the multiple accessors and results have shown that the quality of the judgment sets has increased compared to the previous ones. Here relevancy depends on the distribution of documents and topic pairs among the accessors and not based on the absolute value assigned to the documents (Maddalena et al., 2017).

The agreement between the crowdsourcing and the expert judges has been studied based on different ordinal scales and different datasets based on the system's effectiveness and the topic's quality. Each scale result shows a similar score of the agreements with the ground truth and also shows the most accurate results for each topic level based on this scale. High correlation values show for both easy topics and system rankings. These scales help to get an idea of the various relevance scales or levels of the judgments (Roitero et al., 2021). Crowdsourcing is one of the major ways to collect relevant judgments on a scale. To scale these Information retrieval collections, around 100 to 100,000 workers were used. A new proposed methodology, based on topic set size, calculated using t-test and ANOVA helps to meet the predefined sets of statistical requirements. This can help to estimate the recommended number of accessors needed to judge statistical power and this estimation is dependent on the topic with the limited scale (Roitero et al.,2023).

Based on Frequency

However, crowdsourcing with large data set collection is always challenging. There can be a probability of a high chance of errors in the judgment process due to various disagreements, and issues in the indexing, searching, and even in the process of creating catalogs. The same word with different meanings might affect the quality of the retrieval documents and also the same way, different words with the same means might lead the accessors to choose the documents incorrectly and lead to a reduction in the number of relevant documents in the judgment set (Carpineto & Romano,2012). Pseudo-pseudo-relevance judgment process has been introduced to solve the issues faced during crowdsourcing. This methodology helps to reduce the human accessors' effort by generating a document ranking for the set of relevant documents. Pseudo-relevance judgments consider two important factors as frequency of each document for each run from all the systems runs and at the same time consider the document ranking. In traditional pooling, only pooled documents from the contributed systems were considered. But in this methodology, all the documents from all the systems such as contributed and non-contributed documents were considered (Ravana & Rajagopal,2015).

The magnitude estimation technique is an alternative solution for reducing the human accessors' effort. A scale measurement has been used for the estimation task that has been assigned to the crowdsource to obtain the judgments. This helps to obtain better results compared to classical binary relevance judgments. This estimation task helps to check the consistency of the ranking of documents mainly in terms of topic understandability based on the frequency of terms used in each topic. The results show overall better performance and a more robust evaluation of the relevancy of the documents (Mizzaro et al., 2017).

In some research, evaluation of system effectiveness using some existing methods which have been done by real users are more error-prone and have a big vary in the results compared to the expert judges. Based on the study of the existing methods, to get a better result, instead of depending on a single method, a combination of different best methods helps to get better results and is more effective when applied with machine learning algorithms. It has been done by finding the frequencies of the topic-documents pairs results from these methodologies helps to evaluate the system performance being run even without relevant judgment sets (Roitero et al., 2020).

Topics and topic terms have a greater impact on the quality of the judgments, even if it has done by accessors or groups of accessors. Much research has been done based on the quality of the topics and if it is found not relevant, it would be removed from the test collection. This methodology evaluates the system performance and quality of the topic based on the set of search terms and set of documents. The search terms are taken based on the user query. If the quality of the search terms goes below a threshold value assigned in the methodology, it considers these topics irrelevant and moves from the test collection. The results show that it helps to increase the quality of the retrieved documents with better results, and this can be achieved with the help of human accessors (Zhu et al., 2022).

Pair-wise preference judgments

All relevant documents need to be assigned a rank according to their relevancy. It will help to consider how one document is relevant over another document for a particular topic. Also, can create multiple grades of relevance. It can be done through pair-wise preference judgments or the nominal graded method. Accessors help are needed to judge the documents for both these processes. Accessors prefer pair-wise preference judgment as it requires only binary representation of marking as either relevant or irrelevant. The nominal graded method is needed to assign multiple relevance grades. Accessors can quickly assign the relevancy with pair-wise judgment instead of absolute judgment. So, it's mostly popular among the researchers. Pair-wise judgment uses the Elo rating system to combine or merge the documents (Bashir et al., 2013).

Another pair-wise judgment methodology used a technique to find a fixed number of relevant document pairs that are purely accurate and tried to auto-generate other document pairs similar to those pairs. It helps to generate a large number of preference judgments based on point-wise judgments. This technique helps to reduce the human accessors' involvement in all the documents and topics and also in the evaluation of system effectiveness (Roitero et al., 2022). Differences in the ranks also can be found based on top-ranked results by considering partial preference. It is done by taking top-k ranks of the documents and this process helps to increase the quality of the judgment sets (Clarke et al., 2021).

2.5.1.3 Increasing relevant documents based on topics

Topics play a major role in evaluating the system's performance. Different sets of topics generate different sets of relevant documents. Some topics generate better relevance judgments compared to the other topics. Finding the best topics that can produce more relevant judgments is always a difficult process among researchers (Breto et al., 2013).

Topic difficulty

The traditional approach to the information retrieval evaluation process is to retrieve the maximum number of relevant documents into the judgment list from the document corpus based on the topics (Pang et al., 2019). One of the main roles of predicting the relevancy of a document is based on the topic. One of the main challenges faced by the researchers is the topic difficulty. Based on the number of relevant documents retrieved, the topics can be classified as hard topics, medium topics, and easy topics. Human accessors always prefer to choose the easy topics compared to the hard topics. Due to the accessibility difficulty, relevant documents related to harder topics have not been chosen for the judgment list and it will affect the quality of the judgment sets. Topics can be used to compare the system's performance and through that system effectiveness. Various sets of topics with the same size of topics produce different results and at the same time same sets of topics with different sizes also produce different results (Berto et al., 2013).

Topics can be easy, medium, or hard based on their performance of retrieving quality documents, and based on this criterion the system performance score is assigned. Always human accessors prefer to choose the easy topics which helps to retrieve better results even with lesser pool depth. These harder topics which have the relevant documents and deeper pools might not been considered in the relevance judgment list. Average precision based on a topic can determine the difficulty of a topic. If the average precision of a topic is very high, then that topic is considered an easy topic and if the average precision of a topic is lesser, then the topic is considered a difficult topic for a particular system (Mizzaro, 2008).

The topic size based on the topic difficulty has a great impact on the system evaluation score. For researchers, it will be difficult to judge the topic pairs or topic combinations with a large topic size with large document collection which makes high computational cost and also time-consuming. So as an alternative solution first need to find out the topic difficulty of the topics in that run and find out the best sets of topics that have contributed to the pool and based on that can adjust the topic size. Most of the time easy topics work well with the judgment sets and it helps to increase the effectiveness score (Pang et al., 2019). Another study which shows that makes the earlier work easier by considering the top-k documents from both easier and harder topics and it has shown that even the harder topics also can perform well with the better results. Also based on the different evaluation metrics, almost all the results are consistent (Roitero et al., 2017).

The relevancy of the documents can be determined only after the returned search results. Each query can have different meanings which leads to choosing irrelevant documents in the judgment lists which makes the system performance down. This can be evaluated based on a criterion by incorporating document similarity concepts such as classification and clustering. Correlation coefficient values show the results that some queries are not performing well and it retrieves negative recall values. These queries can be found and can be removed from the evaluation process (Zhu et al., 2022).

Topic hardness is unpredictable sometimes. The same topics with the same sizes might produce different sets of relevant documents based on the participating systems. So, an alternative option is to topic ordering. Topic ordering can be done based on relevancy and needs to be done carefully to evaluate the prediction modelling (Culpepper et al., 2021). Another way of solving the topic difficulty can be done by a testing method with different document collection subsets and repeating these subsets over the same systems. For this evaluation, the same set of topics can be used. The results show that each system has retrieved a different set of documents in each run even though the effect is less (Zampieri et al., 2019). Another method of finding the topic difficulty can be estimated using NDC measures. Using this measure, assign a hardness score to the topics by considering the participating systems' performance. NDCG can even help to choose a particular set of topics which can produce a high recall value for the retrieved documents (Gienapp et al., 2020).

Number of topics considered or topic size

For researchers, the evaluation of the information retrieval systems is a challenging process as the document web collection is getting increased. Usually, the systems' effectiveness can be measured based on the quality of the topics that have been considered for the evaluation and also the number of relevance judgments produced. System effectiveness can be increased either by considering a particular set of topics or by a lesser number of topics or with good quality of the topics (Prabha & Sridevi, 2019). Done evaluation with a greater number of topics might get a greater number of relevance judgments, but the computational cost will be higher and also time-consuming. So most of the research based on topics might prefer evaluation with a lesser topic size and easy topic. It has been proven that even with lesser topics also can achieve better results and also can maintain good effectiveness scores (Carterette et al., 2008, Berto, et al.,2018). For that need to find out the hardness of the topics and also the best topic size. All the topics cannot be effective in retrieving better results. So finding the best topics is always a challenge. One of the effective methods of finding the best topics is with the earlier measures such as precision. Precision @k, was one of the choices, and the k value varies based on the size of the document collection (Dincer,2013).

For better results and at the same time, to reduce the computational cost with lesser topic sizes, much research has been done with lesser topic sizes and varied evaluation depth. Based on the effort-based relevance judgment, better results were achieved with lesser topic size with deeper pool depth or more topic size with lesser pool depth. Gaining the best results based on this concept became an interesting part among the researchers. The number of topics is completely dependent on the user's satisfaction. Accessors always prefer less effort topics for their easy access. Real users won't prefer much time over the hard topics, but expert judges might do it. Due to that many relevant documents won't get moved into the judgment list and it affects the system evaluation score. So, it has been noted that there is no correlation between system evaluation metrics and real users. So, considering low effort or easy topics with various evaluation depths will be preferable among the researchers to maintain the evaluation metric standardized (Rajagopal & Ravana, 2019).

Topic easiness is another study that determines the easiness of the topics, based on the real user's ability to understand the document, find out the relevant documents related to the topic, and understand the concept of the topic title and contents in the documents. The understandability, findability, and readability efforts have a greater impact on the evaluation score of system performance. Users always prefer to neglect the hard topics due to their difficulty in understanding the topic to judge the relevant documents. And they prefer easy topics for their easiness to make predictions. So as a solution, considering topics that are easy and have deeper depth always helps to maintain the quality of the judgment set (Prabha & Sridevi, 2019).

Another solution is to create a model of document topic combination based on the randomly generated clusters of documents. It helps to find out the document-topic pairs based on each system. It helps to reduce the number of topics to be chosen and at the same time can maintain the accuracy of the evaluation process (Voorhees et al., 2017). However, this methodology needs to be done carefully as the results show that the topic-document clusters produce different results on different systems. So the topic needs to be carefully chosen. Topic-document clusters produce different results compared to topic-alone evaluation and due to that evaluation scores also vary (Ferro et al., 2019). A sampling of the topics needs to be chosen carefully; otherwise, it affects the results of the accuracy. Awareness of the topics and balancing will help to choose the topics and passages effectively. It helps to maintain the result same even with different ordering (Hofstetter et al., 2021).

Topic Modelling

To evaluate large dataset collection topic modeling is an option. Topic modeling helps to choose the best subset of the documents and reduce the noise. The matrix factorization method can be used to generate the top modeling. For multi-lingual datasets can use topic modeling concepts for better results. Topic modeling can be used to create topics for even formal datasets, multi-model, or even multi-lingual datasets (Churchill et al.,2021). Different topic modeling using different datasets and different criteria produces different results in the view of accuracy. In this difference, choosing the best evaluation metrics is quite difficult because the accuracy varies based on the topic modeling samples (Rudiger et al., 2022). Also, it shows a clear view of various topic modeling techniques, and which one will be better for different content-based datasets. Also shows a clear view of the various evaluation metrics which can predict the accuracy of the evaluation clearly (Rudiger et al., 2022).

Topic interest among real users has great importance in achieving a good quality evaluation metric. So the aim is to find out which topics are chosen by the real users and through that can increase the system evaluation score. A model called topic modeling has been proposed to find out the interesting topics based on a criterion. This methodology helps to evaluate the quality of the topics and also it has been used in various applications like text classifiers, image classifications, and so on. Some pre-defined keywords can create the topic modeling. With this methodology, it can mine the best topics which can retrieve as many relevant documents as possible. Also, it helps to extract a quality metric based on topics that can predict the number of relevant judgments that can be given by real users about that particular topic (Nikolenko et al., 2017).

The drawback of topic modeling is it generates a set of topics and it may not be as accurate as what is preferred by the human judges. As a solution, a method of retrieving more relevant and accurate topics is based on the topic coverage. This technique computationally calculates the similarity of the topics with the list of topics from the reference list of expert judges. It helps to judge the models or methodologies and also topic quality. The results based on topic coverage show that it helps to find the topic quality, categories of similar topics, and topic model evaluation using various metrics (Korencic et al., 2021).

Considering a document at a time query evaluation algorithms have given a solution for the top-k relevant document sets efficiency (Crane et al., 2013). Learned sparse representations (LSR) is one of the retrieval methods that are used to generate lexical sparse representations of queries and documents in an index. LSR results are based on including document term weighting, query weighting, and document expansion and query expansion. This results in reducing the latency in the query expansion (Nguyen et al., 2023). Applying these two methodologies of the document at a time and score at a time helped to improve the results of system effectiveness by considering the retrieval models based on learned sparse representations. These representations can be used for retrieval model effectiveness (Mackenzie et al., 2023). Topic Modeling techniques have been used in machine learning to build models. A lot of studies have been done with Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). These approaches were mainly used to mine data from unstructured textual data. These approaches work by assigning a probability to each term in the document corpus. Several degrees of coherence were created with these two models and which model generates a higher degree of coherence will be selected. Results show that LDA performs better than LSA (Ateyah & Al-Augby, 2023).

2.5.1.4 Increasing relevant documents based on document similarities

For large document collection, evaluation using traditional techniques like pooling, sampling, and evaluation metrics are better methods to retrieve more relevant documents into the judgment sets. However, these methodologies are time-consuming and have high computational costs. This drawback can be overcome by using well-known techniques called clustering and classification. For pooling or traditional methodologies, the whole dataset needs to be used for the evaluation. However, for clustering and classification techniques, only the documents within the cluster or class need to be considered for the evaluation process. However, with the results, it can be concluded that the quality of the judgment sets is lesser here compared to the traditional methodologies. To overcome this issue, much research has been done using clustering and classification techniques.

Clustering

Much research has been done to increase the overall performance of evaluation using clustering techniques. Clustering can be done based on supervised and unsupervised algorithms (Taha,2023). Among the most popular ones is combining clustering with frequent itemset mining. Based on the relevancy of the documents from the merged ranked list, the documents were clustered based on their similarity by using k-means clustering. From each cluster, similar documents will be paired by calculating the frequency of the document terms. When a user query comes, the terms in that query will be notified and will calculate which cluster has these document pairs that match the most frequent terms. Patterns or more frequent terms matching clusters will be considered and those clustered documents will be moved into the judgment sets (Djenouri et al., 2018).

By combining both clustering and incremental relevance feedback, the search effectiveness can be increased. The relevance feedbacks work by collecting the relevance judgment list from two or three systems and considering it as an initial judgment list. These initial judgments have been sent back to all the other systems and the documents from those systems are based on these initial judgments. By incorporating document clusters with relevance feedback, the initial judgments were categorized based on relevance and irrelevance judgment list Then using clustering techniques, the documents from the other systems were retrieved and easily have separated relevant and irrelevant documents based on the initial judgment clusters. This helps to judge the relevant documents easily and can reduce the methods that focus only on a particular set of topics-based evaluations. In this clustering, all the documents are clustered based on the initial judgment feedback, not on the ranking of the documents. Based on density strategies, the best clusters will be found and the documents in it will be sorted by their relevance score. Top-k documents from these clusters will be considered for the evaluation process (Iwayama,2000).

One of the common approaches to the retrieval evaluation process is to cluster the documents based on user queries. These clustered documents are used for ranking the documents. At the same time, all the clusters are running to retrieve the similar features of these documents by considering the rankings. These document rankings based on various features helped to increase the document similarities in a vector space (Markovskiy et al., 2022). Another approach to improve the retrieval effectiveness based on clustering is by considering topic modeling and each topic in the cluster is evaluated with a set of terms in the document collections from the cluster and find out the frequency of each term that occurred in those documents. Then the topics with the same frequency are considered to represent various themes. Through this methodology, the results show that it helps to retrieve meaningful representations of clusters and also helps to predict the clusters' quality (Yuan et al., 2021).

Clustering the documents based on k-means to group documents have a greater impact in retrieving more relevant documents (Aliwy et al., 2022; Wang, 2021). K-means clustering based on its findability effort is another concept. Clustering the documents as relevant or irrelevant is done based on the effort needed to find the relevant documents. The results show that the performance of the participated systems varies when findability effort is combined with relevance in the system-based evaluation (Rajagopal et al., 2022).

Manifold-based

The inter-document similarity between the documents will be considered in this model called Manifold-based. While the inter-document similarity, the similar documents will be grouped, and new scores will be assigned to all the similar documents. When a new user query comes, the similarity between these grouped documents and the query will be found out and the group has the higher similarity, they consider all the documents in that manifold to be the same and assign a similar score to all the documents and consider that groups into the relevance judgment list. This model shows that both effectiveness and efficiency can be improved with in terms of relevant judgments. This methodology assigns new scores to the documents based on their relevance during the rank aggregation process. The weight matrix Z for the lower ranked documents (d $_{(1\ldots i)}$ gets compared with high ranked documents($_{a(1..j)}$ calculated with

$$Z_{ij} = \frac{\mathrm{sim}(d_i, a_j)}{\sum_{l=1}^{k} \mathrm{sim}(d_i, a_l)}.$$

where the more similar documents (d_i) and (a_j) get the higher weight for Z_{ij}. (Liang et al., 2018).

For pooled documents, to have a better-ranked score, passage-based, and manifold-based document similarity techniques help a lot. In Passage based model, scores have been assigned based on the weightage of the document term frequencies in that passage. In the manifold-based passage model, the term frequency is calculated based on the inter-document similarity by using term modalities. This technique helps to relook the scores in the pooled document list and can help to re-rank the documents with updated scores (Sharga et al.,2020). A rank-based manifold model has been developed to improve the efficiency of the clustering technique. Based on similarity measures, this model helps to create different clusters. An unsupervised similarity checking was done on the document clusters to compute the effective measures in the data collection manifold (Rozin et al., 2021).

Classification

Classification can be done just by classifying the documents based on their similarity is another methodology without considering pooling and system ranking. For each topic, a topic-specific document classification has been considered. This approach is done with an Active learning algorithm which selects the documents first and then classifies the documents based on their similarity. Active learning technique first considers a subset of documents which might selected by the judges. Based on this subset of documents, classify the documents that have not been considered for the pooling. This technique considers both document selection and also labelling of documents that have not been considered in the judgment sets. Comparisons of the subset selected and the documents that have not been considered in the pooled list help to improve the relevance score of the judgment list. However, this methodology might create biases in the evaluation process when considering the subset of the documents. As an alternative, the hybrid combination of human accessors and automated classification techniques has been considered (Rahman et al., 2020).

Usually, in many evaluation processes, only the pooled documents were considered for the evaluation process. The documents that are not in the pooled list are considered irrelevant documents and not considered for the evaluation process. Another methodology was to overcome some of these issues by training a classifier in the pooled list and based on those classified documents, the similar documents from the irrelevant sets were considered and if found a similarity, moved those documents into the judgment sets and it helped to improve the effectiveness of the systems (Buttcher et al., 2007). Another methodology was to find the similarity measures by calculating both frequency terms and sparse data in various dimensions proving that the classified documents performance score increased a lot. Classifying the documents based on frequencies of the terms is the first step and finding the centroids and creating a vector space model to classify the documents. Many evaluation metrics like precision, recall, and f1 score increased with this methodology (Eminagaoglu, 2022).

2.5.1.5 Summarize Findings and Gaps

The literature review aim was to find out the various methodologies that help to increase the number of relevant judgments in the judgment list and through that improve the quality of the judgment list and increase the effectiveness of the systems and accuracy of the evaluation process. The various methodologies considered were pooling, Human accessors, Topics, and Document similarity. One of the main concerns of the information retrieval evaluation process was retrieving the maximum number of relevant documents into the judgment list. Participated systems' performance is evaluated based on how many relevant documents have been retrieved into the judgment set by each system and based on that rankings are allocated for the participated systems. If more relevant documents can be retrieved, better ranks will be assigned through that system and indirectly it helps to increase the accuracy of the evaluation process. Research has used various methodologies to improve the accuracy of the evaluation process.

Pooling is one of the most traditional and popular methodologies considered among the researchers. This technique helps to reduce the computational time of evaluation by reducing the number of documents to be judged. In the pooling technique, the top-k documents from each run will be considered, and these documents will be given for the evaluation process (Spark-Jones, 1975). The results show that the quality of the judgment sets will be maintained almost the same as with the whole judgment sets. Various types of pooling proposed by the researchers based on the criteria such as sampling (Buckley et al.,2007), based on evaluation measures (Li & Kanoulas, 2017) (Moffat et al., 2007) (Lipani et al., 2021) (Tonon et al., 2015), dynamic sampling (Cormack et al., 2018), Multi-armed Bandits (Losada et al., 2016) (Losada et al, 2018) and Shallow pooling based on preference judgments (Arabzadeh et al., 2021).

Human accessors help to achieve a greater number of relevant documents was the other methodology considered by the researchers. Earlier the judgments of the retrieved documents were done by the expert judges. Evaluating the whole list based on the expert judges was not feasible and each time different decisions made by the different or same accessors made the evaluation process repeat every time. These disagreements make the evaluation process so costly (Alonso et al., 2012). So as an alternative option, crowdsourcing was opted to reduce the cost of human accessors. Crowdsourcing has been done with the help of real users. This methodology helped to reduce the computational cost and sometimes more accurate results too (Roitero et al., 2021). But this might create a topic ambiguity, which can be solved by using topic-document pairs (Maddalena et al, 2017). Also, another option to reduce the topic ambiguity was through pseudo-relevance judgments (Ravana & Rajagopal,2015), magnitude estimation, which means the frequency of terms in each term (Mizzaro et al., 2017) (Zhu et al., 2022) and a combination of different best methods described in (Roitero et al., 2020). Pair-wise preference judgment was another popular method (Bashir et al., 2013), Auto generating of pairs based on point-wise judgment (Roitero et al., 2022) and partial preference judgment (Clarke et al., 2021) were the different methods proposed by researchers to reduce the human accessors errors in evaluation process.

Topics have a greater impact on the system's performance. A different set of topics produces different retrieved documents. Finding the best topics to retrieve the greatest number of relevant documents is a quite challenge among the researchers (Breto et al., 2013). For that needs to know the topic difficulty or topic hardness (Pang et al., 2019). Based on the number of relevant documents retrieved, topic hardness can be identified. Based on this hardness, topics can be classified as easy, medium, and hard. (Mizzaro,2008). Most human accessors prefer to choose easy topics (Pang et al., 2019). However, some research shows that there is no evidence that easy topics can retrieve more relevant documents, even harder topics also can (Roitero et al., 2017). Classification and clustering techniques applied based on retrieved web snippets help to find the relevancy of the documents (Zhu et al., 2022). Prediction modeling and topic modeling (Zampieri et al., 2019) (Culpepper et al., 2021) and evaluation metric NDCG's impact (Gienapp et al., 2020) show the testing of topic hardness.

Several topics to be considered were another challenge faced by the researchers. More topics always increase computational cost. The main aim is to reduce the number of topics and at the same time need to choose the best topics (Carterette et al., 2008, Berto, et al.,2018) (Prabha & Sridevi,2019). Sometimes it has been shown that hard topics also produce better results. So the choice was choosing easy topics with lesser depth and hard topics with deeper depth were given better results (Rajagopal and Ravana,2019). Topic easiness was another study that concentrated on low-effort topics (Prabha and Sridevi., 2019). Partitioning documents into different parts and system-topic pair combinations gives better results (Voorhees et al., 2007). However, this partitioning gave different results (Ferro et al., 2019) and topic topic-aware-balancing method was a solution to it (Hofstetter et al.,2021). Topic modeling has been used for large document collection. Topic modeling works differently based on different datasets and different techniques based on various evaluation metrics (Churchill et al., 2021) (Rudiger et al., 2022). Topic interest based on real users using predefined keywords and classification techniques helps to extract topic quality metrics that can predict better judgments (Nikolenko et al., 2017). Topic coverage (Korencic et al., 2021), query evaluation algorithm (Crane et al., 2013), Learned Sparse Representations (LSR) (Nguyen et al., 2023) (Mackenzie et al., 2023), Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) (LSA (Ateyah & Al-Augby, 2023) have given better results based on topic modeling.

Document similarities between the documents helped to reduce the computational cost and time compared to the pooling method. Various types of clustering methodologies proposed such as frequent item set mining based on the k-means algorithm (Djenouri et al., 2018), based on relevance feedback cluster the documents into categories (Iwayama,2020), based on user query cluster documents in different similarity vector space (Markovskiy et al., 2022) and topic modeling with clustering (Yuan et al., 2021).In Manifold-based, inter-document similarity will be considered and a new score assigned based on the similarity found (Liang et al., 2018) combination of both passage-based and manifold-based (Shraga et al., 2020) and rank-based manifold (Rozin et al., 2021) also have given better results. Classification of the documents based on human and automatic judgments (Rahman et al., 2020), trained pooled classifiers to predict unjudged documents (Buttcher et al., 2007), and combination of term frequencies with sparse data in different dimensions on classified documents (Eminagaoglu,2022) given better results based on document similarities.

Judgments through pooling help to retrieve more relevant documents and because of that quality of the judgment set also increases well. Also, as it considers only the documents from top-k, it is cost-effective, and less time is required for evaluation purposes. However, the drawback of the pooling was, that the documents selected, top-k documents from each run will be considered as relevant documents and only these documents are considered for the evaluation process. The documents that have not moved into the pooled list are considered irrelevant and these documents will not be considered for the evaluation process. These documents are called unjudged documents. There might be relevant documents over the unjudged list, but due to some systems inefficiency, these documents were ranked lesser and not moved into the pooled list (Losada et al., 2018) (Cormack et al., 2018). Figure 2.11 shows the overall flow of the pooling techniques.

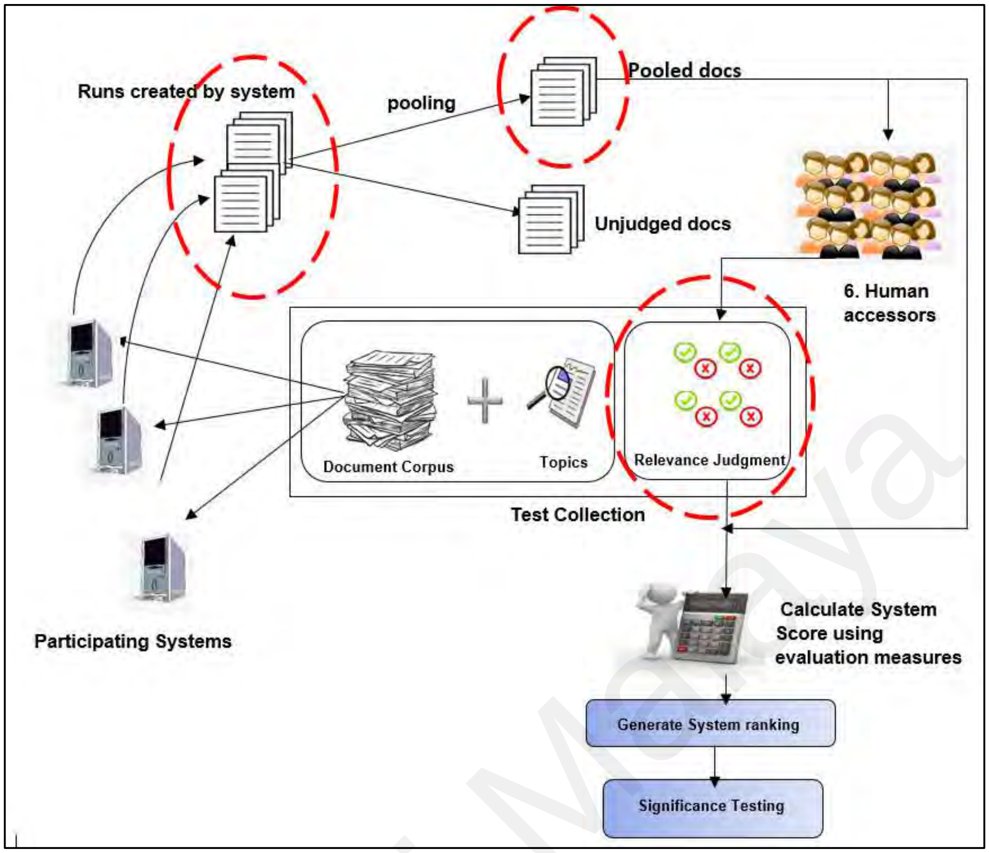


Figure 2.11: Pooling methodology with pooled documents and unjudged documents

For large data collection, document similarity through clustering and classification is faster than the traditional approaches. It's the cluster or class that has a high similar score that is only considered for the evaluation process. However, the drawback here is that the quality of the documents retrieved through clustering and classification is often lower than the traditional methods. Also, the documents within the class or cluster consider it as same and assume it with a similar score. Whenever a new query comes, mostly the document similarities are considered based on the term frequencies. Finding the best cluster or class based on this similarity is a quite challenging process (Djenouri et al.,2021, Djenouri et al.,2018, Rahman et al.,2020). Figure 2.14 shows the overall flow of document similarity using clustering and classification techniques.

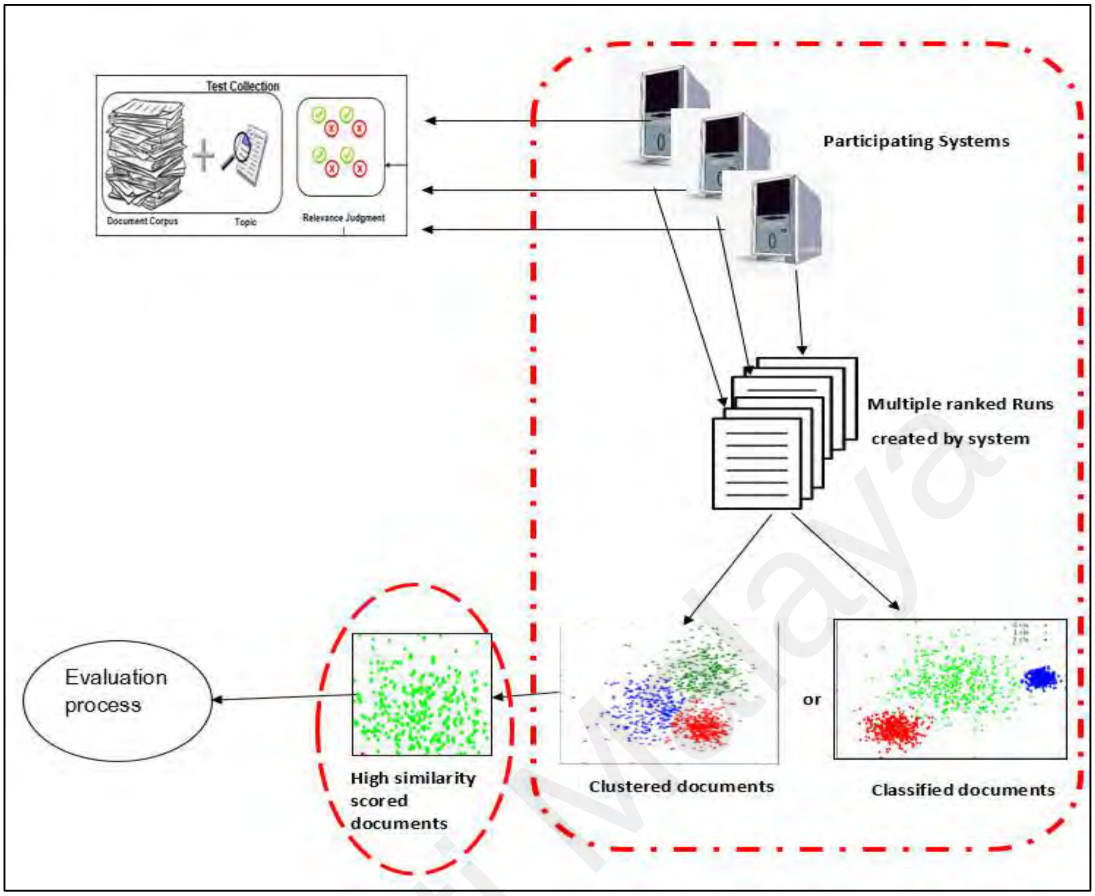


Figure 2.12: Document similarity methodology using clustering and classification technique

At the same time, judgment through human accessors is more error-prone. For a particular topic, judgments made by real users and expert judges are sometimes can be different. It depends on their readability effort, findability effort, and understandability effort and it varies from time to time based on different accessors or the same accessors (Rajagopal & Ravana, 2019). Also, topic hardness has an important role in judgment. Human accessors prefer easy topics compared to hard topics because hard topics might result in different output compared to easy topics. The hardness of the topic is quite a challenge among the researchers (Pang et al., 2019). Many studies have been done to solve the topic of hardness issues and human judgment errors. More research is needed in document similarity and pooling methodology, so this research mainly focuses on these two methodologies and tries to overcome the limitations on these to some extent. The limitation mainly focused on partial relevant judgment, which means most of the relevant documents have not moved into the pooled list. And mainly focuses on how to increase the quality of the judgment sets based on document similarity with the help of pooling.

2.5.2 Improving accuracy by considering test collections

Test collections have a high impact on the quality of the relevant judgment sets. However, the evaluation of information retrieval evaluation systems is a challenging process as the information is getting added to the Web. The main components that have a role in evaluating systems are the topics and the efficiency of the participating systems. In many research, researchers in the information retrieval field have tried to achieve better evaluation accuracy with fewer topics and lesser relevance judgments (Voorhees & Buckley,2002) (Culpepper et al., 2014). This session aims to find out the influence of topics and participating systems in the quality of the relevance judgment in a better way.

2.5.2.1 Improving accuracy based on topics from test collections

Among the test collections, topics have an important role in the system performance. Each subset of documents from the same runs might produce different sets of relevance judgments. Some topics generate a quality judgment list compared to other topics. Finding the best topics that can produce more relevant documents helps to reduce the computational cost (Breto et al., 2013).

Topic Size

For researchers, the evaluation of the information retrieval systems is a challenging process as the documents in the web collection are getting increased. Usually, the effectiveness of the system can be measured based on the quality of the topics that have been considered for the evaluation and also based on the topics and the number of relevance judgments produced. System effectiveness can be increased either by considering a particular set of topics or by a lesser number of topics or with good quality of the topics (Prabha & Sridevi, 2019). Done evaluation with a greater number of topics might get a greater number of relevance judgments, but the computational cost will be higher and also time-consuming. So, most of the research based on topics might prefer evaluation with lesser topic size and easy topics. It has been proven that even with lesser topics also can achieve better results and also can maintain good effectiveness scores (Carterette et al., 2008, Berto, et al.,2018).

The cost of generating relevance judgments is proportional to the number of topics in the collection. One solution was reducing the number of topics required for the evaluation. Each topic has significant importance in the effectiveness of the system. Many studies have been done to find out how many topics need to be included in the test set to achieve a remarkable result. If choosing randomly, how many topics are required to obtain statistical results which have to be reliable? Many discussions have been done among the researchers to choose the best topic sizes. In the early stages, the topic size was 250 and later it was reduced to 75(Sparck Jones et al.,1976). Later it was noticed that even with 50 topics also maintained the quality of the judgment sets (Voorhees et al.,2002).

The results have shown that as long as the topic sizes vary, the evaluation scores also vary (Voorhees & Buckley, 2002) (Buckley & Voorhees, 2000). Later it have been proven that even though test collections had many topics, even with 25 topics can achieve almost similar results (Sakai, 2006). Also, it has been shown that considering statistical significance other than topic size difference has a better result and is more reliable (Sanderson & Zobel, 2005). Also, they have proven that 25 topics are less reliable and that 50 topics can produce reliable results.

These all works lead to a conclusion on the choice of effectiveness metrics. Some metrics can retrieve similar accuracy even with a lesser number of topics. Many studies have been done with graded relevance judgments and results show that measures are stable and accurate (Sakhi,2007). Also, it has been shown that Average Precision is better than Precision at 10 among the evaluation metrics to show a better result (Sakhi, 2006) (Webber et al., 2008b). Paired t-test, one-way ANOVA, and confidence intervals were used to design the topic set sample size. These metrics required topic-by-run score matrices from past test collections to determine the performance of each system population variance for each evaluation measure (Sakhi 2016).

Several studies have been done to find out the most effective way to reduce the computational cost. One among them is to reduce the topic size or choose the best topics for the evaluation process. But the choice was, to collect many relevance judgments for fewer topics, Narrow and Deep (NaD judging) or few relevance judgments for many topics, Wide and Shallow (WaS judging) (Carterette et al., 2009). Best and worst subsets in a "bottom-up" approach describe the topic set reduction approach which proves that some sets of topics or topic subsets can retrieve more relevant documents compared to others (Mizzaro & Robertson,2007) (Guiver et al., 2009). Also "top to bottom" approach shows better evaluation using many queries judged shallowly and fewer queries in detail (Carterette et al., 2009b). Previous works have not considered the cost impact of judgment with deeper depth, the judging speed. WaS were not considered topic construction time (Voorhees,2006). By considering all these drawbacks, an intelligent topic selection algorithm has been proposed based on learning to learning-to-rank method and these methods helped with better topic selection and better judgment (Kutlu et al.,2018).

For better results and at the same time, to reduce the computational cost with lesser topic sizes, much research has been done with lesser topic sizes and varied evaluation depth. Based on the effort-based relevance judgment, better results were achieved with lesser topic size with deeper pool depth or more topic size with lesser pool depth. Gaining the best results based on this concept became an interesting part among the researchers. Always the number of topics is completely dependent on the user's satisfaction. Accessors always prefer less effort topics for their easy access. Real users won't prefer much time over the hard topics, but expert judges might do it. Due to that many relevant documents won't get moved into the judgment list and it affects the system evaluation score. So, it has been noted that there is no correlation between system evaluation metrics and real users. So, considering low effort or easy topics with various evaluation depths will be preferable among the researchers to maintain the evaluation metric standardized (Rajagopal & Ravana, 2019).

Earlier in TREC evaluation, systems were evaluated by accessing the set of topics, and based on these topics, the effectiveness of the system was evaluated. Generally, Average Precision was used to measure the system performance with every topic, and the mean score of every topic was considered to rank the systems (Sanderson,2010). However, this method doesn't compare the system performance on different collections. Systems that performed well on one collection might not produce the same on the other collection. Each topic is important as the topic's results are highly variable. If the results of topics are almost the same, then the retrieval process would be easier and more reliable too. The use of standardized scores is one way to reduce topic variance (Webbar et al., 2008). It calculates per-topic standardization scores that scale the scores to nearly 0.5 and sets the minimum scores to 0 and the maximum to 1. Another simple linear transformation of the score over the non-linear has been proposed by (Sakhi, 2016). It calculates the topic standardization score falls within the range of 0.05 and 0.95 and it was more accurate to predict the topic scores. Standardization score was introduced to reduce the errors in topic variability. However, the results show use of standardization scores does not reduce the score deltas, especially in the comparisons with paired significance tests (Vorhees,2019).

Compared to this, an empirical distribution as an alternative shows that these standardized ranking methods are done topic by topic, comparisons have been done on different scales altogether, and the standardization scores to make statements about new scores (Urbano et al., 2019). Correlation analysis has shown the relationship between evaluation measures and interval-scaled versions, overcoming most of the standardization scores issues faced earlier on the topic of performance on the system effectiveness (Ferrante et al., 2021).

Topic easiness is another study that determines the easiness of the topics, based on the real user's ability to understand the document, find out the relevant documents related to the topic, and understand the concept of the topic title and contents in the documents. The understandability, findability, and readability efforts have a greater impact on the evaluation score of system performance. Users always prefer to neglect the hard topics due to the difficulty in understanding the topic to judge the relevant documents. And they prefer easy topics for their easiness to make predictions. So as a solution, considering topics that are easy and have deeper depth always helps to maintain the quality of the judgment set (Prabha & Sridevi, 2019).

Another solution is to create a model of document topic combination based on the randomly generated clusters of documents. It helps to find out the document-topic pairs based on each system. It helps to reduce the number of topics to be chosen and at the same time can maintain the accuracy of the evaluation process (Voorhees et al., 2017). However, this methodology needs to be done carefully as the results show that the topic-document clusters produce different results on different systems. So, the topic needs to be carefully chosen. Topic-document clusters produce different results compared to topic-alone evaluation and due to that evaluation scores also vary (Ferro et al., 2019). A sampling of the topics needs to be chosen carefully; otherwise, it affects the results of the accuracy. Awareness of the topics and balancing will help to choose the topics and passages effectively. It helps to maintain the result same even with different ordering (Hofstetter et al., 2021).

Topic Hardness

Determining topic hardness is an important feature in the information retrieval systems evaluation as it has a great impact on the system rankings. TREC-6 evaluated the topic difficulty based on the view of the human accessors and later it was proven that it did not correlate with the computational difficulty based on the evaluation of search results (Voorhees & Harman, 1997). So the topic hardness or topic difficulty have not to be biased. Not too easy and not too difficult to maintain the reliability of the test collection (Eguchi et al., 2002).

To identify the actual topic difficulty or hardness, a median of the average precision has been considered by many research and it shows the actual retrieval effectiveness of the systems. Using TF-IDF, topic terms were considered for the evaluation, and categorized the topics into easy, medium, and best. System ranks of top runs have significantly changed the topic hardness, but still, the difference in the total ranking is not as significant as the results of statistical tests. These results prove that topic hardness or topic difficulty has an impact on the systems rankings (Eguchi et al., 2002). Categorizing the topics into two groups for the best and worst queries gives a great difference in the average precision will be reflected in some attributes and easier to determine the prediction compared to the random sampling. The best combination of inverse document frequency and term frequencies average for the topic title can predict the average of 1/3 or to ½ of the correct best and worst topics (Kwok, 2005).

If a topic based on an information retrieval system is easy, means for that particular topic, it is easy to determine if the document is relevant or non-relevant. Then the participating system performs better on that document, the system gets a boost on its overall effectiveness which is equal to the same rating it would get if those systems can perform well on the difficult topic. GMAP or Geometric Mean Average Precision has been used to gain more weight on the topic effectiveness scale (Mizzaro,2008). Another methodology has been used based on the mean or maximum average precision. The topics have been split up into four groups based on their difficulty. The splitting was done not based on the number of relevant documents found on each topic. It measures the correlation of each group's results with the full set and has noticed that the easiest topic groups have a higher correlation than the harder topics. This experiment was mainly done on the multilingual sub-task (Mandl,2009).

Always easy topics won’t be able to retrieve effective results in the consideration of the system's effectiveness. Easy topics help to differentiate the best and least effective systems but from the view of the best effective system, the topics were not easy with the evaluation of geometric mean average precision (Roitero et al., 2017). Based on the permutation algorithm, topic difficulty was explored and results show that topic difficulty is variant and needs to be careful when relying on the ordering of the topic when evaluating the performance of the evaluation models. Also, it helps to measure the system performance with the topics and query formulations on the unique pooled documents (Culpepper et al., 2021).

Apart from the Mean or Average of average precision, another measure is based on the topic difficulty score using the best average precision. The topic difficulty score methodology gives the conclusion that the evaluation performance goes down due to a sub-collection of topics. It shows the results that hard topics are hard for all the systems and hard for all the other document subcollections (Hu et al.,2003). In other work, topics were classified based on the topic difficulty scores. The low median score for the difficult topics and high median scores for the easy topics along with an outlier (Voorhees, 2003). Another approach to estimate the topic hardness was based on the NDCG topic difficulty score. This method is based on the ratio of the NDCG score to the pooled judged documents. It is more stabilized than the outlier systems too (Gienapp et al., 2020).

2.5.2.2 Improving accuracy based on participated systems from the test collections

The web collection is getting added in real-time and it causes the lack of inconsistency in performing the information retrieval evaluation process. As the data collection is getting increase many relevant and irrelevant documents are getting added. At the same time, many proxies also getting added to the data collection in the matter of topic titles, missing documents, topic descriptions, and unrelated titles (Rasmussen,2003). Due to that bias happens to the systems during the pooling process and only particular systems documents were considered for the judgment process and it affects the evaluation of information retrieval systems.

Generating relevant judgments from large test collections like TREC, with the help of human accessors is time-consuming and more error-prone throughout the judgment process (Smucker and Jethani, 2012). Relevance judgments created by human accessors are time-consuming and highly costly. Still, these documents are not feasible every time as each time these human accessors might produce different judgment decisions at different times with the same accessors or different accessors. Disagreement among the accessors is one of the major issues noticed among the researchers during the evaluation process (Alonso et al., 2012). The next major issue was the high cost of utilizing these human accessors every time for each round of the evaluation process. Many studies have been done by researchers to reduce the cost by considering the documents only from the pooled list instead of evaluating the whole document list retrieved (Carterette et al., 2008), (Cormack., 1998). Also, another alternative solution was reducing the number of topics accessed (Prabha & Sridevi, 2019).

To reduce the involvement of human accessors, the alternative solution was crowdsourcing. Crowdsourcing has a lot of advantages which mainly include replacing human accessors' help and through that can be cost-effective and flexible. The previous research shows that the disagreement between the users and human accessors is not so high if they work individually, but it is quite large if they work as a group. Sometimes it shows that crowd-sourcing has produced better results compared to the expert judges. Results show that during the TREC collection evaluation process, the judgments were done faster with good results at low-cost Alonso et al., 2012). However, crowdsourcing with large data set collection is always challenging. There can be a probability of a high chance of errors in the judgment process due to various disagreements, and issues in the indexing, searching, and even in the process of creating catalogs. The same word with different meanings might affect the quality of the retrieval documents and also the same way, different words with the same means might lead the accessors to choose the documents incorrectly and lead to a reduction in the number of relevant documents in the judgment set (Carpineto & Romano,2012).

The magnitude estimation technique is an alternative solution for reducing the human accessors' effort. A scale measurement has been used for the estimation task that has been assigned to the crowdsource to obtain the judgments. This helps to obtain better results compared to classical binary relevance judgments. This estimation task helps to check the consistency of the ranking of documents mainly in terms of topic understandability based on the frequency of terms used in each topic. The results show overall better performance and a more robust evaluation of the relevancy of the documents (Mizzaro et al., 2017). pair-wise judgment methodology used a technique to find out a fixed number of relevant document pairs that are purely accurate and tried to auto-generate other document pairs similar to those pairs. It helps to generate a large number of preference judgments based on point-wise judgments. This technique helps to reduce the human accessors' involvement in all the documents and topics and also in the evaluation of system effectiveness (Roitero et al., 2022). Differences in the ranks also can be found based on top-ranked results by considering partial preference. It is done by taking top-k ranks of the documents and this process helps to increase the quality of the judgment sets (Clarke et al., 2021).

Building test collection judgment sets are expensive based on human accessors. At the same time, pooling with fixed pooled size failed to produce the judgment sets as expected due to the document collection size increases. This results in generating biased judgment sets and due to that the correct system effectiveness cannot be measured. It is due to unfairly ranking the documents in the systems and failing to predict the system behavior correctly (Buckley,2006). Many methods have been introduced by the researchers to overcome this issue of incompleteness judgments. It includes alternative strategies to increase the number of relevant documents in the judgment sets by constructing new judgment sets. One method is to increase the pool size of each query and apply simple regression on each query number of new relevant documents found on each pool depth. This technique helps to achieve more relevant documents with the given effort and increases the reliability of the systems (Zobel,1998). The next method considers topics with a greater number of relevant documents and judged them based on different pool depths. It helps to increase the effectiveness of the systems evaluation (Jayasinghe et al., 2014). Some observations were based on the sub-topics. Topics subtopics might be relatively similar to one another. In the case of irrelevant documents, the subtopics might be widely different content. A language model has been used to find the documents possibly related to the subtopics of a query. Based on the results, gain values were calculated. Calculating the gain value in different ways and aggregating them into a single measure helps models generate the effectiveness of the judgment list (Hui et al., 2017). The relevance feedback method to prioritize the documents in the pooled list is another method that indirectly selects which documents from the document collection are relevant to the human assessments. It helps to retrieve more relevant documents with lesser assessment effort in the relevance judgment set (Otero et al., 2023).

The next approach was to reduce the number of judgments needed for the evaluation of the retrieval systems. Two different ways in which those judgment documents were selected. The first method was static selection, which choose a set of documents in advance based on the scoring. If a document is considered with a high score by multiple retrieval systems, that document will be considered on the judgment list. Scoring functions have been used to find the documents with high potential among the system runs. The second method is dynamic selection, which chooses the documents based on the previously completed judgments. These judgment lists decide which document should be chosen next (Moffat et al., 2007).

For larger test collections, the documents in the pooled or relevant judgment sets are incomplete. This concept might be true with the small document collection. By applying various evaluation metrics, the system's effectiveness can be evaluated. However, due to this incompleteness, the results produced by the evaluation measures were varying. Different evaluation measures have different evaluation criteria concerning how much it is correlated to user satisfaction. Most of the evaluation measures are derived from precision and recall. Precision at 100 documents finds the number of relevant documents in the top 100 ranked list of each topic. However, it has been proven that the error rate is higher than the mean average precision. Mean Average Precision is the mean of the precision scores after each relevant document found in the relevance judgment sets. It's a stable measure compared to other measures.

Based on the evaluation metrics, many studies have been done to improve the incompleteness in the relevance judgment sets. Preference judgment is one of the most impacts ones. The preference is based on a particular topic. It helps to prevent the accessors from choosing one particular category of the document (Frei and Schauble, 1991). A new measure called bpref helps to find out the fraction of judged non-relevant documents retrieved over the relevant documents. It has shown that MAP and bpref system ranking are almost equivalent and an average of these two measures agree to retrieve the better system. Only judged documents with better help to build more relevant documents in the pooled list (Buckley and Voorhees,2004).

But the drawback of the bpref is that judgment becomes more incomplete as long as it keeps on evaluating, at it deviates from the average precision score. It affects the average precision value and also the ranking of the systems. Average precision is considered a gold standard in the matter of incomplete judgments (Yilmaz and Aslam,2007).Evaluation measures with average precision such as induced average precision which considers sample documents from the pool and calculates the average precision once all the unjudged documents are removed. Subcollection average precision considers the samples and also considers the samples from the unjudged document list. Inferred average precision considers the average precision of all the documents from the pool and have noticed that inferred average precision has achieved better results compared to the whole document pool. This value is almost similar to the actual Average precision (Yilmaz and Aslam, 2006).

2.5.3 Improving accuracy by considering incompleteness in judgment sets and the biasness in ranking

There might be many cases in which the judgment sets might be incomplete. For example, When the document collection is dynamic, documents are added to the document collection (document corpus) over time. During that time, judgments can become a smaller subset when compared with the whole document collection. At that time, the judgment set became incomplete. Also. for large test collections, only by pooling technique with depth-100 cannot retrieve all the relevant documents into the judgment list. As an alternative solution, judging all the documents in the system runs will become costly too. So only by using the pooling technique to find out the relevance judgment from large test collections always becomes incomplete or biased (Eguchi et al., 2002).

In many cases, the system's effectiveness cannot be measured due to this incompleteness. One of the main reasons for the incompleteness is when the evaluation depth becomes greater than the pool depth. When evaluating depth=1000 and pool depth=100, many relevant documents might not be listed in the pooled list. If the pool depth==200, is compared to pool depth=100 more relevant documents might move into the judgment sets. However, if the pool depth increases, the evaluation cost also increases. At the same time, if a participating system fails to retrieve a relevant document into the pooled list, the contribution of that document to the system effectiveness scores becomes zero. It happens when most evaluation approaches consider that document as irrelevant (Yilmaz and Aslam,2006).

If all the relevant documents have not been moved into the judgment sets, then the judgment sets are considered incomplete or biased to the pooled systems (Webber et al., 2010). Much research has been done to overcome this issue. Pooling is based on non-uniform distribution over different sampling probabilities (Aslam et al., 2006). Judging the documents based on relevant nuggets of information helps to automatically generate relevance judgments (Pavlu et al., 2012). Another method chose only a subset of topics to be accessed completely, not only pooled by ignoring irrelevant ones. The pool of topics is randomly chosen and it helps to correct the biasness of the un-pooled systems (Webber & Park, 2009). Other finding shows that a lack of participation systems leads to biasness in pooling. So merging the pooled and unpooled documents helps to reduce the bias by applying a linear combination of ranks of each document to each run. The effect is measured based on the precision values and helps to increase the quality of the unjudged documents (Lipani et al., 2015).

Due to the incompleteness in the relevance judgment sets, two different ways in which research has been done to manage the issue of relevance judgments. They are metric adjustments for incompleteness and document selection for relevance judgments.

2.5.3.1 Metric adjustments for incompleteness

Metrics play an important role in comparing the evaluation of the system's performance. Different metrics perform differently and evaluate different aspects of effectiveness. The most common evaluation metrics are precision and recall (Robertson et al., 2010). Metric adjustment means either the development of new effective metrics or enhancement in the already existing ones. The standard evaluation measures such as average precision and R-precision are not robust to the incompleteness relevance judgment. That’s why a new measure, bpref, was proposed by them which is highly correlated with average precision when the relevance judgments are complete and more robust to incomplete relevance judgments (Buckley and Voorhees,2004). Many studies have used this bpref metric (Sakai,2007) (Sakai and Kando,2008) and it has proven that NDCG and AP to condensed list have given better results than bpref (Sakai, 2007) (Bompada et al., 2007).

Many measures like Q-measures, Normalized Discounted Cumulative Gain (NDCG) or AveP to condensed list can be obtained by filtering out all the unjudged documents from the core ranked judgment list which indirectly helps with a better solution for the incompleteness than bpref. Also, applying the graded relevance to obtain helps to have a better result based on the incompleteness, and through that Q-measure and NDCG have also proven that it helps to reduce the biasness (Sakai,2007). Another metric RBP, assumes, measures the probability of the user moving from one document to another ranked document. If more relevant documents are in the judgment sets, it increases the RBP value (Sakai and Kando, 2008).

Take the samples of the system runs and take the approximate values of average precision and score variance can be evaluated based on the number of samples (Aslam and Yilaz,2007) (Aslam et al., 2006). Adjusting the scores based on the degree of the biasness against the unpooled documents is one way to reduce incompleteness. For this not only based on the documents (pooled and unpooled) but all systems are also evaluated against all topics and find out which all topics in the unpooled document list have created biasness in the unpooled systems were found and adjust the scores on these topics-based documents (Webber and Park, 2009).

Many metrics are worked on with average precision. Average precision is a system-oriented evaluation metric that has top-heavy bias. It works with probabilistic interpretation. But when it plots with a precision-recall curve, the biasness can be viewed easily. It has a good performance ranking function when it works with learning to rank as its objective. Due to that most of the metrics use multi-graded relevance judgment concepts, especially in the area of learning to rank to optimize for ndcg. Based on the relevancy of the documents in the judgment list, a relevancy score is assigned based on the NDCG. This score is considered as a gain returned to the accessors based on a relevant document. Another metric called Gap graded average precision is a continuation of Average precision that helps to reduce the biasness by using multi-graded relevance judgment (Robertson et al., 2010).

Most of the information retrieval evaluation tasks are conducted based on the precision evaluation measure. Most of the information retrieval tasks are evaluated using a standard evaluation metric called Mean average precision. The evaluation based on patent retrieval is a wide research area in which its main objective is to find the relevant documents in the runs as early as possible by assigning higher ranks to the relevant documents to reduce the human effort. Patent retrieval is generally considered as a recall-oriented task but reassigning the score estimation with mean average precision helps to achieve a better patent score for the retrieved documents and helps to increase the quality of the judgment set (Magdy and Jones,2010). Some other score-based evaluation metrics helped to overcome the issues related to incompleteness judgment such as inferred average precision, infAP. This metric helped to show the robustness of the incompleteness judgments, and they are focused on the precision-based results (Magdy and Jones,2010).

Sampling techniques help to estimate the information retrieval metric of the incomplete judgments. As per this method, the relevancy of a document can be estimated based on a particular topic, but with a cost. By considering the annotation of the samples of the corpus, choosing documents to accurately estimate the metric scores. The fairness of the system ranking can be evaluated based on this score (Kirnap et al., 2021). The incompleteness can be categorized based on the assumption of the non-relevant documents, relevancy prediction, score estimation issue, and condensed list. Also, the issue is in estimating the gain value of the unjudged documents. Also, the error bound of unjudged documents also the scoring in the condensed list. Also, the main concern with the systems is judgments were done without omitting the irrelevant ones. These issues can be solved by adjusting the pool depth and evaluation depth based on the various evaluation metrics (Lu et al.,2016). Previous research shows that point-wise evaluation has a lot of limitations and is based on preference metrics. For highly effective systems, it shows that metric evaluations are more effective for offline evaluation. However, these evaluation measures reliability in the presence of errors in the judgment sets needs to be verified. These errors might be due to the human judges who provide the judgments as assessments of the relevance of a given document to a specific query. Random qrels flip strategy helps to system orderings and system-versus-system significance testings help to reduce the judgment error rate (Rashidi et al.,2023)

2.5.3.2 Document selection for relevance judgments

To increase the quality of the relevance judgment sets, researchers found an alternative way by considering documents as nominees for the relevance judgment. These have been done by either changing the order of the documents in the run list or by creating the test collection. Adjusting the order of the documents for the judgment based on the effectiveness (Moffat et al., 2007). Score adjustments based on the bias on the unjudged document list is another solution. Biasness estimation has been done with a leave-one-out method based on the pooled document list and adjusting the score based on it (Webber and Park,2009). One method is by adjusting the documents based on the ranks of the participating systems helps to increase the effectiveness. The document from the good systems helps to produce better effectiveness on relevance judgments compared to poor systems. The results show that the documents were ordered based on good systems help to achieve better performance (Sanderson,2010). Sampling distributions and variance-optimizing strategies help to make the system pair-wise comparisons and this strategy helped to achieve better results with some evaluation measures (Schnabel et al., 2016). Document adjudication for pooling-based evaluation helps to early identify relevant documents in the pool with the help of multi-armed bandit problem models. This model is one of the best adjudication strategies (Lozada et al., 2016). The rank fusion approach based on rank fusion models with the distribution of retrieval scores plays a critical role in creating samples by combining multiple search results (Lozada et al., 2018). For the pooled documents, Many document prioritization methods have been proposed earlier to get better-pooled documents to reduce the human accessors' effort. However, how many relevant documents can predict a better-pooled list was not researched much at that time. The diversified stopping method helped to determine when to stop making relevant judgments have studied and results show that these methods help to reduce 95% of the accessors' effort (Lozada et al., 2019). Documents are selected based on pseudo-relevance and based on prioritization. Otherwise, randomization is based on the level of accessors. The pooled document quality can be increased by examining the level of accessor quality, inter-accessor agreement, similarity of the system ranking, and systems robustness based on the topics (Sakai et al., 2023).

However, most of these methods show that the process of selecting documents based on relevancy failed to produce enough relevant documents and through that failed to produce test collection properly. It is mainly because most of the topics have too many relevant documents and the pooling technique fails to retrieve all these documents into the judgment sets. Assessors have to create new strategies and tools to create test collections based on the Cranfield paradigm by reassigning the criteria of “truly relevant” to the desired participated systems (Voorhees et al., 2022). Most of the metrics based on graded judgments restrict the relevant documents into the judgment sets, preference-based judgments help to retrieve more relevant documents, and the test collection standard will increase based on it and it can be easily reusable (Clarke et al., 2023).

2.6 Evaluation Metrics

Information Retrieval evaluation uses various evaluation metrics to find out the participated systems' effectiveness. Metrics help to assign scores to the systems based on how many relevant and non-relevant documents are retrieved based on a specific query given by the user. Many metrics are available for both quantitative and qualitative evaluation (Dalianis,2018). Some quantitative-based evaluations used in my research are recall, precision, average precision, mean average precision, normalized distributed cumulative gain, and rank-biased precision.

Precision and Recall

Precision and Recall are the dependent terms for the evaluation. Precision is calculated based on number of documents retrieved. Precision is calculated based on several retrieved relevant documents, R for a query Q, and total number of documents retrieved, D. Precision is calculated then as R/D. Recall is based on the relevant documents in the collection. To evaluate the performance of the systems by considering the relevant documents, a recall metric needs to be used. Recall is calculated based on several relevant documents retrieved, R with a total number of relevant documents in the document collection, C. Recall is calculated as R/C (Sanderson and Zobel,2005) (Arora et al., 2016). These two metric terms calculations are shown here. Equation 2.1 indicates Precision and Equation 2.2 indicates Recall

Equation 2.1

$$\boldsymbol{Precision} = \frac{|Relavant\_Retrieved|}{|Retrieved|}$$

Equation 2.2

$$\boldsymbol{Recall} = \frac{|Relavant\_Retrieved|}{|Total\_Relavant\_Collection|}$$

For example, if the participated system identifies 40 documents from 100 documents, in which 35 are relevant and 5 are irrelevant base on a topic, Precision, and Recall are calculated as,

Precision=35/40=0.875 (indicating 87.5 % were relevant)

and

Recall = 35/100=0.35(indicating 35% of relevant documents were found)

Pooling techniques help to obtain several retrieved documents, so precision is considered a more effective metric compared to recall. Pooling concepts considers only a subset of the documents for the evaluation process. The effectiveness of the participated systems was calculated based on a cut-off rank, P@k. k indicates the top k documents from the run list. This technique calculates the effectiveness based on the relevancy of the document based on its ranking as shown as r_i in Equation 2.3. r_i assumed as a binary relevant judgment.

ie.,

$$r_i = \begin{bmatrix} 1 & \quad if\ the\ document\ is\ relevant \\ 0 & \quad if\ the\ document\ is\ irrelevant \end{bmatrix}$$

Equation 2.3

$$P@k = \frac{1}{k}\sum_{i=1}^{k} r_i$$

P@k can be interpreted as an average of values, r_i ….. r_m .The cut-off rank score get varies based on the k values as it depends on the ranking of the documents based on the query given by the user.

Average Precision

Average precision is defined as the mean of precision score obtained after each relevant document is found and assigned zero as the precision core if the relevant document is not retrieved (Kishida,2005). The quality of the retrieved document list based on the query given by the user matters the sequence in which the relevant and non-relevant documents are retrieved and the total number of relevant documents for that given query. Average precision varies based on the quality of the retrieved document list (Aslam et al., 2005). Average precision is calculated based on the precision value of the document and also the relevancy of the document for a particular topic. The average precision is shown in Equation 2.4.

Similar to precision, Average Precision also has cut-off ranks to make the score higher. If enough relevant documents are not there in the retrieved list, the cut-off rank can be increased. Average precision

Equation 2.4

$$\boldsymbol{Average\ Precision, AP} = \frac{\sum_{i=1}^{k}(P(k)\ X\ relv(k))}{R}$$

Here k indicates the cut off rate or top k documents, P(k) indicates the precision of k documents and R indicates the total number of relevant documents for that query. relv(k) indicates the relevancy of the document for that topic or query. relv(k) shown in binary representation such as

$$relv(k) = \begin{bmatrix} 0 & if\ P(i)\ is\ 0 \\ 1 & otherwise \end{bmatrix}$$

For example, from a retrieved list, top k=10 documents have considered. Out of these 10 documents 6 documents are relevant. The average precision calculated shown in below Table 2.3.

Table 2.3: Example of calculating Average Precision based on one query

DOC ID	RELEVANCY	PRECISION SCORE
D1	R	1
D3	R	1
D2	NR	0.66
D6	NR	0.5
D12	R	0.6
D8	NR	0.5
D4	R	0.57
D9	R	0.62
D10	NR	0.55
D7	R	0.6

Average precision of these documents is calculated based on

$$\text{AP@10} = \frac{(1+1+0.6+0.57+0.62+0.6)}{6} = 0.731$$

For all the non-relevant documents, the relevancy score is 0. The Average precision score increases based on the number of relevant documents on the top of the list.

Mean Average Precision

The average precision is calculated based on the topics. The performance of a system effectiveness based on several topics is calculated using Mean Average Precision (MAP) (Voorhees, 2007). Mean Average Precision is calculated based on the mean of all the average precision of topics over all the topics. The calculation shows in Equation 2.5.

Equation 2.5

$$MAP = \frac{\sum_{t=1}^{N} AP(t)}{N}$$

Where, t indicates the topic, AP(t) indicates the average precision of that topic and N represents the total number of topics in the system runs.

For the evaluation of the systems effectiveness, mean average precision is an effective metric which is used by the researchers. MAP retrieve the results based on the number of relevant documents retrieved and it is precision biased (Harman, 2011).

Table 2.4: Example of calculating Mean Average Precision

Topics	**Average Precision**
540	.69

541	.47
542	.02
543	.04
544	.83
545	0.31
546	0.11
547	0.30
548	0.09
549	0.80
550	0.47

An example of calculating Mean average precision is shown in the Table 2.4. Each topics average precision is shown the table. The average precision value will vary based on the topics performance and mean average precision is calculated based on the system performance over all the topics in that system runs. The Mean average precision is calculated as follows.

$$MAP = \frac{(0.69 + 0.47 + 0.02 + 0.04 + 0.83 + 0.31 + 0.11 + 0.30 + 0.09 + 0.80 + 0.47)}{11}$$

$$MAP = 0.37$$

Normalized Discounted Cumulative Gain (NDCG)

NDCG is also an evaluation metric used to measure how well a participated system works not only by considering how relevant the results are but also considering how good the order is. NDCG is considering how well the participating systems ranked the documents in the runs. Mean average Precision only calculates the precision, whether the documents are relevant or not. However, it is not considering how relevant the recommended results are.

NDCG shows how the graded relevance affects the results of information retrieval evaluation. Gain is calculated based on the relevancy score of each document. The relevancy score is calculated based on two factors such as highly relevant documents are more valuable than marginally relevant documents. The greater the ranked position of the relevant document, the less valuable it is, as the user will be less likely to examine the document (Kekäläinen,2005).

Gain is calculated based on the four-scale assessment such as 0 for irrelevant documents, 1 for marginally relevant, 2 for fairly relevant documents and 3 for highly relevant documents (Sormunen, 2002).

For example,

$Gain,\ G' = \langle 3, 2, 3, 0,0, \ldots \rangle$ (Järvelin & Kekäläinen, 2002)

Cumulative Gain, CG is calculated as sum of gains of the first k items recommended. Its calculated by summing from first position until kth position.

$$\boldsymbol{CG[i] = \begin{cases} G[i], if\ i = 1 \\ CG[i-1] + G[i], \quad otherwise \end{cases}}$$

For example, calculation is as follows:

$Cumulative\ Gain,\ CG' = \langle 3, 5, 8, 8, 8, \ldots \rangle$ (Järvelin & Kekäläinen, 2002)

Discounted Cumulative Gain, DCG assign relevance score based on the position. The position of the relevant document is higher, the less valuable it for the user, because user less likely to go through the documents due to the effort. The relevance documents on the top gets the higher rank and lower ranked relevant documents gets the lower rank.

$$\boldsymbol{DCG@k = \sum_{i=1}^{K} \frac{G_i}{\log_2(i+1)}}$$

Or

$$\boldsymbol{DCG[i] = \begin{cases} CG[i]\,, & if\ i < b \\ DCG[i-1] + \frac{G[i]}{b_{log\, i}}\,, & if\ i \geq b \end{cases}}$$

The simple way of doing the discounting this requirement is to divide the document score by log of its rank.

For example, b=2 . So DCG is calculated as

$$Discounted\ Cumulative\ Gain,\ DCG' = \langle 3, 5, 6.89, 6.89, 6.89, \ldots \rangle$$

(Järvelin & Kekäläinen, 2002)

NDCG is normalized discounted cumulative gain is the DCG with a normalization factor in the denominator. The denominator is the ideal DCG score, when most relevant documents come first. NDCG calculation shown in Equation 2.6

Equation 2.6

$$\boldsymbol{NDCG@k = \frac{DCG@k}{IDCG@k}}$$

were,

$$IDCG@k = \sum_{i=1}^{K} \frac{G_i}{\log_2(i+1)}$$

For example,

$$Normalized\ Discounted\ Cumulative\ Gain,\ NDCG' =$$

$$\langle 1, 0.83, 0.87, 0.78, 0.71, \ldots \rangle$$

(Järvelin & Kekäläinen, 2002)

NDCG can be used to apply to a condensed list, means ranked list of documents obtained by removing all unjudged documents from the original list. It is a simple solution to evaluate the incompleteness of the relevance data (Sakai, 2007a).

Rank Biased Precision

One of the limitations in the metrics such as average precision and recall were lacking in evaluating the user's behaviour. This drawback has been overcome by the rank biased precision (RBP). RBP is a metric that measure the utility gained based on the user persistence. In RBP metric, a parameter named *p* represent the user behaviour on probability to proceed to the next rank (Moffat & Zobel, 2008). It shows how the user examining each document from the top of the list and will proceed to the next document with the probability of *p*, or finishes the searches with the probability of *1-p* (Park & Zhang, 2007).

RBP is calculated as shown in Equation 2.7

Equation 2.7

$$RBP(p) = (1-p) \sum_{i=1}^{d} r_i \cdot p^{(i-1)}$$

where $r_i \in [0,1]$, relevance judgment of i^th element and (1-*p*) is the factor used to scale the RBP measure within [0,1]. If the user is with low persistence, (close to 0) means user not likely to examine after the first document, and high persistence, (close to 1) means users might examine many documents.

For an example, with p=0.5,

ranked relevant judgment list shown as (1,1 ,0,1,?, 0, 0, 1). Here 1 indicates relevant, 0 indicates irrelevant, and ? indicates not judged.

RBP (0.5) = (1-0.5) X ($0.5^0 + 0.5^1 + 0.5^3 + 0.5^7$) =0.816 (Park & Zhang, 2007)

2.7 Statistical Significance Tests

One of the main aims of the information retrieval researchers is to find out the best retrieval methods which can increase the effectiveness of the participated systems. Given two participating systems, during the evaluation process we need to find out the best system which can perform better on a particular retrieval method. TREC based evaluation follows the typical way of collecting a set of documents, create a set of queries and find the relevant judgements based on it and measure the effectiveness based on the evaluation metrics (Voorhees & Harman, 2005).

Sometimes, there will be some noises in the evaluation process. Such as topic difficulty, judges' behaviour, document collection limitations, and all. These noises sometimes produce false positive results. Statistical significance tests help to overcome this drawback by finding out the better retrieval methods or techniques that can perform truly better instead of by chance performing well. Statistical significance test provides information about whether the observed evaluation score difference is really meaningful or not (Hull, 1993).

Two categories of significance testing are there. Parametric or model-based test and non-parametric. Parametric based testing involves precision and recall, they make a number of specific assumptions about the distribution of measurements and their errors. Non-parametric-based testing includes Student's paired t-test, Wilcoxon signed rank, sign test shifted bootstrap, and randomization tests (Smucker et al., 2007). A significance test consists of a statistical test to judge the two systems based on the difference in any one evaluation metric, mostly Mean Average precision. Second, it consists of a null hypothesis. Thirdly a significance level that determines the system performance based on this value is either above or below the significance level (Box et al., 2005). All the above-mentioned significance tests have their criteria and null hypothesis. All these tests aim to measure the probability of the similarity in the performance of the two systems are same or occurring by chance only (Smucker et al.,2007).

The null hypothesis, $H_{o,}$ or the initial assumption of the statistical testing will be that all the retrieval methods' performance will be the same. The testing aim is to disapprove this hypothesis by assigning a p-value. This p-value will be considered as the probability that the difference could occur in the performance of these methods (Hull, 1993). Before the testing a significance level will be assigned, indicated with the symbol α. Usually, α value will be 0.05 (Andrade, 2019). If the p-value is less than the significance level, the null hypothesis can be rejected, and if the p-value is greater than the significance level, an alternative view of estimation of the likelihood of the two methods merely be different (Hull, 1993). The interpretation of these significance levels and p-values is explained in (Andrade, 2019).

In Fisher's Randomization test, a null hypothesis is created as two systems are identical and have no effect over each other on their Mean average precision. There is a total 50 topics were there, so 2^{50} ways to label the results under the null hypothesis. Randomization test measures the differences between two systems' Mean average precision in each permutation. If 2^{50} permutations were created, it could measure the number of times the differences in MAP would be greater or lesser. This number divided by 2^{50} would give an exact p-value. This kind of test is known as the randomization test or permutation test (Smucker et al.,2007), (Basu, 2011).

The Wilcoxon Signed Rank test also has the same null hypothesis as of Randomization test. However, in randomization test, the test can use any test statistic, but Wilcoxon must have a specific test statistic. This test takes the difference in the paired score and ranks them in ascending order based on the value. The minimum sum of both positive and negative rank is the test statistics. The Wilcoxon test statistics throw away the true differences and replace its with a magnitude of differences. This helps to make the computation easy and distribution of rank sums. This helps to determine the p-value (Zimmerman & Zumbo, 1993), (Smucker et al.,2007).

As same as the above two tests, the Sign test also have a null hypothesis which states that the two systems are similar and have no effect on each other. Sign test statistics is the number of pairs of one system will be better than the other system. It has a binomial distribution with the number of trials being the total number of pairs. This distribution is obtained by counting the number of successes in 2^{50} permutations of the scores of 50 topics. However, the Wilcoxon and sign test were not preferred by the researchers due to test statistics (Smucker et al., 2007),(Parapar et al., 2020).

The Student t-test measures the difference in the mean of the two systems. The null hypothesis states that both means are equal whereas the alternative hypothesis states that one will be greater or lesser than another or both are not equal. Three types of t-tests are there. One sample t-test, independent samples t-test, and paired samples t-tests. In One sample t-test, measures the mean value of a sample is statistically the same or different from the mean value of the parent data from where this sample was taken. For this mean value, standard deviation, and t-value are used for the calculation. One sample t-test will be used if the sample size is less than 30 (Winter, 2019). The independent t-sample test is an unpaired t-test that determines the means of two unrelated groups. Two categorical variables and one normally distributed variable are used for the test. Significance level is calculated based on mean, standard deviation, and number of observations. Paired t-test measures the two dependent systems. The student-t-distribution is similar to the continuous, normal distribution with bell-shaped and symmetrical. Mean differences, standard deviation, and t-test statics were calculated. If the t value is less than 0.05. can reject the null hypothesis (Mishra et al., 2019), (Wilkerson, 2008).

2.8 Summary

This chapter has described the overall literature review of Information Retrieval, Information retrieval Evaluation, the types of evaluation, the benefits of doing the evaluation, and the evaluation process have described in detail. Followed the TREC collection and components of the test collection have been described in detail. The overview of the different TREC versions is mentioned here. A detailed description of the test collection which has been used for this research is shown here. Other test collections like Cranfield, CLEF, NTCIR, FIRE, and INEX overviews are also shown. During the evaluation process in these test collections, it has been noticed that the number of relevant documents retrieved in the relevance judgment sets is lesser. It affected the overall quality of the judgment sets. While doing the evaluation process, due to the lesser quality of the judgment sets, the accuracy of the evaluation process is also lesser. This literature studied that much research has been conducted to improve the accuracy of the IR evaluation process. It considered based on a number of relevant documents, based on the test collections and based on incompleteness of the judgment sets and biasness in the ranking are described in detail. The limitations of these existing methodologies are also highlighted in this session. These limitations need to be overcome in order to improve the quality of the judgment set and also through that it increases user satisfaction and rely on the contributed systems. Also, various evaluation metrics used in this research to evaluate the proposed methodology have been explained in detail, and also given the statistical significance tests that have been used here to measure the performance of the systems using the proposed methodology are also listed.

CHAPTER 3: ENHANCING INFORMATION RETRIEVAL ACCURACY BY INCREASING RELEVANT DOCUMENTS

This chapter covers an overview of how to improve the accuracy of the information retrieval evaluation process by increasing the quality of the judgment sets. This can be achieved by increasing the number of relevant documents in the relevant judgment sets. In order to achieve this accuracy, an experimental methodology has been proposed here. These experimental methodologies focus on increasing the number of relevant documents in the judgment sets and through that increase the effectiveness of the information retrieval evaluation process. Section 3.1 starts with explaining the research approach on the importance of the quality of the judgment sets and the limitations on the baseline works. The research framework and proposed experimental methodology have been shown in Section 3.2. The various research techniques used in this experiment are described in Section 3.3.

The evaluation of information retrieval systems' performance is not only based on their efficiency but also their effectiveness. Effectiveness is calculated based on several relevant documents retrieved by the participating systems. The ability of the systems to retrieve as much of relevant documents and at the same time suppress the irrelevant ones (Ferro, 2017). The main aim of information retrieval evaluation is to increase the accuracy of the information retrieval evaluation by increasing the quality of the relevant judgment sets and it can be achieved by increasing the number of relevant documents in the judgment sets.

3.1 Research Approach

This section describes how the unjudged clustered or classified documents increase the number of relevant documents in the pooled list based on relevance judgment and through that increase the quality of the judgment sets. In document similarity, only clustered or classified documents are considered for the evaluation process, and pooling only considers pooled documents for the evaluation process. As mentioned in the literature review, the pooling methodology provides better quality results compared to the document similarity. The baseline results of these two methodologies are shown in Figure 3.1. For this baseline experiment, three methodologies were considered. 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. Document similarity has been done 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 in the cluster. Whenever a new user query comes, the patterns are discovered in each cluster and find out the most relevant clusters that match the user query and the clustered documents are considered for the evaluation process (Djenouri et al.,2021, Djenouri et al.,2018). 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 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 and TREC-10 Web Track collection. The details of these TREC collections are shown in Table 3.1.

Table 3.1: Datasets Overview

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

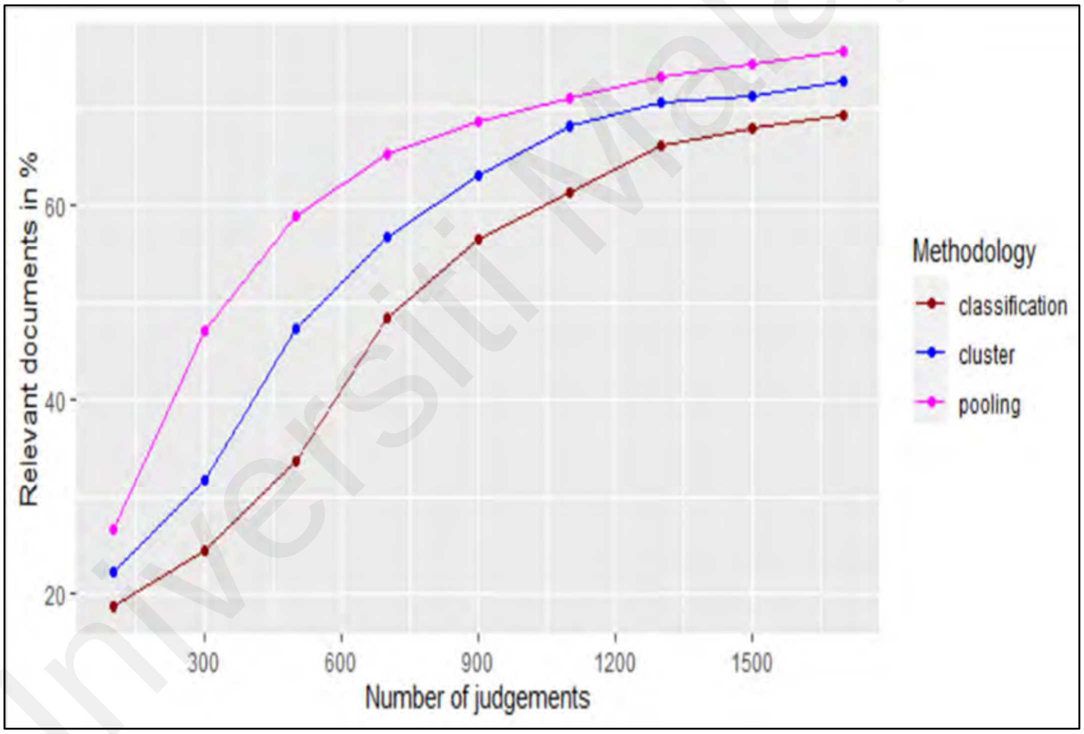


Figure 3.1: Baseline results of pooling and document similarity results based on TREC-8 collection

cluster-(ICIR) -(Djenouri et al.,2021), **classif-(CAL)-** (Rahman et al.,2020), **pooling**-(Losada et al.,2018**)**

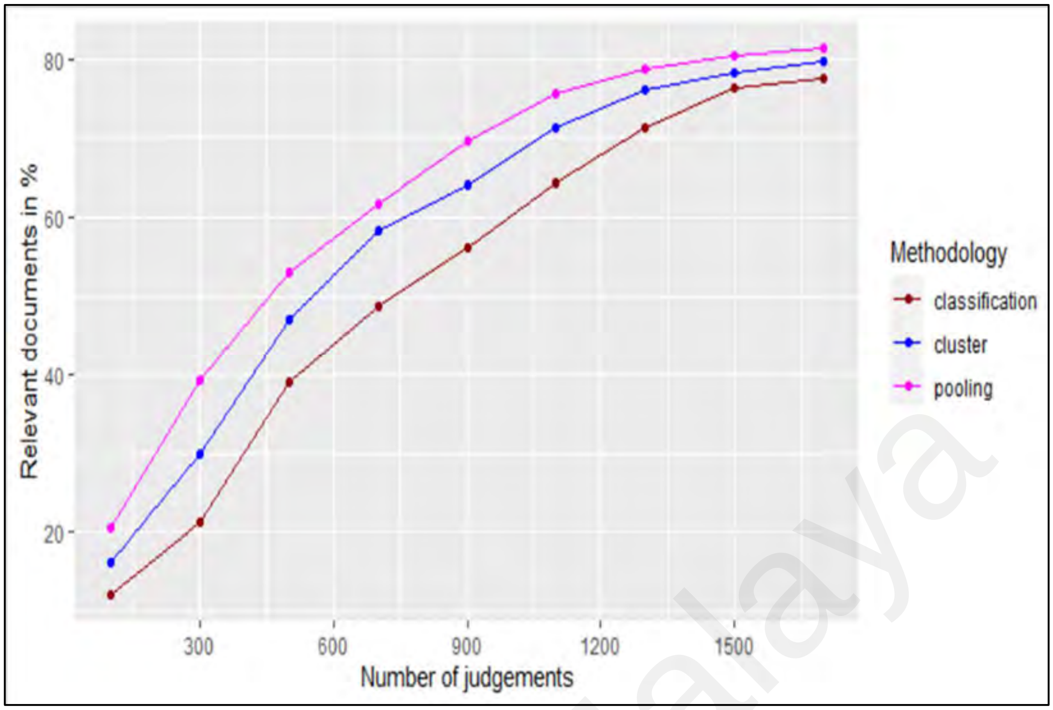


Figure 3.2: Baseline results of pooling and document similarity results based on TREC-8 collection

cluster-(ICIR) -(Djenouri et al.,2021), **classif-(CAL)-** (Rahman et al.,2020), **pooling**-(Losada et al.,2018**)**

Fig 3.1 and Fig 3.2 show the baseline results of several relevant documents (in percentage) retrieved by each methodology based on the range of several judgments. The X-axis shows the number of judgment ranges and the y-axis shows the number of relevant documents retrieved in percentage. As mentioned earlier, the results show that the pooling methodology performed better compared to the document similarity-based methodology.

The main problem concern here is that considering document similarity through a classifier or cluster globally can achieve the number of relevant documents, but the quality of these documents is comparatively lesser compared to the traditional methodologies (Djenouri et al.,2021, Djenouri et al.,2018, Rahman et al.,2020).

To overcome this problem, a methodology has been proposed by combining the pooling technique and document similarity. For the experimental design, TREC data collection has been used. TREC-8 Adhoc collection and TREC-10 Web collection. The Adhoc retrieval task investigates the performance of the participated systems that search a static set of documents using topics or queries. NIST provides participants with around 2 gigabytes of document collection and a set of 50 topic statements to participants to create a set of queries based on the topic statements and run these queries against the document collection. Participants have retrieved the best 1000 documents for each topic for evaluation purposes. The output of these runs is the official test results for the ad-hoc tasks. The relevance judgments were not known to the participants when they generated the runs. Participants have used only documents, topics, and relevance judgments from previous TREC were used to generate their runs. Fifty topics from 401-450 were created in the TREC-8 dataset (Hawking et al., 1999). The Web Track has been created with 100GB of test collection and a smaller Web task of 2 GB. The Web Track started in 1999 and was evaluated every year until 2003. Then re-continued in the year of 2009. TREC-2001 Web Track mainly focused on the topic relevance task. In this web track also 50 topics were generated from 501 to 550. TREC 2001 uses a GOV test collection which consists of 1.25 million pages.

3.2 Research Framework and Methodology

For this experiment, the TREC-8 and TREC-10 datasets were used. Data cleaning was done, and baseline experiments were done with pooling and document similarity techniques. Implementation of the proposed methodology has been done by incorporating pooling and document similarity techniques like clustering and classification. The overall framework for the implementation of this proposed methodology is shown in Figure 3.3.

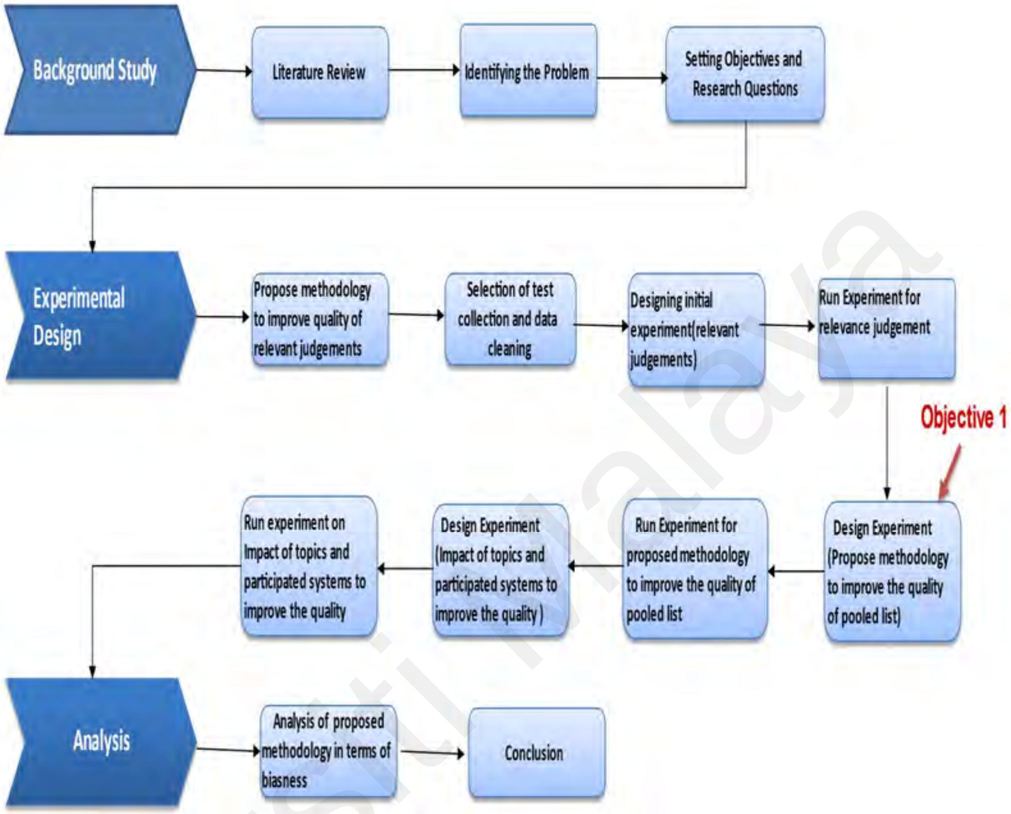

Background Study
Literature Review
Identifying the Problem
Setting Objectives and Research Questions
Experimental Design
Propose methodology to improve quality of relevant judgements
Selection of test collection and data cleaning
Designing initial experiment(relevant judgements)
Run Experiment for relevance judgement
Objective 1
Design Experiment (Propose methodology to improve the quality of pooled list)
Run Experiment for proposed methodology to improve the quality of pooled list
Design Experiment (Impact of topics and participated systems to improve the quality)
Run experiment on Impact of topics and participated systems to improve the quality
Analysis
Analysis of proposed methodology in terms of biasness
Conclusion

Figure 3.3: Research methodology flow diagram

To improve the quality of the relevant judgment set, more relevant documents have to be in the judgment list. For the experiment, as a baseline, the pooling technique has been used based on top-k documents from each run that have been considered, and these documents are called pooled documents. Documents that have not been considered in the pooled list are always considered irrelevant and those documents here are called an unjudged list. The pooling was done based on the Combsum rank aggregation technique (Losada et al., 2018). The documents from the unjudged list are clustered based on ICIR

(Djenouri et al.,2021, Djenouri et al.,2018) and classified based on CAL (Rahman et al.,2020).

Based on a test collection, each participating system retrieves a set of ranked lists of documents from the document corpus based on the user query. These ranked documents are called runs. If N participated systems are there, N runs have been created. These runs were merged with the help of a rank aggregation technique called Combsum. Judging the whole merged ranked list of documents is considered high-cost and time-consuming. So only a subset of documents is considered by using the depth-k technique. With the pool depth of k, the top relevant documents from all the runs were considered based on the topics and merged and this subset is called as pooled documents or judgment list documents (p_1,p_2…p_n). Remaining all the documents that are not considered for the judgment process are called unjudged documents ($U_{Ci \dots Cn, di \dots dn}$). These unjudged documents were plotted either by using clustering or classification techniques. The clustering technique is based on an agglomerative hierarchical clustering model and by the k-means clustering technique by considering the similarity of the documents. The next step is to find the similarity of the documents between the pooled list and the unjudged list. Next, find the highest-scored document from the pooled list. Find the similarity between the p_i and the unjudged clustered documents. Like d_i is checking the similarity with p_1, based on considering the topic title using the TF-IDF technique. The same p_1 will do the similarity checking with all the top documents in each cluster. Find out the cluster which has the highest similarity score. The top-k documents from that cluster will be assigned a new score based on equation 3.1 and move those documents into the pooled list.

$$\text{New Score assigning=}\frac{sim(di,U_{Cij})}{\sum_{l=1}^{k} sim(di,U_{Cij})} \qquad \textbf{Equation 3.1}$$

(Liang et al.,2018)

The same process continues with the second high-ranked document from the pooled list. Once the similarity is found, those top-k clustered documents also move into the judgment list by assigning new scores. The same process continues with all the pooled documents. Once the iteration is completed, the pooled documents will be sent for the evaluation process. The overall framework of the proposed methodology is shown in Figure 3.4.

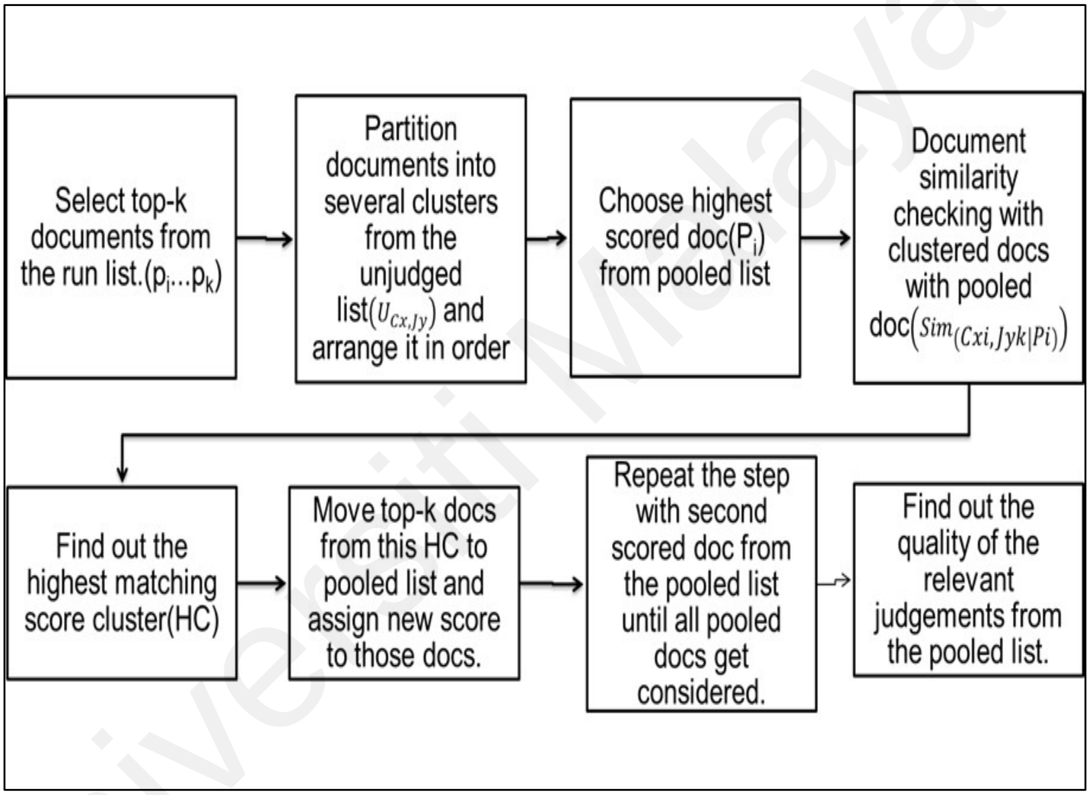

Select top-k documents from the run list.($p_i...p_k$)
Partition documents into several clusters from the unjudged list($U_{Cx,Jy}$) and arrange it in order
Choose highest scored doc(P_i) from pooled list
Document similarity checking with clustered docs with pooled doc($Sim_{(Cxi,Jyk|Pi)}$)
Find out the highest matching score cluster(HC)
Move top-k docs from this HC to pooled list and assign new score to those docs.
Repeat the step with second scored doc from the pooled list until all pooled docs get considered.
Find out the quality of the relevant judgements from the pooled list.

Figure 3.4: Experimental Methodology based on pooling and document similarity using clustering technique

The same experiment was done with the classification technique. The documents were classified based on the similarity of the documents in the unjudged list. Document similarity checking between the pooled document list and each class has been done.

Which class has produced the high similarity score, top-k documents from this class have moved into the pooled list. The second high-scored document from the pooled list has continued with the same process. Figure 3.5 shows the overall framework of the proposed methodology using the classification technique.

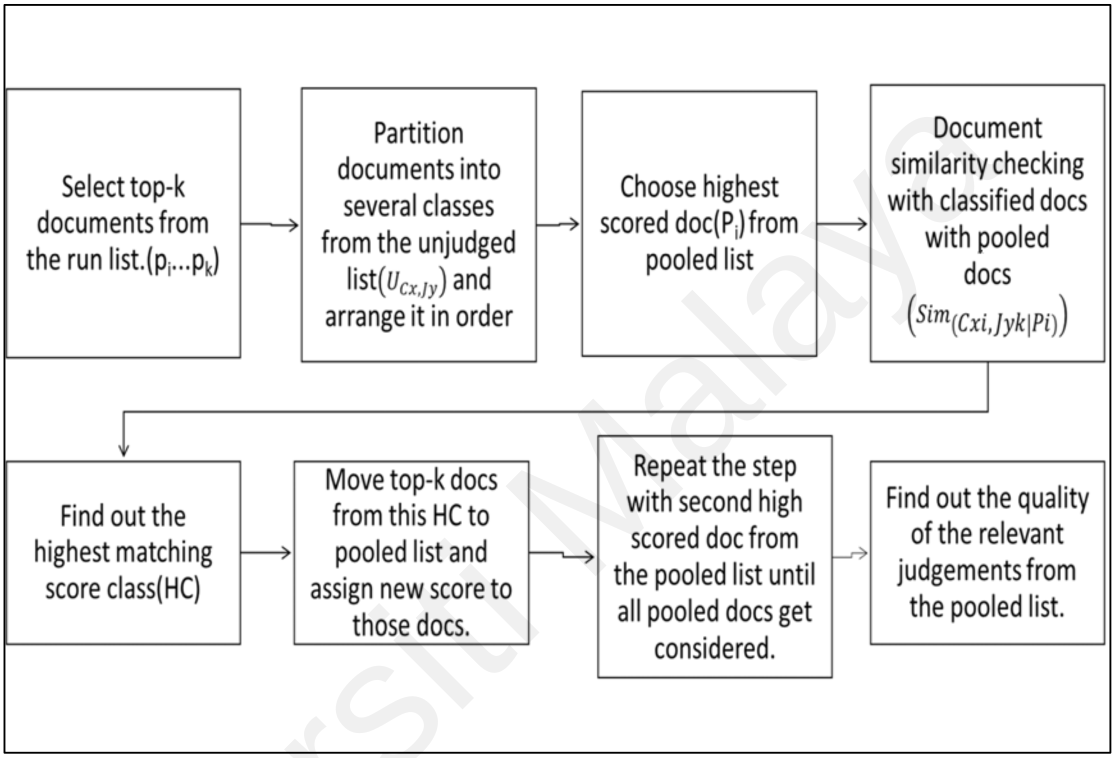

Select top-k documents from the run list.(p_i...p_k)
Partition documents into several classes from the unjudged list($U_{Cx,Jy}$) and arrange it in order
Choose highest scored doc(P_i) from pooled list
Document similarity checking with classified docs with pooled docs ($Sim_{(Cxi,Jyk|Pi)}$)
Find out the highest matching score class(HC)
Move top-k docs from this HC to pooled list and assign new score to those docs.
Repeat the step with second high scored doc from the pooled list until all pooled docs get considered.
Find out the quality of the relevant judgements from the pooled list.

Figure 3.5: Experimental Methodology framework based on pooling and document similarity using classification technique

Step-by-step process of proposed methodology:

1. Runs have been created from the participating systems
2. These runs were merged using a rank aggregation technique
3. Choose top-k documents from these merged lists, pooled document list
4. Using clustering/classification technique partition unjudged docs based on the similarity

5. Choose the highest-scored doc from the pooled list
6. Do document similarity checking with this high-scored pooled doc with each cluster/class document.
7. Find out the highest-scored cluster/class that matches with that doc.
8. Move top-k documents from this HC to the pooled list by assigning a new score.
9. Repeat the step with the second high-scored doc from the pooled list until all pooled docs are considered.
10. This updated pooled list has been given for the evaluation process.

3.3 Research Techniques

The various techniques used in this proposed methodology have been explained here.

K-means clustering

Clustering analysis is a technique used to group similar objects into clusters. K-means clustering is the most used cluster analysis which helps to partition the objects into k number of clusters. One of the main properties of k-means clustering is to make sure that all objects in the cluster are similar to each other. In the same way, the objects from the different clusters should be as different as possible. We assign objects to a cluster based on the distance of that object from the centroid.

The main aim of the k-means clustering is to divide M objects in N dimensions into K clusters. The algorithm requires an input of M points in N dimensions and K initial clusters in N dimensions. Euclidean distance has been used to find the distance between the objects to the cluster. The general procedure is to search for k-partitions with locally optimal within-cluster sum of squares by moving points from one cluster to another. The generalized k-means cluster is shown here.

Steps for k-means clustering:

1. Choose the number of clusters, k
2. Select k random points from the objects as centroids
3. Assign all the objects to the closest centroid.
4. Recompute the centroids based on the newly formed cluster
5. Repeat steps 3 and 4.

This iteration goes until the centroids of the clusters are not changed.

(Hartigan & Wong, 1979) (Likas et al., 2003)

Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering algorithms can be characterized as greedy, in the algorithmic sense. An irreversible algorithm has been used to create the data structure (Murtagh & Contreras,2012). Agglomerative Hierarchical Clustering helps to group documents into clusters based on their similarity. This clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters (Sasireka & Baby,2013). This algorithm starts considering each document as a singleton cluster. Next, pairs of clusters are successively merged based on their similarities until all the clusters are merged into one big cluster containing all the documents. The result is a tree-based representation of the documents based on its similarity. It means all the similar documents will be merged under one branch.

Efficient Fusion technique

The cluster hypothesis states that the documents within the same cluster or manifold are likely to have the same degree of relevance to the same information needed in a given query (Liang & Rijke,2015). These concepts have given a solution to many ranking problems, but only to a limited extent. It sometimes led to a negative impact on the efficiency. A novel manifold-based data fusion approach helps to provide support by using inter-document similarities within the document manifold of documents being fused. This fusion technique has been used for the fusion score regularization. This fusion method integrates with the standard unsupervised data fusion method such as CombSUM. This method aims to calculate a fusion score for each document that appears in the input result lists to be fused. This one is mainly used for the inter-document similarities of all the documents in the collection. The new score assigned to the documents is based on the similarity of the document and the document from a particular clustered documents list (Equation 3.1).

$$\text{New Score assigning}=\frac{sim(di,U_{Cij})}{\sum_{l=1}^{k} sim(dl,U_{Cij})}$$

Hence the more similar document di and U_{Cij} (i^th document from the j^th cluster), the higher the score of the documents. (Liang et al.,2018)

Combsum (Rank aggregation technique)

Rank aggregation approaches also called data fusion approaches, combine the results of participating systems to produce a new better ranking. Rank aggregation methods improve the performance of the evaluation process concerning those of the input methods (Bartell et al., 1994). Various Rank aggregation techniques have already been introduced like Combsum, CombMNZ, Borda, etc. It is used to merge all the multiple ranked documents. Combsum is the most popular rank aggregation algorithm which has been used in our experiments. Combsum sums the document's retrieval scores from all search systems that retrieved the documents. It helps the documents that are relevant get high scores and appear in the judgment list. In the pooled list, we computed the score of the documents based on the Combsum and ranked the documents in the decreasing order of that score (Shaw & Fox, 1994). Each run will be merged together and ranked according to its relevancy. These documents are given for the evaluation process.

CHAPTER 4: INCREASING THE QUALITY OF RELEVANT JUDGEMENTS BY CONSIDERING TOPICS AND PARTICIPATING SYSTEMS FROM TEST COLLECTIONS

Test collections have a high impact on the quality of the relevance judgment sets. However, the evaluation of information retrieval evaluation systems is a challenging process as the information is getting added to the Web. The main components that have a role in evaluating systems are the topics and the efficiency of the participating systems. In many research, researchers in the information retrieval field have tried to achieve better evaluation accuracy with fewer topics and lesser relevance judgments (Voorhees & Buckley,2002) (Culpepper et al., 2014). This session aims to find out the influence of topics and participating systems in the quality of the relevance judgment in a better way. Section 4.1 explains how the proposed methodology can be beneficial by considering the topics from the test collections in a cost-effective way and Section 4.2. explains how the proposed methodology can further enhance the result on the quality of the judgment sets by considering participated systems. The performance of the proposed methodologies in terms of incompleteness and biasness in the ranking of documents is shown in detail in Section 4.3

4.1 Effect of Topic size to improve the quality of the pooled list

Among the test collections, topics have an important role in the system performance. Each subset of documents from the same runs might produce different sets of relevance judgments. Some topics generate a quality judgment list compared to other topics. Finding the best topics that can produce more relevant documents helps to reduce the computational cost (Breto et al., 2013).

It is essential to consider the topics while evaluating the participating systems. As the topic size increased, the cost of the evaluation also increased. So need to reduce the number of topics to be evaluated without affecting the quality of the judgment sets. Also, choosing the best topics matters. Some topics can retrieve more relevant documents compared to others. Topic difficulty or topic hardness can be defined as how well a participating system can perform in retrieving more relevant documents when evaluating a topic. If the topic is high scored, then most of the participating systems achieve high scores when evaluating these systems against that topic and that topic can be considered as an easy topic. At the same time, the low-scored topics make the systems scores less which considered those topics as hard topics. This session covers how effectively can maintain the quality of the relevance judgment sets with reduced topic size and at the same time by considering topic hardness.

4.1.1 Research Approach

Each topic generates a different number of relevant documents based on its performance. The variation in the topic's performance can be due to various factors, such as variation in the number of relevant documents per topic in the test collections. Some topics might have more relevant documents compared to other topics.

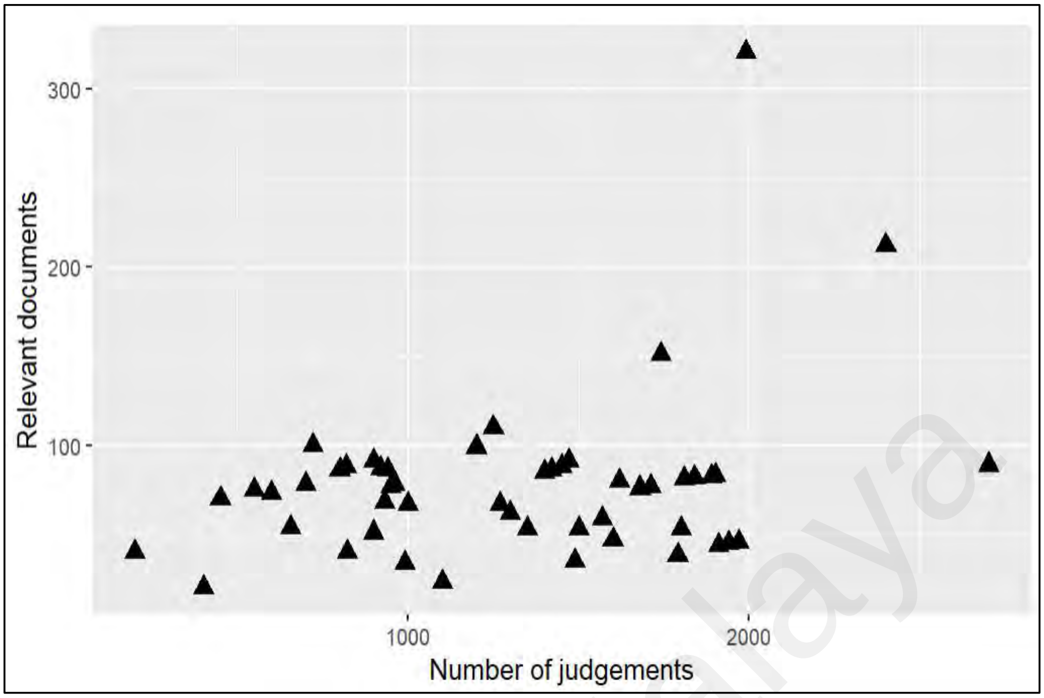


Figure 4.1: Number of documents judged per topic in TREC-8 Adhoc Track

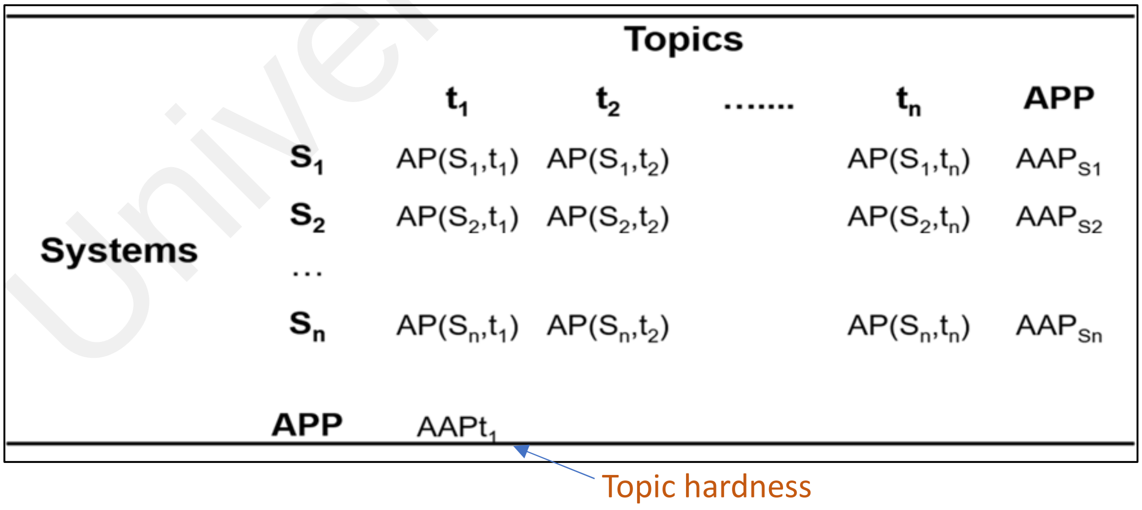
As an example, Figure 4.1 shows the topic performance over the whole test collection run. Experiments have been done based on the TREC-8 Adhoc track. It consists of 50 topics from 401 until 450 and 129 participated systems. The horizontal lines indicate the number of documents judged per topic and the y axis indicates the number of relevant documents found based on per topic. The results show that the number of documents judged by each topic varies much to the judgment pool. At the same time, it has been noticed that even with a lesser judgment pool itself can achieve most of the relevant documents.

The topics that have contributed less to the judgment pool might be due to the topic's hardness or the topic's difficulty. Topic difficulty can be found based on the topic difficulty score. Calculating the topic difficulty score of all topics helps to determine the difficulty of each topic and also helps to select the topics based on their score. The hardness of the topics can be calculated either based on the Average of Average Precision (AAP) or Topic Difficulty Score (TD)(Carterette et al., 2009) (Berto et al.,2013) (Ting Pang et al.,2019) (Gienapp et al.,2021). Average of Average Precision can be calculated based on the summation of the average precision of all the systems on a particular topic and also the total number of systems. Figure 4.2 shows how to evaluate the topic's performance over the systems. The average of average precision (AAP) is calculated to find the topic hardness. For *n* systems, APt₁ shows the Average precision of topic t₁ over all the systems from S_1 until S_n. AAPt₁ is the summation of all Average Precision of topic t_1 over all the systems.

Based on the Average of Average Precision of topics, the topic difficulty was calculated as

$$\textbf{Topic Difficulty } (\mathbf{TD_t}) = \frac{\sum_{i=1}^{n} AP(S_{n,t})}{S_n} \qquad \textbf{Equation 4.1}$$

(Carterette et al., 2009) (Berto et al.,2013) (Ting Pang et al.,2019) (Gienapp et al.,2021)



		Topics				
		t_1	t_2	....	t_n	APP
Systems	S_1	$AP(S_1,t_1)$	$AP(S_1,t_2)$		$AP(S_1,t_n)$	AAP_{S1}
	S_2	$AP(S_2,t_1)$	$AP(S_2,t_2)$		$AP(S_2,t_n)$	AAP_{S2}
	...					
	S_n	$AP(S_n,t_1)$	$AP(S_n,t_2)$		$AP(S_n,t_n)$	AAP_{Sn}
	APP	$AAPt_1$				

Topic hardness

Figure 4.2: Average of Average Precision to Calculate Topic Hardness

Figure 4.3 shows how the topic's performance varies based on the topic hardness using equation 4.1. The experiment was done with TREC-8 Adhoc collection which consists of 129 systems with 50 topics. AAP scores of 129 systems over 50 topics range shown in the x-axis and the AP score of two systems' performance is shown in the y-axis.

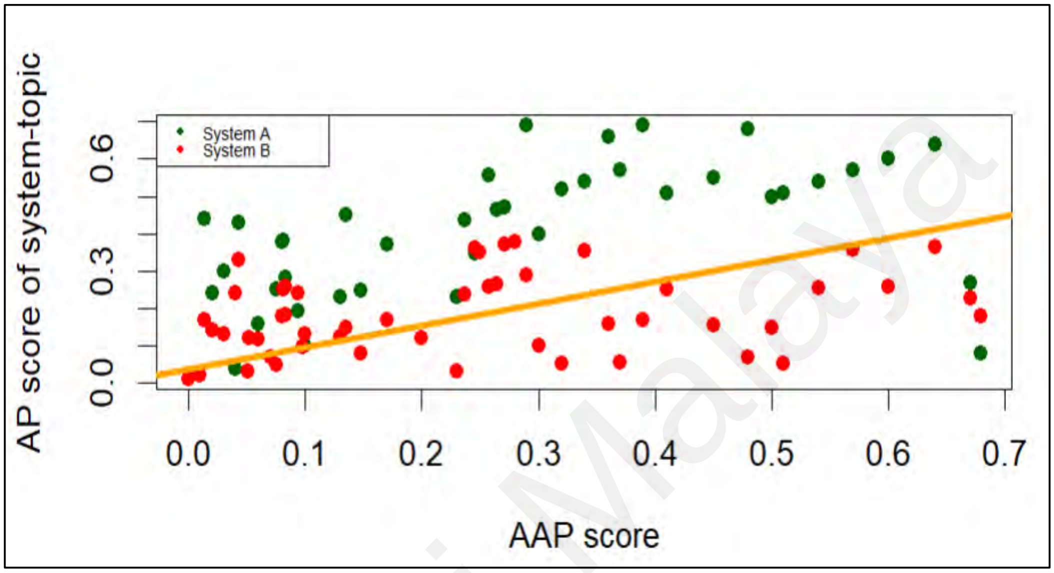


Figure 4.3: Example of two systems' performance based on topic hardness

The results show that System A performs in achieving a high topic score compared to System B. System B is affected by the topic hardness due to the low average precision score of the topics. Another method of calculating topics is based on aggregation techniques. Many studies have shown the results that hard topics exist, and they are hard for all the systems. All the hard topics were classified with a low mean average precision score and with at least one high outlier score (Voorhees,2003). To define the topic *t* difficulty D_t as:

$$D_t = \frac{max_t - mean_t}{SD_t} \qquad \textbf{Equation 4.2}$$

where,

D_t= Topic Difficulty score

max_t=high score of average precision

$mean_t$= median average precision score

SD_t=Standard Deviation

where,

$$SD_t = \frac{\sqrt{\sum(maxt-\overline{mean})}^2}{n-1}$$

where, n=number of topics

(Ravana et al., 2009)

High average precision and high AAP or TD score of a topic means the topics are performing better and are considered as Easy topics. Low average precision and low AAP means the systems are performing badly on that topic, and it is considered as Hard topic.

4.1.2 Research Framework and Methodology

The experiment was done with the TREC-8 test collection which consists of 129 systems and 50 topics and the TREC-10 test collection with 97 systems with 50 topics. This experiment was conducted as a continuation of the proposed methodology. The proposed methodology has been done with pooling and document similarity techniques using clustering and classification concepts. The results show that compared to the baseline works, this proposed methodology performed well in retrieving more relevant documents into the judgment sets and through that increased the quality of the judgment sets. The impact of topics in the proposed methodology has been evaluated here and how effectively these topics can perform on the proposed methodology has been evaluated here. The overall aim is to achieve or maintain the quality of the relevant judgment sets with reduced topic size and at the same time choose the best topics with minimal topic hardness. The overall structure of the evaluation of topics is shown in Figure 4.4.

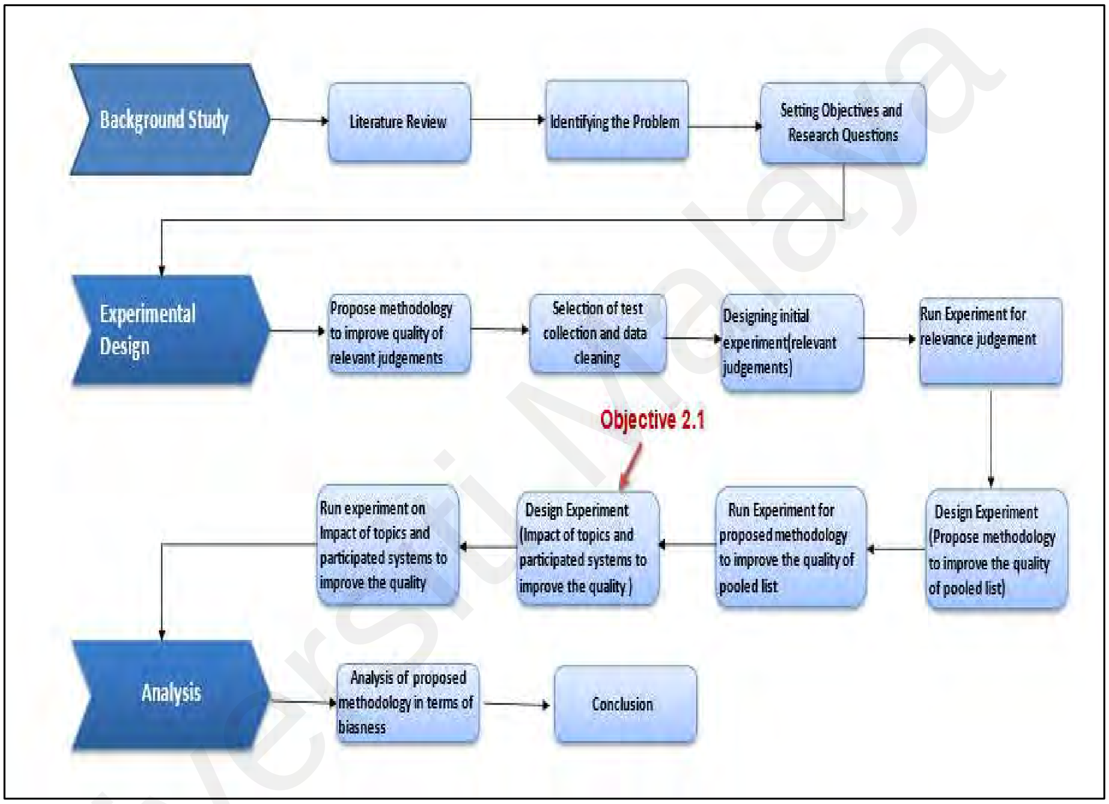

Background Study
Literature Review
Identifying the Problem
Setting Objectives and Research Questions
Experimental Design
Propose methodology to improve quality of relevant judgements
Selection of test collection and data cleaning
Designing initial experiment(relevant judgements)
Run Experiment for relevance judgement
Objective 2.1
Design Experiment (Propose methodology to improve the quality of pooled list)
Run Experiment for proposed methodology to improve the quality of pooled list
Design Experiment (Impact of topics and participated systems to improve the quality)
Run experiment on Impact of topics and participated systems to improve the quality
Analysis
Analysis of proposed methodology in terms of biasness
Conclusion

Figure 4.4: Research framework on the effect of topics in the evaluation process with the proposed methodology

The experiment based on the topic has been done with the proposed methodology and categorizes documents into easy and hard topics. Evaluation metrics of average precision and mean average precision have been calculated based on these topic categories. Kendal tau correlation was used to find the correlation of the overall mean average precision of all the topics with the average precision of the two categories of the topics (Ting Peng et al., 2019). The experiment was done with both clustered and classified documents. The results show that both have categorized the topics almost a similar way and correlation also in the same category.

The topics were categorized based on the difficulty of the Average of the average precision score. The score estimation is done based on a single query over all the participated system runs. The topics were categorized based on these AAP scores and classified into three intervals. These intervals have been chosen to distribute the topics evenly (Carterette et al., 2009). In our experiment, the intervals have split into two categories. Easy topics and Hard topics. The intervals have been distributed based on the (Carterette et al., 2009) and

$$\text{Hard topics: AAP} \in [0,0.17]$$

and

$$\text{Easy topics: AAP} \in [0.17, \text{max}]$$

70% of the queries were hard which includes more queries because it includes most of the queries where no relevant documents were found. 30% of the topics were easy consisted of most of the relevant documents were found as per the TREC-8 data collection. The experimental methodology of the topic evaluation is shown in Figure 4.5. The experiment was conducted as a continuation of the proposed methodology. Calculate the mean average precision of all the topics over all the systems. Partition documents into easy and hard topics. Calculate the average precision of k-topics for some random choices (Tang et al.2019). The same experiment was done with both classified and clustered documents. The detailed experiment of the evaluation of topics is shown below.

- Calculate the MAP of all topic's overall systems (actual MAP)

- Based on the AAP scores, topics were split into hard and easy topics
- Randomly choose k-topics from each category
- Calculate the Average Precision of those topics with all systems
- Repeat this step for several times (10 or 100)
- Calculate the mean average precision of these repeated processes of the topics separately
- Calculate Kendal's tau correlation with MAP and actual MAP

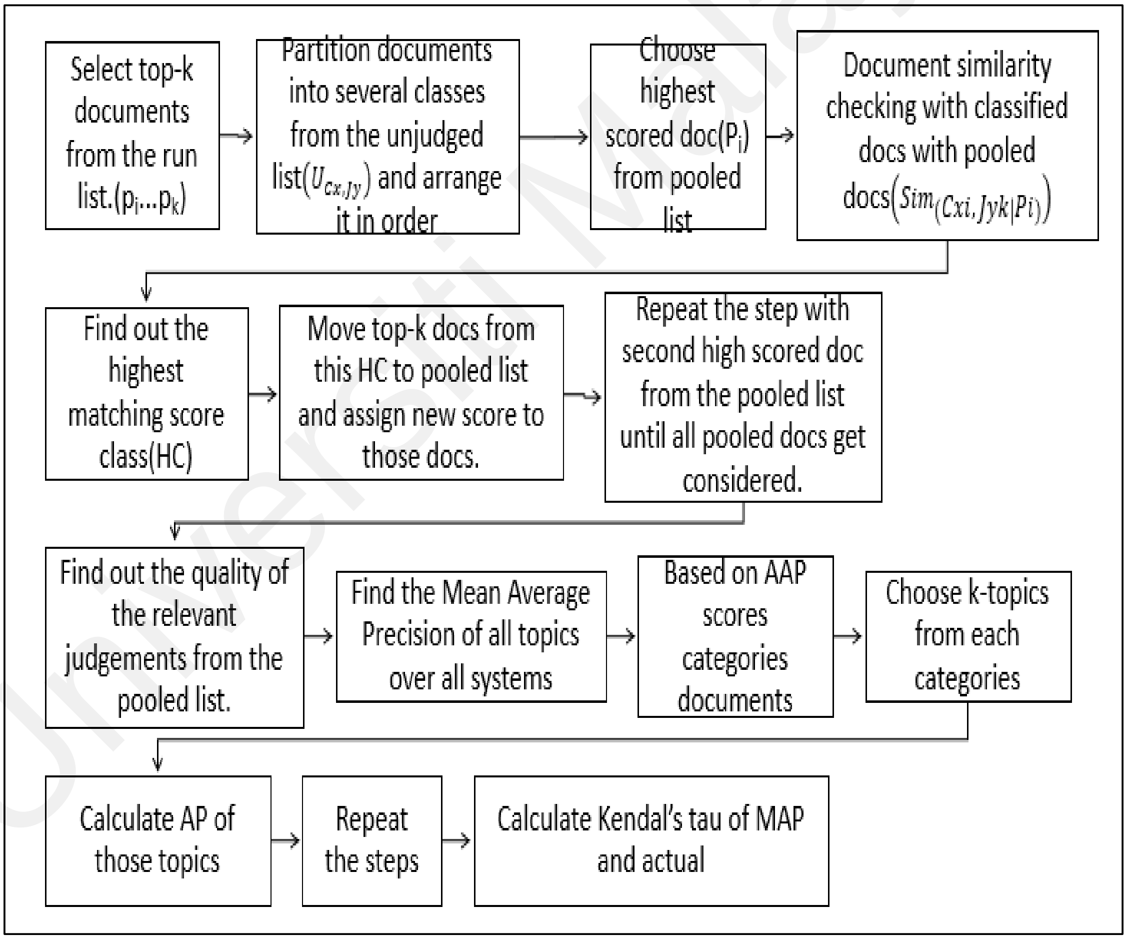

Select top-k documents from the run list.(p_i...p_k)
Partition documents into several classes from the unjudged list($U_{cx,Jy}$) and arrange it in order
Choose highest scored doc(P_i) from pooled list
Document similarity checking with classified docs with pooled docs($Sim_{(Cxi,Jyk|Pi)}$)
Find out the highest matching score class(HC)
Move top-k docs from this HC to pooled list and assign new score to those docs.
Repeat the step with second high scored doc from the pooled list until all pooled docs get considered.
Find out the quality of the relevant judgements from the pooled list.
Find the Mean Average Precision of all topics over all systems
Based on AAP scores categories documents
Choose k-topics from each categories
Calculate AP of those topics
Repeat the steps
Calculate Kendal's tau of MAP and actual

Figure 4.5: Experimental methodology of the evaluation process based on the topic hardness.

4.2 Effect of Participated systems to improve the quality of the pooled list

The web collection is getting added in real-time and it causes the lack of inconsistency in performing the information retrieval evaluation process. As the data collection is getting increase many relevant and irrelevant documents are getting added. At the same time, many proxies also getting added to the data collection in the matter of topic titles, missing documents, topic descriptions, and unrelated titles (Rasmussen,2003). Due to that biasness happens to the systems during the pooling process and only particular systems documents were considered for the judgment process and it affects the evaluation of information retrieval systems.

System performance has a greater impact on the quality of the relevant judgment sets. The test collection called TREC, is one of the most commonly used data collection to evaluate the performance of the retrieval systems. Generally, the general assumption of the Cranfield paradigm is that the documents in the relevant judgments are all relevant. This means the documents are complete. But for larger collections, obtaining all the documents in the judgment become relevant is impossible due to the large effort of human accessors.

4.2.1 Research Approach

In the area of research in information retrieval and machine learning, the evaluation of retrieval systems is an important part that can contribute many advantages to the information retrieval process. Many evaluation measures have been used to compare the performance of the retrieval systems (Raghavan et al., 1989). To evaluate the system performance, Cranfield Paradigm, also generally known as test collection has been used. In Cranfield methodology, based on the document collection and topics, a set of relevant judgments were found. Based on these relevant judgment sets, the performance of the participated systems was evaluated (Voorhees,2002).

As per the Cranfield Paradigm methodology, the general assumption is that the documents in the relevance judgment sets are all relevant. It means that all relevant documents based on each topic have been identified in the relevance judgment sets. But judging the whole document is practically difficult due to the human accessors cost. The judging of the whole relevance judgment set is time-consuming also. Many greedy techniques were proposed by the researchers to reduce the judgment sets and at the same time maintain the quality of the judgment sets (Aslam et al.,2003).

The most popular technique to evaluate the retrieval or the participated systems with lesser relevance judgment sets is through the depth pooling method. TREC test collection is one of the best choices for these kinds of evaluations. In the depth pooling method with k means, only the top k documents from the relevance judgment sets were considered for the evaluation process and the rest of the documents were considered irrelevant and we called those documents an unjudged document list. The number of relevant documents to be judged varies based on the pool depth. Based on Table 3.1, the total number of relevant documents to be judged based on different pool depths is shown in Table 4.1.

Table 4.1: Number of documents to be judged based on various pool depth

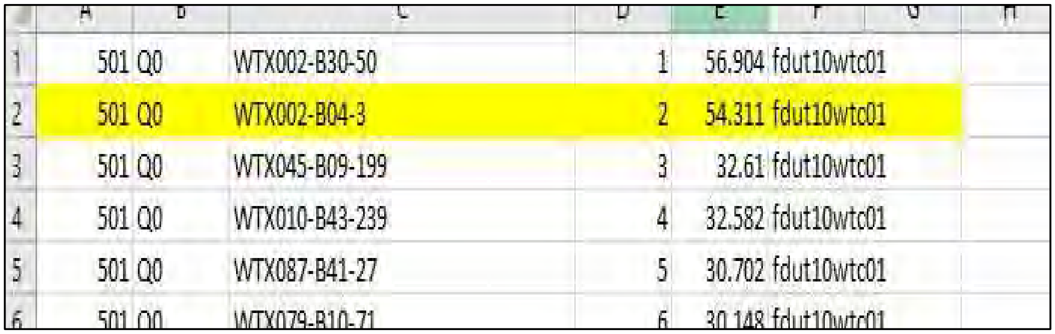
Pool Depth-k	TREC-8	TREC-10
10	200	184
20	379	339
50	713	698
100	1712	1414

A pool depth of 100, is a standard choice, and in many research, it has been proven that most of the relevant documents are found at these pool depths. And this depth is enough for the evaluation of the retrieval systems.

But in the case of dynamic document collection, this concept sometimes won't be correct. In dynamic document collection, the document collection is changing, and new documents are kept getting added or deleted from the test collection. In such cases, the results based on the evaluation of the retrieval systems might be incorrect. It happens due to the updated test collection with added or deleted documents was not considered for the evaluation process. So the pool depth might not be able to retrieve many documents in the dynamic document collection.

On the other hand, sometimes the relevance judgment sets can be incomplete. For the smaller datasets, the concept of fixed pool depth might work. But for larger test collections, with a fixed pool depth, sometimes the number of relevant documents retrieved might not be correct. As the pool depth increases, the retrieval systems are retrieving more relevant documents into the judgment sets. Due to this, we cannot rely completely on a fixed pool depth with an assumption that all the relevant documents have been retrieved, and relevance judgment sets are complete.

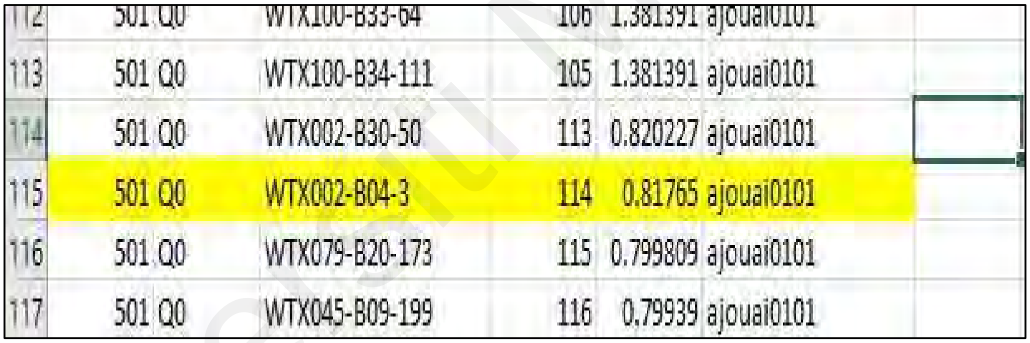
One of the main reasons for this drawback is the performance of the retrieval systems. Each retrieval system assigns different ranks for the documents that are actually relevant. Hence, with the standard pool depth, most of the relevant documents have not been considered in the judgment sets. As an example, Figure 4.6 shows the performance of two systems by assigning different ranks for a document for the same topic.



	A	B	C	D	E	F	G	H
1	501	Q0	WTX002-B30-50	1	56.904	fdut10wtc01		
2	501	Q0	WTX002-B04-3	2	54.311	fdut10wtc01		
3	501	Q0	WTX045-B09-199	3	32.61	fdut10wtc01		
4	501	Q0	WTX010-B43-239	4	32.582	fdut10wtc01		
5	501	Q0	WTX087-B41-27	5	30.702	fdut10wtc01		
6	501	Q0	WTX079-B10-71	6	30.148	fdut10wtc01		

Figure 4.6: Rank assigned for a document for a topic by system fdut10wtc01

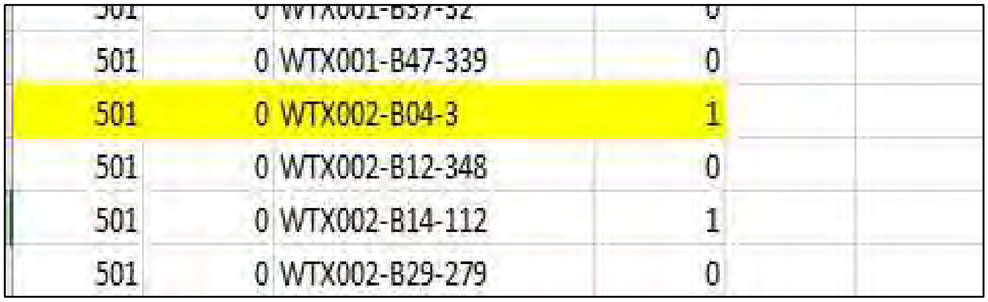
From the TREC-10 test collection Figure 4.6 shows that, based on a system run named fdut10wtc01, for topic 501, the document named WT002-B-04-3, the system has assigned rank 2 for the document based on the topic relevancy.



112	501	Q0	WTX100-B33-64	106	1.381391	ajouai0101	
113	501	Q0	WTX100-B34-111	105	1.381391	ajouai0101	
114	501	Q0	WTX002-B30-50	113	0.820227	ajouai0101	
115	501	Q0	WTX002-B04-3	114	0.81765	ajouai0101	
116	501	Q0	WTX079-B20-173	115	0.799809	ajouai0101	
117	501	Q0	WTX045-B09-199	116	0.79939	ajouai0101	

Figure 4.7: Rank assigned for a document for a topic by system ajouai0101

At the same time, from the TREC-10 test collection Figure 4.7 shows that, based on another system run named fdut10wtc01, for the topic 501, the document named WT002-B-04-3, the system has assigned rank 114 for the document based on the topic relevancy.



501	0	WTX001-B37-32	0		
501	0	WTX001-B47-339	0		
501	0	WTX002-B04-3	1		
501	0	WTX002-B12-348	0		
501	0	WTX002-B14-112	1		
501	0	WTX002-B29-279	0		

Figure 4.8: Judged relevance judgment set of TREC-10 qrels.

As per the TREC-10 qrels Figure 4.8 shows that the already judged relevance judgment set, the document WTX002-B-04-3 for a topic 510 is shown as actually relevant. When the pooling is done with the top 10 or top 50 or classic top 100 document list, the system named fdut10wtc01 was able to move the document into the pooled list, but the system named fdut10wtc01could not send that document into the pooled list due to the rank assigned is lesser and the evaluation process consider this document as irrelevant. Due to that, the performance of the system varies based on the ranking.

Various evaluation measures have been used by the researchers to find the performance of the retrieval systems. The standard evaluation measures used for this purpose are average precision and R-precision (Buckley and Voorhees, 2004). b-pref is also used to evaluate the incompleteness of the relevance judgment sets. B-pref is highly correlated to average precision. But it has been proven that by using b-pref the judgement sets get more and more incomplete. So, average precision is considered a gold standard and the best evaluation measure to evaluate the system performance based on the judgment sets (Yilmaz and Aslam, 2006). Figure 4.9 and Figure 4.10 shows the performance of the random relevant systems based on the complete judgment sets. Figure 4.9 and Figure 4.10 show two random sample sets of systems performances. The experiment was done with the TREC-8 dataset and consisted of 129 systems and 50 topics. The mean average precision of all topics with all systems was calculated and considered as actual MAP. 10% of random samples of systems were considered from these 129 systems and calculated the mean average precision for the 13 systems. The correlation coefficient of the actual MAP and the MAP of these random sample systems are plotted here in Figure 4.9. Another random sample of the system performance have shown in Figure 4.10. It has shown that some systems perform very well even with the same documents and same topics. Some systems can contribute well to the judgment set and others cannot. Choosing documents from performance systems will affect the quality of the judgment sets and later these documents considered for the evaluation process will affect the accuracy of the evaluation process. The effectiveness or performance of the systems can be measured with various evaluation measures such as MAP, AP, NDCG Rbp, etc. The variation in the effectiveness scores could contribute to the reliability of the participated systems. Based on these scores the retrieval or participated systems can categorized based on their performance.

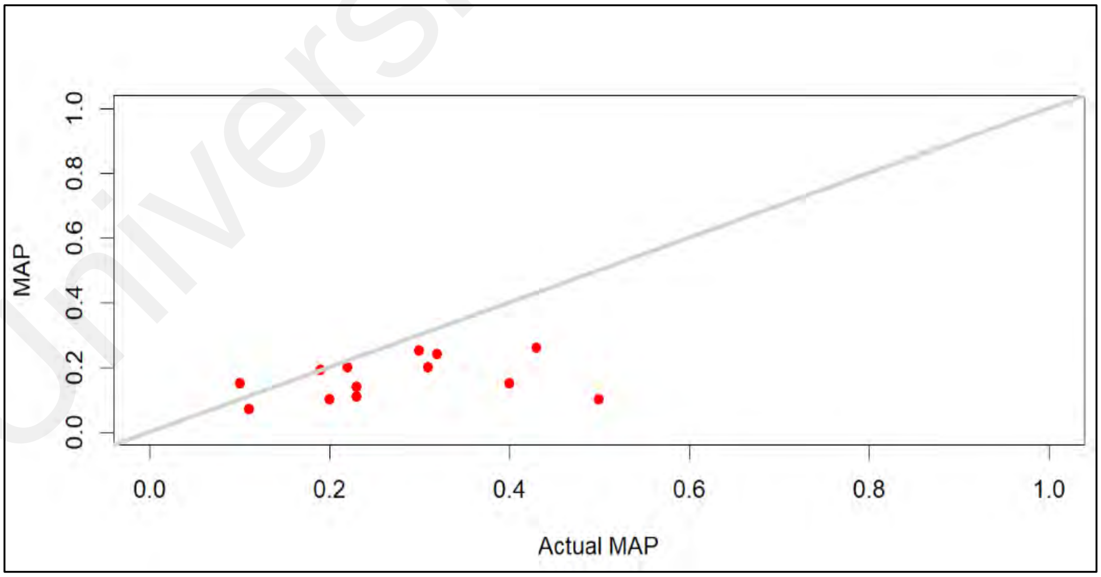


Figure 4.9: MAP vs Actual MAP for 10% of random retrieval systems from TREC-8

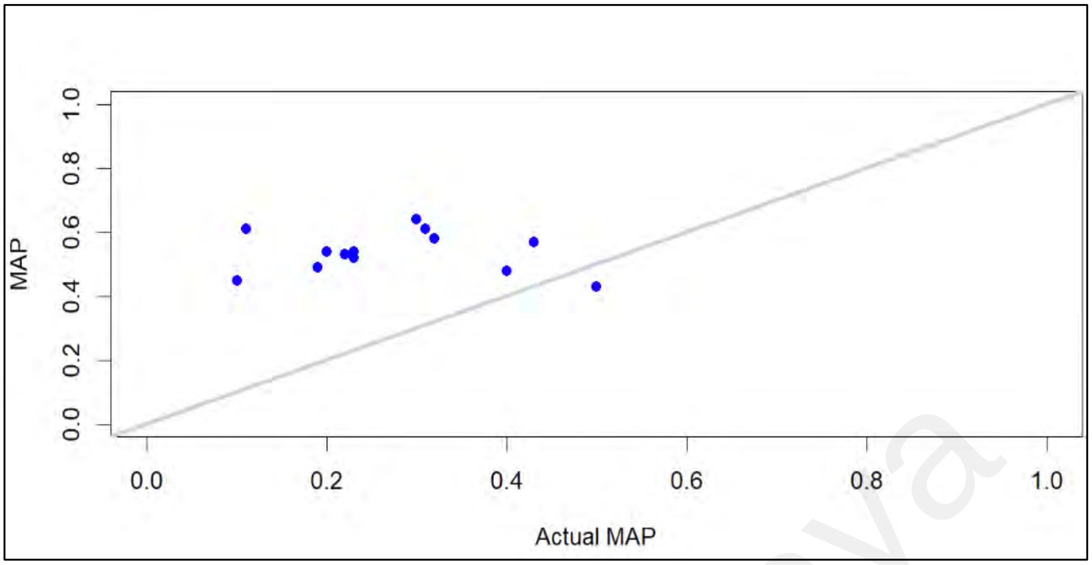


Figure 4.10: MAP vs Actual MAP for another 10% of random retrieval systems from TREC-8

Good contributing systems: These systems can assign better ranks for the relevant documents and at the same time suppress the irrelevant ones.

Less contributing systems: The systems that assign lower ranks to the documents or do not consider the documents that are actually relevant.

The main concern is less contributing systems assign lower ranks to the documents that are actually relevant. The documents that are chosen from these types of systems or the irrelevant ones into the pooled list or relevance judgment sets and later considered documents for the evaluation process will affect the quality of the relevance judgment sets and through that affect the accuracy of the evaluation process (Iwayama,2000, Djenouri et al.,2018, Rahman et al.,2020).

4.2.2 Research Framework and Methodology

Based on the performance of the retrieval systems, the quality of the judgment sets varies. If the contributed or retrieval systems can retrieve relevant documents into the judgment sets and at the same time assign a better rank for these documents helps to increase the quality of the judgment sets. At the same time, good systems can suppress the irrelevant ones too. The main aim of this research is to increase the quality of the judgment sets most efficiently. For that, a better methodology is proposed based on the methodology proposed in Section 3.1. In this methodology, the system's performances were considered. The system performances were calculated based on the evaluation measures. The evaluation measures assign a score for each participating system based on its performance in retrieving relevant documents into the judgment set. Categorization of the systems was done based on these evaluation scores. The documents from the high-score assigned systems received a high preference for the judgment list. The documents from the low-scored systems were considered as an unjudged list. The experiment has been done based on the proposed methodology mentioned earlier and has done document similarity checking with both clustering and classification techniques. The experiment was done with TREC-8 and TREC-10 datasets. The framework of the proposed methodology is shown in Figure 4.11.

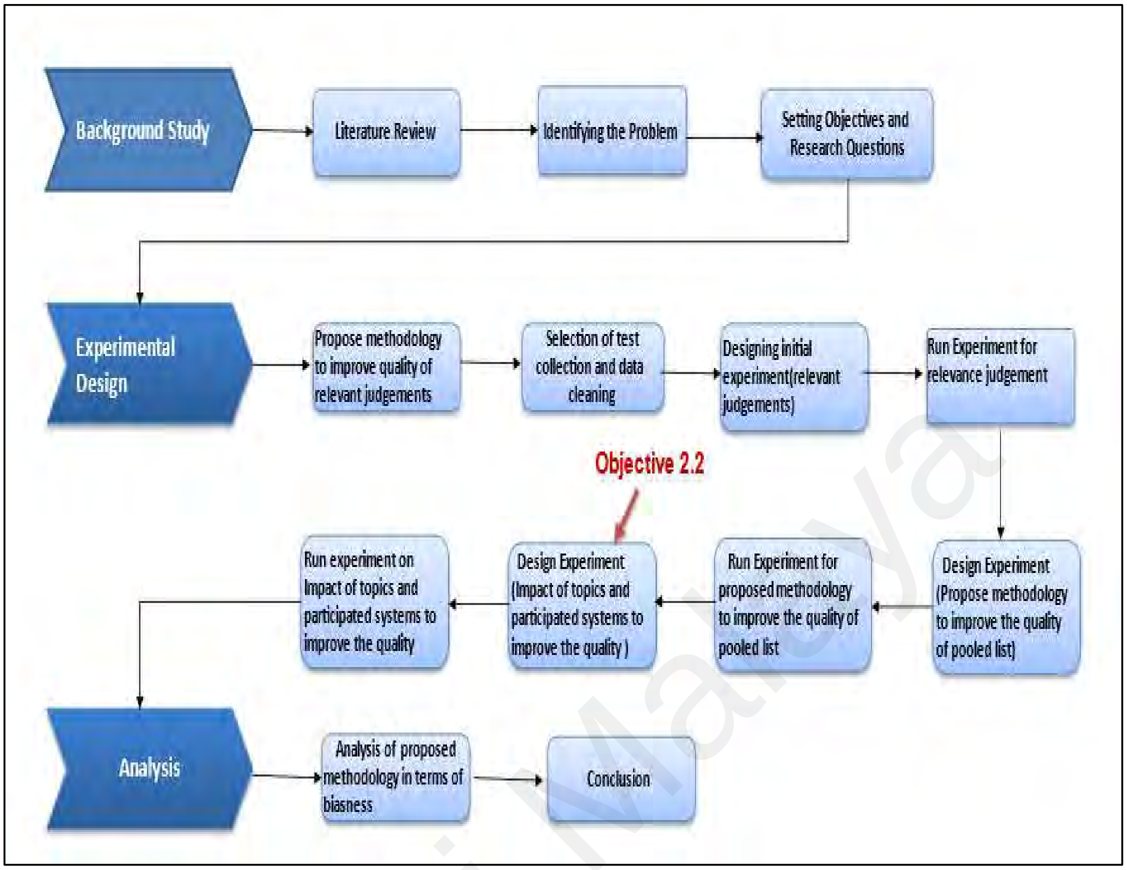

Background Study
Literature Review
Identifying the Problem
Setting Objectives and Research Questions
Experimental Design
Propose methodology to improve quality of relevant judgements
Selection of test collection and data cleaning
Designing initial experiment(relevant judgements)
Run Experiment for relevance judgement
Objective 2.2
Run experiment on Impact of topics and participated systems to improve the quality
Design Experiment (Impact of topics and participated systems to improve the quality)
Run Experiment for proposed methodology to improve the quality of pooled list
Design Experiment (Propose methodology to improve the quality of pooled list)
Analysis
Analysis of proposed methodology in terms of biasness
Conclusion

Figure 4.11: Research methodology flow diagram

The methodology works as follows.

Based on a test collection, each participating system retrieves a set of ranked lists of documents from the document corpus based on the user query. These ranked documents are called runs. If N participated systems are there, N runs have been created. These runs were merged together with the help of a rank aggregation technique called Combsum. Judging the whole merged ranked list of documents is considered high-cost and time-consuming. So only a subset of documents is considered by using the depth-k technique. With the pool depth of k, the top relevant documents from all the runs were considered based on the topics and merged and this subset is called pooled documents or judgment list documents (p_1,p_2…p_n). Remaining all the documents that are not considered for the judgment process are called unjudged documents ($U_{Ci \ldots Cn, di \ldots dn}$). Evaluation measures have been applied to these pooled documents to find out the better-performed systems. For this experiment, average precision-based evaluation measures were used. As many irrelevant documents are there in the pooled list, the average precision values vary, and it affects the ranking of the retrieval systems. Induced Average Precision is used in order to make the judgment correctly. For a topic with R-relevant documents, induced AP works almost similar to the average precision with only a slight change. In the induced AP, the documents that are irrelevant to the judgment list or the pooled list were removed and not considered for evaluation. Once these irrelevant documents were removed, induced AP was calculated the same way as average precision. Induced AP provides a better precision-recall curve with the pooled documents as it considers only relevant documents (Yilmaz and Aslam, 2006), (Yilmaz and Aslam, 2008).

The modified version of the induced AP based on our methodology is shown below. Here are the documents from the pooled list considered for the evaluation purpose. It is done by ignoring the irrelevant documents from the pooled list and calculating the average precision to find the system performance and through that find out which system are producing the better results.

Given a topic T with R relevant documents in the pooled list P, induced AP can be calculated as

$$\text{Induced AP}=\frac{1}{R}\sum\nolimits_r \frac{number\ of\ relevants\ up\ to\ rank\ (r)}{rank(r)} \qquad \textbf{Equation 3.4}$$

Where *r* is a relevant document and *rank(r)* is the rank assigned for the document of a retrieval system

(Yilmaz and Aslam, 2006), (Yilmaz and Aslam, 2008)

Based on the evaluation score, the systems were classified into good contributing and less contributing systems. The documents from the good contributing systems were considered as pooled documents by reshuffling the pooled list and the documents from the low participated systems were considered as irrelevant documents and considered as unjudged documents. These unjudged documents from the less participated systems were plotted either by using clustering or classification techniques. The clustering technique is based on an agglomerative hierarchical clustering model and by the k-means clustering technique by considering the similarity of the documents. The next step is to find the similarity of the documents between the pooled list and the unjudged list. Next, find the highest scored document from the pooled list. Find the similarity between the p_i and the unjudged clustered documents. Like d_i is checking the similarity with p_1, based on considering the topic title using the TF-IDF technique. The same p_1 will do the similarity checking with all the top documents in each cluster. Find out the cluster which has the highest similarity score. The top-k documents from that cluster will be assigned a new score and move those documents into the pooled list.

$$\text{New Score assigning}= \frac{sim(di,U_{Cij})}{\sum_{l=1}^{k} sim(di,U_{Cij})}$$

(Liang et al.,2018)

The same process continues with the second high-ranked document from the pooled list. Once the similarity is found, those top-k clustered documents or pooled documents also move into the judgment list by assigning new scores. The same process continues with all the pooled documents. Once the iteration is completed, the pooled documents will be sent for the evaluation process. The overall framework of the proposed methodology is shown in Figure 4.12.

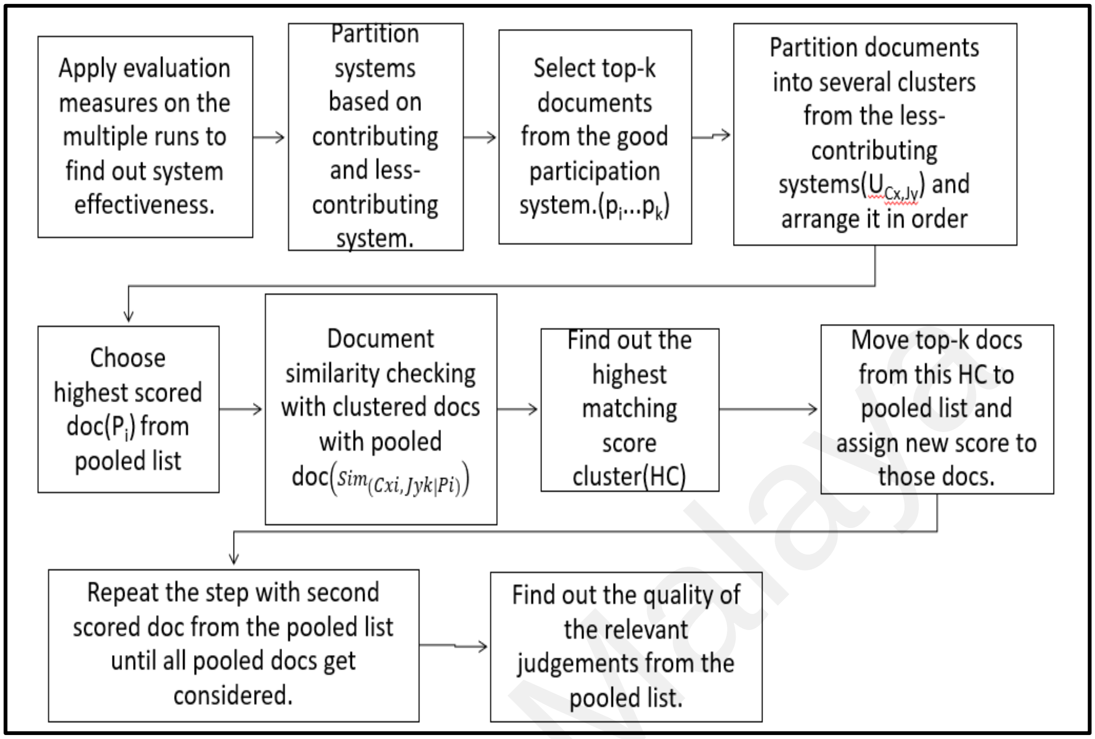

Apply evaluation measures on the multiple runs to find out system effectiveness.
Partition systems based on contributing and less-contributing system.
Select top-k documents from the good participation system.($p_i ... p_k$)
Partition documents into several clusters from the less-contributing systems($U_{Cx,Jy}$) and arrange it in order
Choose highest scored doc(P_i) from pooled list
Document similarity checking with clustered docs with pooled doc($Sim_{(Cxi,Jyk|Pi)}$)
Find out the highest matching score cluster(HC)
Move top-k docs from this HC to pooled list and assign new score to those docs.
Repeat the step with second scored doc from the pooled list until all pooled docs get considered.
Find out the quality of the relevant judgements from the pooled list.

Figure 4.12: Experimental Methodology framework based on pooling and document similarity with evaluation scores using clustering technique

The same experiment was done with the classification techniques. For clustering, k-means clustering, and hierarchical clustering techniques have been used. For classification, active learning classification techniques have been used. The documents were classified based on the similarity of the documents in the unjudged list. Document similarity checking between the pooled document list and each class has been done. Which class has produced the high similarity score, top-k documents from this class have moved into the pooled list. The second high-scored document from the pooled list has continued with the same process. Figure 4.13 shows the overall framework of the proposed methodology using the classification technique.

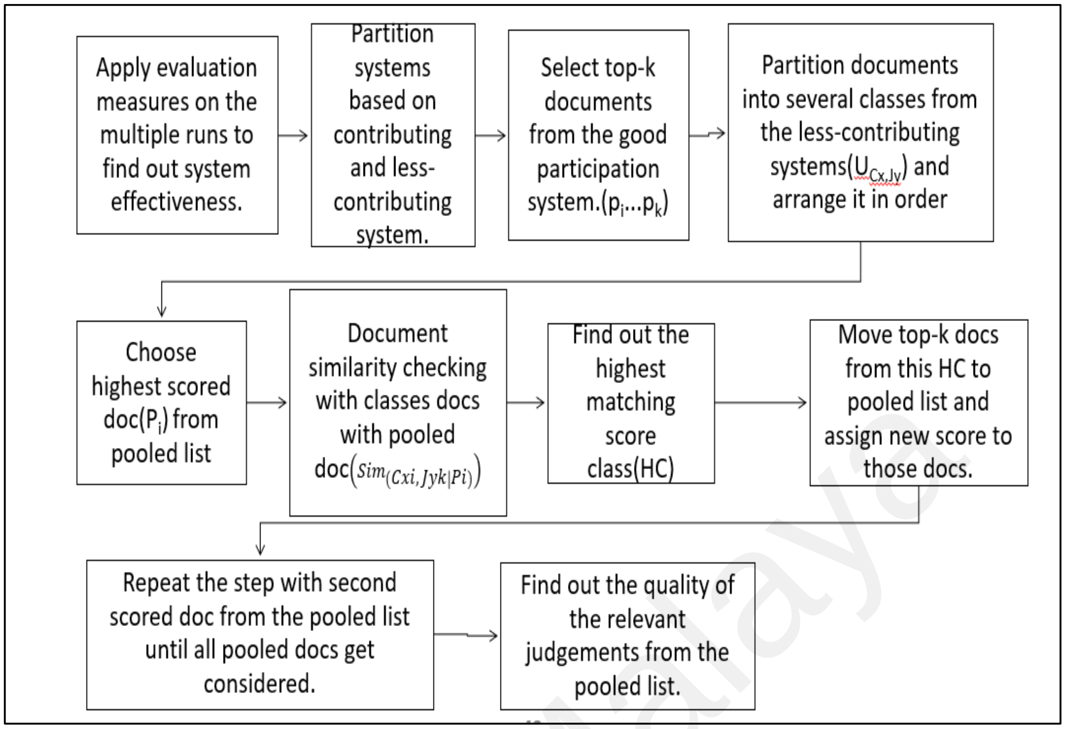

Apply evaluation measures on the multiple runs to find out system effectiveness.
Partition systems based on contributing and less-contributing system.
Select top-k documents from the good participation system.(p_i...p_k)
Partition documents into several classes from the less-contributing systems($U_{Cx,Jy}$) and arrange it in order
Choose highest scored doc(P_i) from pooled list
Document similarity checking with classes docs with pooled doc($Sim_{(Cxi,Jyk|Pi)}$)
Find out the highest matching score class(HC)
Move top-k docs from this HC to pooled list and assign new score to those docs.
Repeat the step with second scored doc from the pooled list until all pooled docs get considered.
Find out the quality of the relevant judgements from the pooled list.

Figure 4.13: Experimental Methodology framework based on pooling and document similarity with evaluation scores using classification technique

The step-by-step flow of the methodology is shown here.

- 1. Apply Induced AP@100 to the pooled list.
- 2. Based on the score, find out good contributing systems and less contributing systems.
- 3. Reallocate the pooled list by choosing top-k documents from the good participation systems into the pooled list
- 2. Using K-means clustering or active learning classification to partition unjudged docs (from less contributing systems) based on similarity using TF-IDF
- 3. Choose the highest-scored doc from the pooled list

- 4. Do document similarity checking with this high-scored pooled doc with each top-k document from each cluster or class.
- 5. Find out the highest-scored cluster or class that matches with that doc.
- 6. Move top-k documents from this HC to the pooled list
- 7. Repeat the step with the second high-scored doc from the pooled list until all pooled docs are considered.
- 8. Once all the iterations are done, the pooled list is given for the evaluation process.

4.3 Evaluating the effectiveness of the proposed evaluation methodology in terms of incompleteness and biasness in ranking

The basic assumption of the Cranfield paradigm is that the relevance judgment sets are complete. i.e., every document based on each topic is relevant. For smaller test collections, this assumption might be correct. But for large test collections like TREC, this assumption might not be true. The results might be closer to the completeness. To know the degree of completeness, an evaluation of retrieval systems is needed. TREC-based test collection is one of the most commonly used methodologies to evaluate retrieval systems. In this methodology, three components are there, document collection, topics, and a set of relevant documents. (Yilmaz and Aslam, 2006).

To avoid the judgment of the entire collection, the pooling method was used. Depth-100 pooling means, that only the top 100 documents for each topic from each system run were considered for the judgment process. Other remaining documents considered are irrelevant. The standard assumption of the test collection is that with this depth itself, most of the relevant documents can be achieved and can maintain the effectiveness of the systems (Harman,1995) (Zobel,1998). Even though even after pooling, relevant judgment sets are incomplete and biased. Much research has been done to address this issue with the large test collections. Many evaluation measures are there to find out the biasness or incompleteness of the judgment sets when the number of documents in the judgment set is limited (Yilmaz and Aslam, 2006).

4.3.1 Research approach

Pooling has been used to identify a subset of relevant documents from the multiple-ranked list to reduce the human accessor's effort in system-based evaluation. Still, with the help of evaluation metrics, it has been proven that many relevant documents have not moved into the pooled list due to many reasons. If most of the relevant documents have not moved into the judgment set, then that set can be called an incompleteness judgment. Much research has been done to reduce the incompleteness. (Buckley and Voorhees,2004) shows that the most commonly used evaluation measures such as average precision, R-precision, Precision@10, and MAP are not robust to the incompleteness relevance judgment. They proposed one of the best evaluation measures to find the incompleteness effectiveness is with bpref@k. This measure is finding faulty documents, which means judging irrelevant documents that are assigned with a higher rank than the documents that are relevant. This evaluation is done only with the judgment set documents.

For a topic *T*, with *R* relevant documents where *r* is a relevant document and *n* is the first non-relevant document of the of R which retrieved by the systems S, bpref is calculated as

$$bpref = \frac{1}{R}\sum_r 1 - \frac{|n\,ranked\,higher\,than\,r|}{R}$$ **Equation 4.3**

Based on equation 4.3, bpref works as follows Figure 4.14 and Figure 4.15 shows the performance of the two systems. Each system produced a set of documents which are ranked according to their relevancy. Assume 10 judged documents,4 Relevant docs (R), and 6 Irrelevant docs (N), eval depth=5

System A	**Ranks**				
	1	**2**	**3**	**4**	**5**
Relevance vector r	1	0	0	0	1
$|dn\,ranked\,higher\,than\,dr|$	0	-	-	-	3

$$bpref@5 = \frac{1}{4}\left[\left(1-\frac{0}{4}\right)+\left(1-\frac{3}{4}\right)\right] = 0.312$$

Figure 4.14: Bpref calculation for system A

In this example, assume R=4 and N=6 and evaluation depth=5. Bpref is calculated based on the relevant documents which are represented as “1” in the relevance vector. The quantity |*dn* ranked higher than *dr*| shows the number of non-relevant documents judged with a high rank before the actual relevant document.

For system B, with the same number of relevant and non-relevance as same as system A, System B retrieved the document and ranked as shown below.

System B	**Ranks**				
	1	**2**	**3**	**4**	**5**
Relevance vector r	1	1	0	1	0
$\lvert dn\ ranked\ higher\ than\ dr \rvert$	0	0	-	1	-

$$bpref@5 = 0.687$$

Figure 4.15: Bpref calculation for system B

As per the above two systems' performance, System A assigned higher ranks for the documents which are irrelevant so that the bpref score becomes lesser. System B, performed better in ranking the documents correctly. Hence, it scored a good bpref result.

Average Precision and Mean Average precision have been used to measure the effectiveness of the systems. When the relevance judgment is complete, bpref will be highly correlated with average precision. But bpref is a better evaluation metric than Ap when compared to the incompleteness of the relevance judgment (Buckley and Voorhees, 2004). But sometimes when bpref deviates from actual precision more incompleteness happens in the relevance judgment sets. Other evaluation measures such as induced AP, subcollection AP, and inferred AP have been proposed to overcome this incompleteness (Yilmaz and Aslam,2007), but it is highly correlated in terms of system rankings when the judgment set is complete. In this research bpref measures have been used to evaluate how the proposed methodologies work well in the matter of incompleteness.

4.3.2 Research framework and methodology

This section highlights the issues of biasness in the pooled list. It calculates the bias of the pooled list and adjusts the scores of the documents from the unpooled documents which matches the similarity with the pooled documents. For the score adjustment, the whole system is not considered, only the documents from the pooled list are evaluated. The relevant documents from the unpooled systems are not considered in the pooled list due to the ranking biasness. The systems have assigned lower ranks to the documents which are actually relevant. The ranks assigned for the unpooled systems are not considered in this evaluation. If a similarity of any document is found with pooled documents, the new score is assigned to it based on the pooled document rank. Similarity checking between pooled documents and unjudged documents has been done. If a similarity is found, a new score is assigned based on the relevant documents in the pooled list scores. In this pooled list the documents from the good participating systems are considered as pooled lists and documents from the less contributing systems are considered as unjudged document lists. Once all the document similarity with the pooled document list is done, the mean score of the pooled list and the unpooled list is calculated. Based on this mean score biasness in ranking is calculated. Various pool depth and qrels sets have to be considered and the experiment has to be repeated multiple times to know the variation in the biasness in ranking with different qrels sizes. The various qrels size will be used until less biasness is found in the differences in the pooled and unjudged document lists. Bpref measures have been used for the experiment and this measure works effectively with the incompleteness of the judgment sets. Previous results show that as the qrels size increases bpref value becomes inconsistent. But with reduced qrels, bpref works well to show the incompleteness of the judgment sets. The overall framework of the biasness calculation in terms of ranking is shown in Figure 4.16.

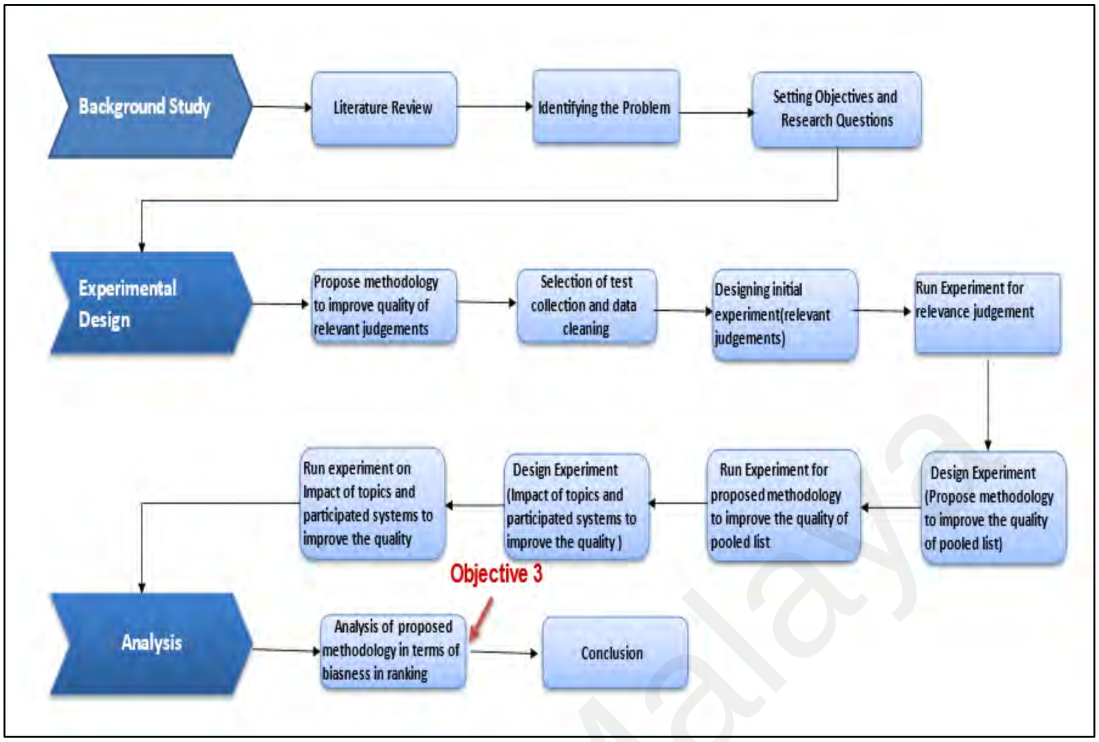

Background Study
Literature Review
Identifying the Problem
Setting Objectives and Research Questions
Experimental Design
Propose methodology to improve quality of relevant judgements
Selection of test collection and data cleaning
Designing initial experiment(relevant judgements)
Run Experiment for relevance judgement
Design Experiment (Propose methodology to improve the quality of pooled list)
Run Experiment for proposed methodology to improve the quality of pooled list
Design Experiment (Impact of topics and participated systems to improve the quality)
Run experiment on Impact of topics and participated systems to improve the quality
Objective 3
Analysis
Analysis of proposed methodology in terms of biasness in ranking
Conclusion

Figure 4.16: Research methodology flow diagram

The measures that depend only on the ranks of the relevant documents have used evaluation measures such as MAP and P@10. The biasness in the incompleteness of the judgment set have calculated with the bpref measure. Figure 4.17 shows how the relevance judgments are performed in the matter of relevant documents' effectiveness and also the matter of incompleteness of the judgment sets.

For this experiment, the TREC-8 dataset was used with 528k documents (1.9GB) 50 topics (401-450), and 124 runs. For this experiment, various qrels sizes have been used to measure the effectiveness changes as the qrels value varies. 6 qrels sets were used for the experiments, The results show that as the qrels size increases, MAP and P@10 value increase while at the same time, as the qrels size increases bpref value decreases. It happens due to the incompleteness in the relevance judgments as the document size increases. Ranking inconsistency is very much increasing as the relevance judgment size is increasing. This makes the performance effectiveness of the participating systems go down and affects the accuracy of the evaluation process.

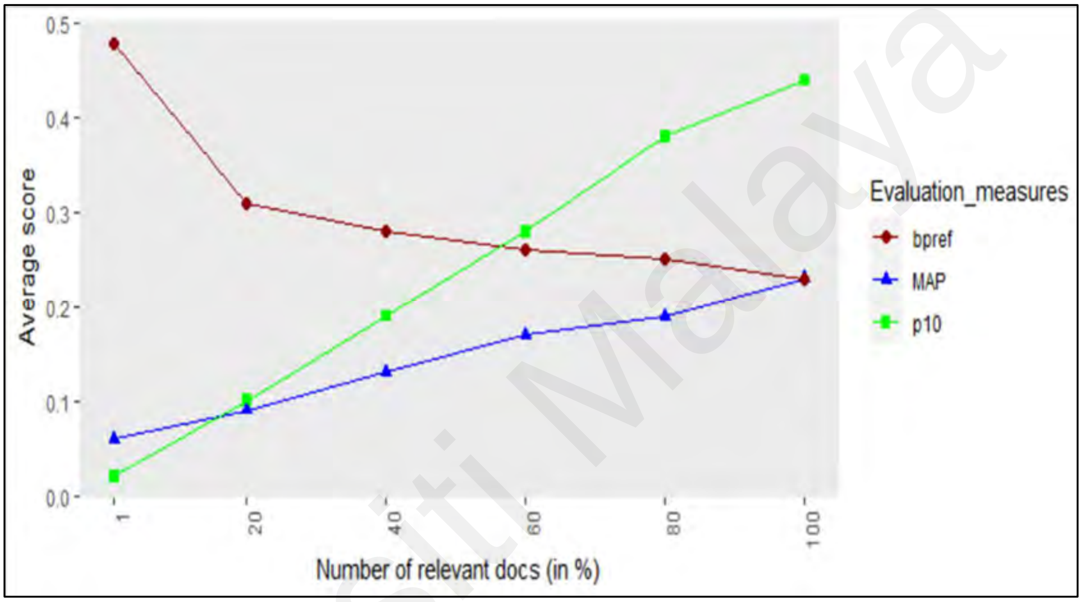


Figure 4.17: Changes in the average score of the evaluation measures based on different judgments set in TREC-8

As the bpref measure decreases, the inconsistency occurs. Consistency of the scores is an important feature for the practical implementation of the test collection. To maintain consistency, we need to increase the relevant judgment sets quality, for that, we need to adjust the order of the documents by assigning higher ranks to the unjudged documents and move into the pooled list.

Algorithm 1 shows how the biasness is calculated with pooled and unjudged document lists and how the biasness is measured. The algorithm used here aims to show how the similarity between pooled and clustered unjudged documents has found the similarity of how a new score is assigned to a similar document found in the judgment list and how these documents have moved to the pooled list with the new score assigned. The steps continue with different qrels sets until the biasness is lesser. The biasness in the relevance judgment set is calculated with bpref measure.

Algorithm 1: Reduce the biasness in ranking by assigning new scores to the unjudged documents

```
$T \leftarrow set\ of\ Topics\ from\ pooled\ list$
$P \leftarrow set\ of\ pooled\ docs\ from\ systems$
$(U_{Cx,Jy}) \leftarrow clustered\ unjudged\ documents$
$Q$
$\leftarrow set\ of\ qrels\ on\ T\ based\ on\ S\ from\ pooled\ list$
$P' \leftarrow set\ of\ unique\ pooled\ docs$
$Q' \leftarrow set\ of\ qrels\ with\ different\ size$
for $Q \in Q'$ do
  for $p \in P'$ do
      $Sim_{(Cx,Jy|p)}$                                  Find the similarity of the cluster x unjudged document Jy with the pooled document
  if matches found, $Score_{Jy} = \frac{sim(p, C_{x,jy})}{\sum_{l=1}^{k} sim(p_l, U_{Cjy})}$        New score assigning based on pooled doc rank
$P' \leftarrow C_{x,Jy} + P'$
repeat until $p \in P'$ docs consider
   end for
$P_s \leftarrow mean\ of\ pooled\ score$
```

```
$(U_{Cx,Jy})_s \leftarrow \boldsymbol{mean\ of\ unpooled\ score}$
$\boldsymbol{\beta_s} \leftarrow \boldsymbol{P_s} - (\boldsymbol{U_{Cx,Jy}})_s$                    Calculate biasness
return $\boldsymbol{\beta_s}$
repeat until enough $\boldsymbol{Q} \in \boldsymbol{Q'}$ consider
end for
Calculate bpref measure and MAP
```

CHAPTER 5: RESULTS AND DISCUSSIONS

The main aim of this thesis is to increase the quality of the judgment sets by increasing the number of relevant documents in the judgment sets and through that increase the accuracy of the evaluation process. For that, a methodology has been proposed and evaluated based on the baseline works. This section covers the results and discussions of the methodologies proposed in Chapter 3 and Chapter 4. Section 5.1 covers the performance of the methodology proposed in Chapter 3 has been evaluated by comparing it with the baseline methodologies. The effectiveness of this methodology has been evaluated using various evaluation metrics by considering different pool depths and evaluation depths. and also found out which clustering technique might perform better results in the matter of increasing the quality of the judgment sets. Section 5.2 covers the effectiveness of the improved proposed methodologies from Section 4.1 and Section 4.2, based on reduced topic size and considered documents from systems based on evaluation scores also explored in this session. Section 5.3 covers the effectiveness of the proposed methodologies in order to reduce the biasness in ranking documents compared to the baselines and at the same time how effectively increased the quality of the judgment sets are shown in detail section 4.3.

5.1 Improving the accuracy of the Information Retrieval Evaluation process by increasing the number of relevant documents

5.1.1 Results and Discussions

The methodology by combining pooling and document similarity techniques, proposed in Chapter 3 has mainly aimed to retrieve the maximum number of relevant documents into the judgment list and through that increase the quality of the judgment sets. For the document similarity checking, both classification and clustering techniques have been used. For the pooling technique, only pooled documents are considered for the evaluation process. For the experiment, baseline works have been used for the comparison and evaluation of the performance of the proposed methodology. For this baseline experiment, three methodologies were considered. 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. Document similarity has been done 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 in the cluster. Whenever a new user query comes, the patterns discovered in each cluster and find out the most relevant clusters that match the user query and the clustered documents are considered for the evaluation process (Djenouri et al.,2021, Djenouri et al.,2018). The classification-based methodology, namely CAL (Continuous Active Learning) which 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 documents that are unjudged (Rahman et al.,2020). The performance of these methodologies is shown in Figure 3.1.

For the experiment, TREC dataset collection was used for evaluation purposes. TREC-8 and TREC-10 collections have been used. The ad hoc retrieval task investigates the performance of the participating systems that search a static set of documents using topics or queries (Hawking et al., 1999). And Web Track mainly focused on the topic relevance task(Hawking and Voorhees, 2001). Participants have used only documents, topics, and relevance judgments from previous TREC were used to generate their runs. The details of each TREC collection used in this experiment are shown in Table 5.1.

Table 5.1: Datasets Overview

Dataset	**TREC-8**	**TREC-10**
Track	Adhoc Track	Web Track
Number of Topics	50	50
Topics	401-450	501-550
Total systems	129	97
Systems considered	124	77
Document collection in numbers	528k	1700k
Document collection in GB	2 GB	10GB

The baseline work implementation is shown in Figure 3.1. The proposed methodology has been implemented based on pooling and document similarity techniques using clustering and classification techniques. The results of the proposed methodologies (both based on clustering and classification) compared to the baseline works are shown below. Figure 5.1 shows the performance of the proposed methodologies based on the TREC- 8 test collection. The proposed methodologies were compared with the baseline works such as **base_cluster-(ICIR**) -(Djenouri et al.,2021), **base_classification-(CAL)-** (Rahman et al.,2020), **pooling**-(Losada et al.,2018). The proposed methodology aimed to find out how many relevant documents were retrieved with different numbers of judgments. The x-axis shows the number of relevant documents considered and the y-axis shows the number of relevant documents retrieved in percentages. The experiment was done with pool depth=100 and evaluation depth=1000.

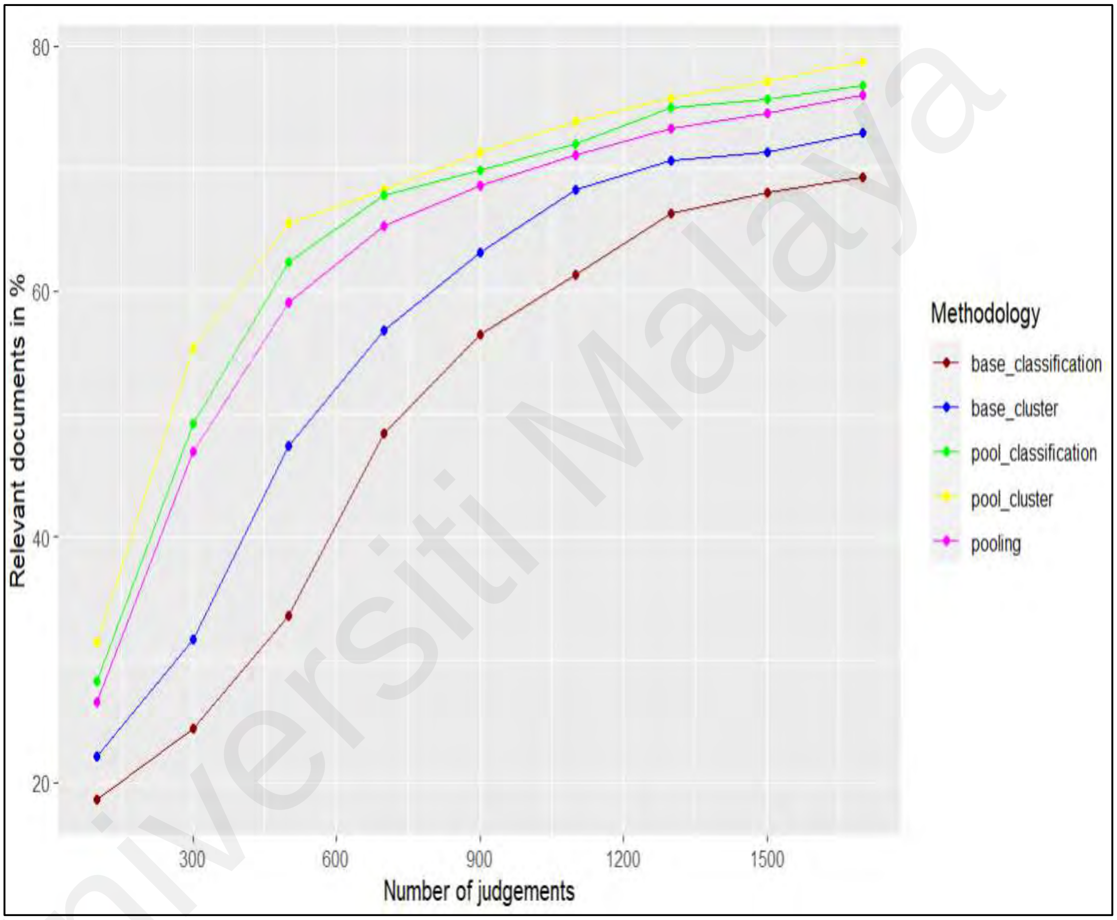


Figure 5.1: Relevant documents retrieved using various methodologies (in %) using the TREC-8 dataset

base_cluster-(ICIR) -(Djenouri et al.,2021), **base_classification-(CAL)-** (Rahman et al.,2020), **pooling**-(Losada et al.,2018**), pool_classification-** (proposed methodology based on classification technique), **pool_cluster**- (proposed methodology based on clustering technique)

The same experiment of the proposed methodologies was done with TRCE-10 dataset collection and was compared against the baseline works. The performance of the proposed methodologies in retrieving the number of relevant documents based on the TREC-10 collections is shown in Figure 5.2. The experiment was done the same way as the TREC-8 test collection with pool depth =100 and evaluation depth = 1000.

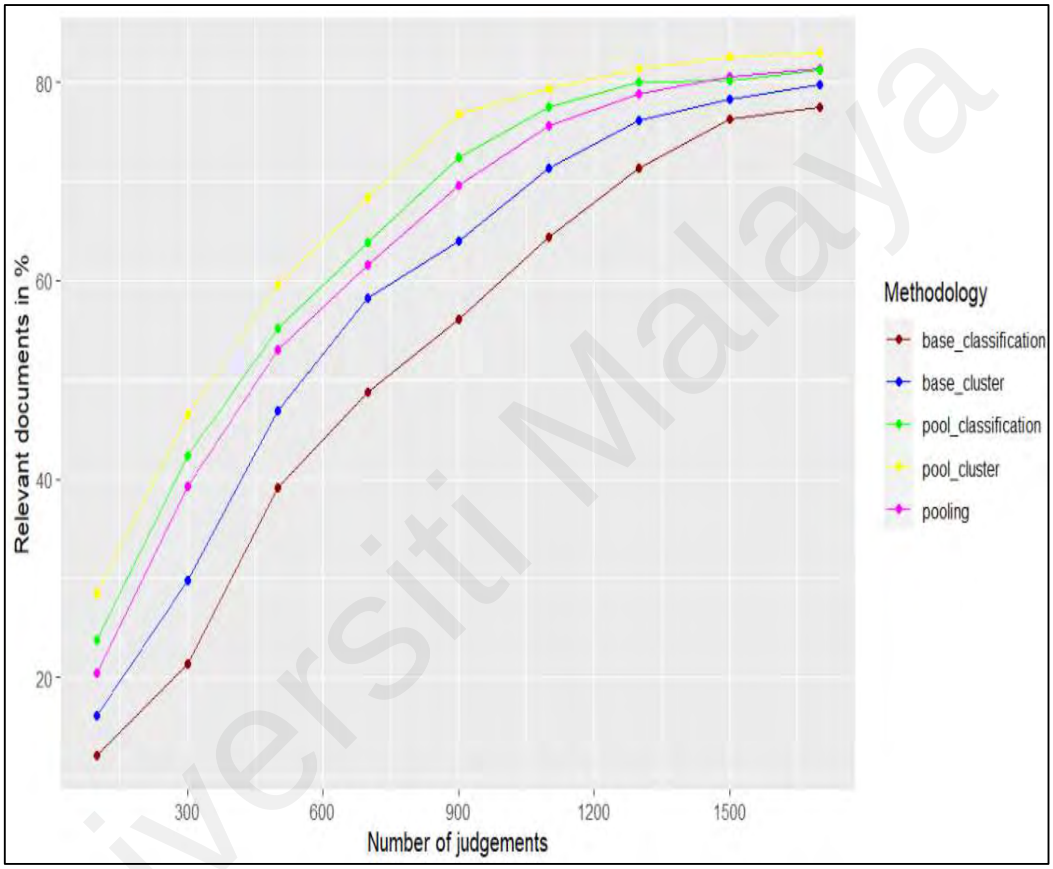


Figure 5.2: Relevant documents retrieved using various methodologies (in %) using the TREC-10 dataset

base_cluster-(ICIR) -(Djenouri et al.,2021), **base_classification-(CAL)-** (Rahman et al.,2020), **pooling**-(Losada et al.,2018**), pool_classification-** (proposed methodology based on classification technique), **pool_cluster**- (proposed methodology based on clustering technique)

The results show that the proposed methodologies, based on clustering and classification techniques outperformed compared to the baseline line works. This means the proposed methodology could be able to retrieve a greater number of relevant documents into the judgment sets even with the lesser pool depth. The results show the same with both the test collections. And the methodology helped to retrieve more relevant documents even with the lesser pool depth which is cost-effective and more efficient.

The clear view of the number of relevant documents retrieved by the proposed methodology based on pooling and document similarity using clustering compared to the baseline works such as **clust (ICIR**)-(Djenouri et al.,2021) and **Pool Diff**-(Losada et al.,2018) is shown in Figure 5.3.

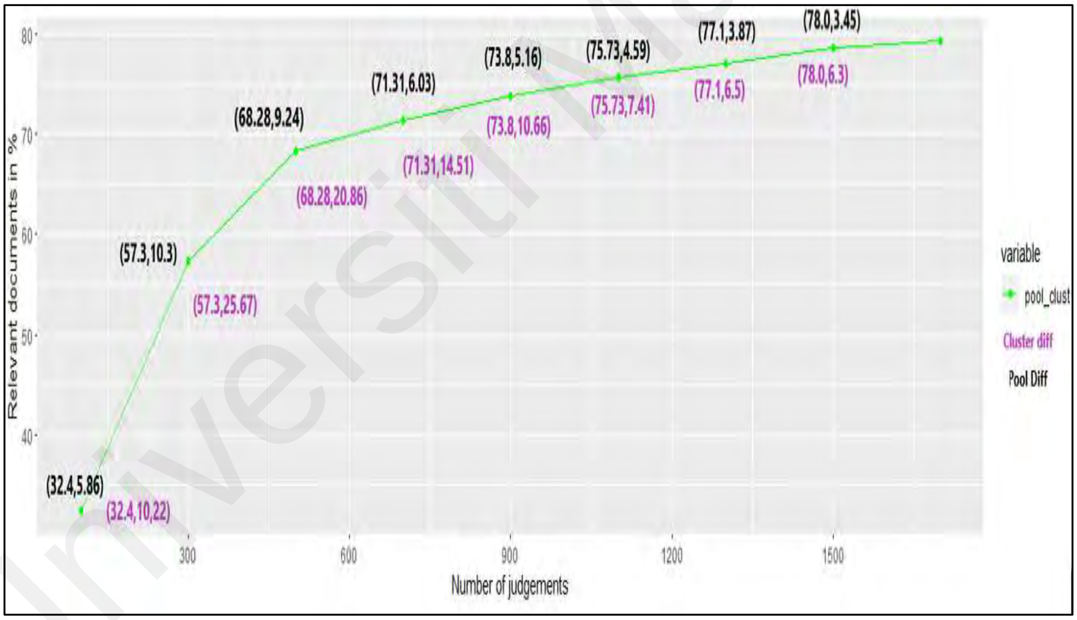


Figure 5.3: Additional number of relevant documents retrieved using a proposed methodology (in %) using the TREC-8 dataset

clust(ICIR)-(Djenouri et al.,2021), **Pool Diff**-(Losada et al.,2018**), pool_clust -**

(proposed)

Cluster diff shows how many relevant documents were retrieved in percentages by the proposed methodology and how many extra documents were retrieved compared to the cluster (ICIR) methodology. For example, In TREC-8 data collection, for the top 300 documents, 57.3% of relevant documents were retrieved by the proposed methodology and compared to the baseline work based on clustering, ICIR, 25.67% of extra documents were found. But with the top 900 relevant documents. 73.8 % of relevant documents were found, but only 10.66% of extra documents only found. It has been noticed that these proposed methodologies were able to retrieve most of the relevant documents with lesser pool depth compared to ICIR.

Pool diff shows how many relevant documents were retrieved by the proposed methodology in percentages and how many extra documents were retrieved compared to the pooling methodology. Compared to previous works, the difference between the proposed one and the pooling technique difference is not so much, but still, this methodology helped to achieve better results. It's because only the documents from the unjudged list were considered and those found as relevant based on pooled documents, moved into the pooled list.

Various evaluation measures have been used to evaluate the performance of the system based on the proposed methodology. Different evaluation measures have different criteria and different properties based on the user's satisfaction. MAP, Mean Average Precision is calculated based on the mean of all the average precision of topics over all the topics. Mean average precision is the mean of the precision values scored after each relevant document was found and at the same time by considering relevant documents that are not considered by ignoring it and assigning value 0 for it.

The performance of the methodologies in order to retrieve the number of relevant documents into the judgment sets is shown in Table 5.2. The comparison of the proposed methodologies with clustering and pooling with baseline works has been done. The average precision of topics overall topics was calculated based on the pool depth of 100 and evaluation depth of 1000. Later, mean average precision is calculated based on the mean of the average precision of the topics over all the topics were done. The experiments have been done with both TREC-8 and TREC-10 test collections. The results were ordered according to the ascending order of their performance scores.

Table 5.2: Mean Average Precision (MAP) results

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

The results show that the performance of the proposed methodology is performed well compared to the baseline works. Pooling+Classification methodology outperformed Classification, CAL technique. And the same way, the Pooling+Clustering methodology outperformed the Clustering, ICIR technique. Both these methodologies performed better than the traditional Pooling method.

Metric stability has a greater impact on the methodologies' performance in system evaluation. That is the consistency of the metrics results from one set of topics or systems to another set. Discounted cumulative gain (DCG) calculated with each rank of the document has a fixed weight generally called a discount, which is multiplied with multi-graded relevance, also called gain of the document (Järvelin and Kekäläinen,2002). Later, they included normalization, Normalized Discounted Cumulative Gain (NDCG), to normalize each topic score with the ideal score. Mean Average Precision assumes binary relevance. Graded relevance is done by NDCG in which the ranking score is related to the relevance vector. A detailed explanation of the NDCG calculation of Equation 5.1 is shown in Equation 2.6 in the previous chapters.

Equation 5.1

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

$$\text{were,} IDCG@k = \sum_{i=1}^{K} \frac{G_i}{\log_2(i+1)}$$

The effect of evaluation depth and the pool depth on metric behavior is evaluated here using NDCG. The correlation of the NDCG metric over pooled depth 10 and pool depth 100 over various evaluation depths using TREC-8 test collection has been done in Figure 5.4. Various combinations of evaluation and pool depth on TREC-8 data collection have been considered here. Pool depth and evaluation depth of 10 are considered as shallow, pool depth and evaluation depth of 100 is considered as deep, and pool depth of 10 and 100 with evaluation depth of 1000 are considered as extended. The correlation of pool depth and evaluation depth with 10 and pool depth with 10 and different evaluation depths have been plotted here. Also, the correlation of pool depth and evaluation depth with 100 and pool depth with 100 and different evaluation depths also plotted in this graph to know the difference of systems performance in shallow, deep, and extended depths.

The results show that if the pool depth is greater than the evaluation depth and the pool depth is equal to the evaluation depth, the correlation score increases. Also, noticed that when the pool depth is lesser than the evaluation depth, a decrease in the system correlation can be found. This indicates that for this test collection, extending the evaluation depth to 1000 on a pool depth of 10 and 100 gives a reliable result that the ranking has been misled by some irrelevant documents in the relevance judgment sets.

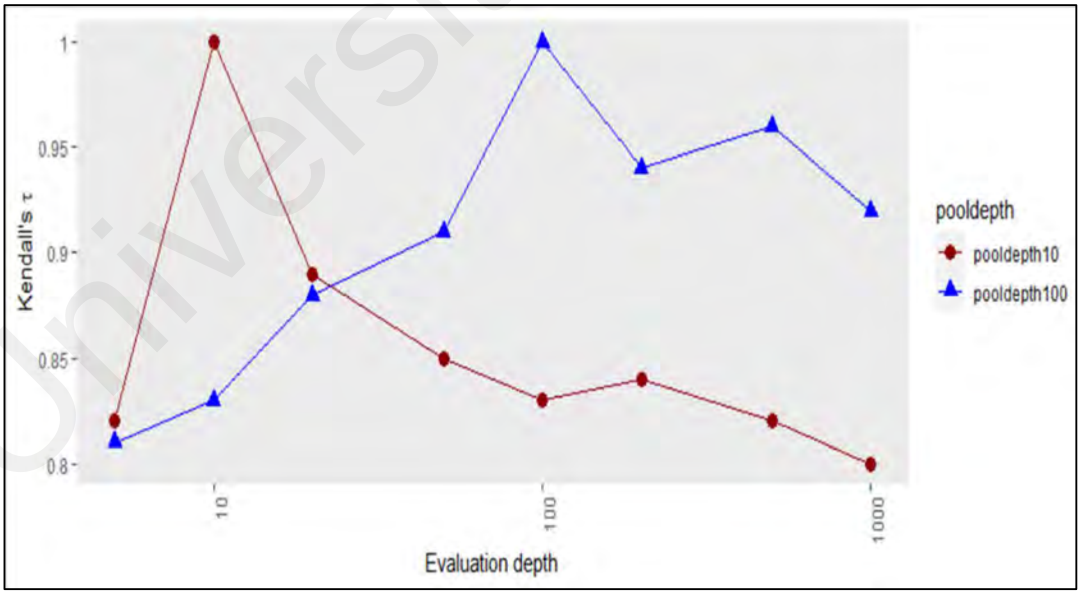


Figure 5.4: NDCG@d, where pool depth, d=10 and 100 using TREC-8 dataset

NDCG scores are between a defined range due to the ranks weight in logarithmic is not convergent. A convergent and geometric weighted metric called Rank Biased Precision, Rbp. In the Rbp metric, a parameter named *p* represents the user behavior on the probability of proceeding to the next rank (Moffat & Zobel, 2008). It shows how the user examines each document from the top of the list and will proceed to the next document with the probability of *p* or finish the searches with the probability of *1-p* (Park & Zhang, 2007).

RBP is calculated as shown in Equation 5.2

Equation 5.2

$$\boldsymbol{RBP(p) = (1-p)\sum_{i=1}^{d} r_i \,. p^{(i-1)}}$$

where $r_i \in [0,1]$, relevance judgment of ith element, and (1-*p*) is the factor used to scale the RBP measure within [0,1]. If the user is with low persistence, (close to 0) means the user not likely to examine after the first document, and high persistence, (close to 1) means users might examine many documents.

NDCG and Rbp are different in their ranking weights. Rbp is not dependent on the evaluation depth. DCG weights go down sharply once the evaluation depth goes down. Rbp weights go down in proportion at each rank. The correlation of the Rbp metric with p=0.977 over pooled depth 10 and pool depth 100 over various evaluation depths using TREC-8 test collection have been done in Figure 5.5.

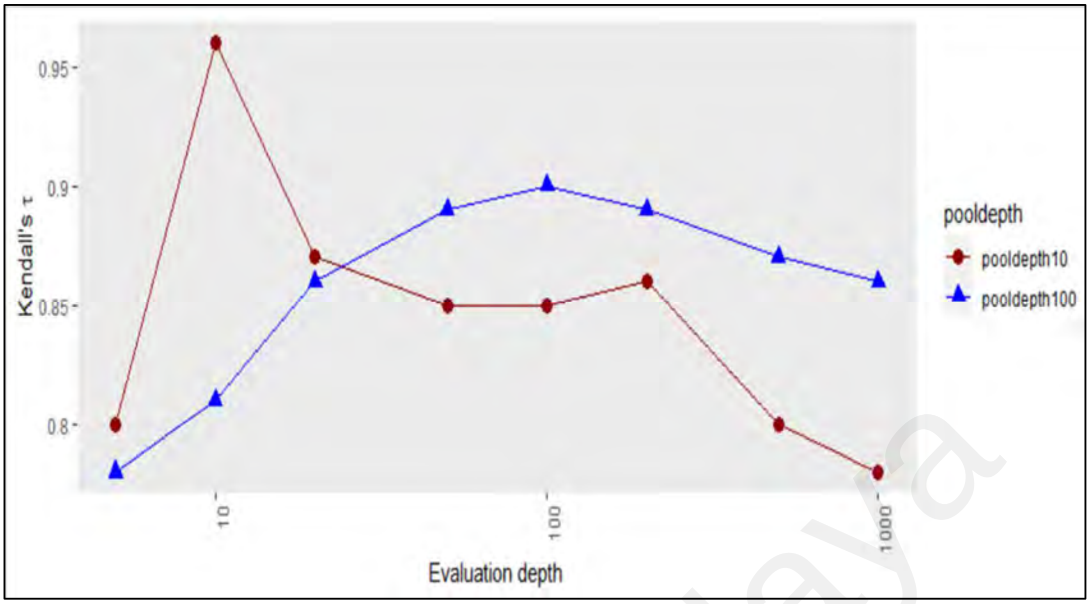


Figure 5.5: Rbp, where pool depth d=10 and 100 using the TREC-8 dataset

The results show almost similar to NDCG, that if the pool depth is greater than the evaluation depth and the pool depth is equal to the evaluation depth, the correlation score increases steadily. Also, noticed that when the pool depth is lesser than the evaluation depth, a slow decrease in the system correlation can be found. This decrease is almost proportional to the evaluation depth. This indicates that for this test collection, extending the evaluation depth to 1000 on a pool depth of 10 and 100 gives a reliable result that the ranking has been misled by some irrelevant documents in the relevance judgment sets.

Discriminative power is one of the measures to evaluate the stability of the evaluation metrics. Discriminative measures calculate the proportion of a set of evaluation systems, in which the difference in their effectiveness is found statistically significant (Sakai 2006b, 2007b). Based on the standard pool depth of 100 and the evaluation depth of 1000, with TREC-8 dataset collection with randomly paired systems, the comparison of the metric stability using discriminative measures is shown in Table 5.3.

Table 5.3: Discriminative power of various metrics on TREC-8 and TREC-10 collection

Metric	**TREC-8**	**TREC-10**
	Adhoc	Web
MAP@1000	0.768	0.694
ndcg@1000	0.741	0.689
Rbp, p=0.8	0.671	0.632
Rbp, p=0.95	0.706	0.641

The table shows that discriminative power varies based on TREC collection. The results show that ndcg got a more discriminative score than Rbp. The variation happens due to the different evaluation depths and pool depths. Based on different evaluation depths and pool depths, significant tests have been done. Here the significant test was done with a two-tailed t-test, paired with a significant level α=0.05. Figure 5.6 and Figure 5.7 shows the mean average precision metric stability measured using discriminative power. TREC-8 data collection was used with various pool depths and evaluation depths.

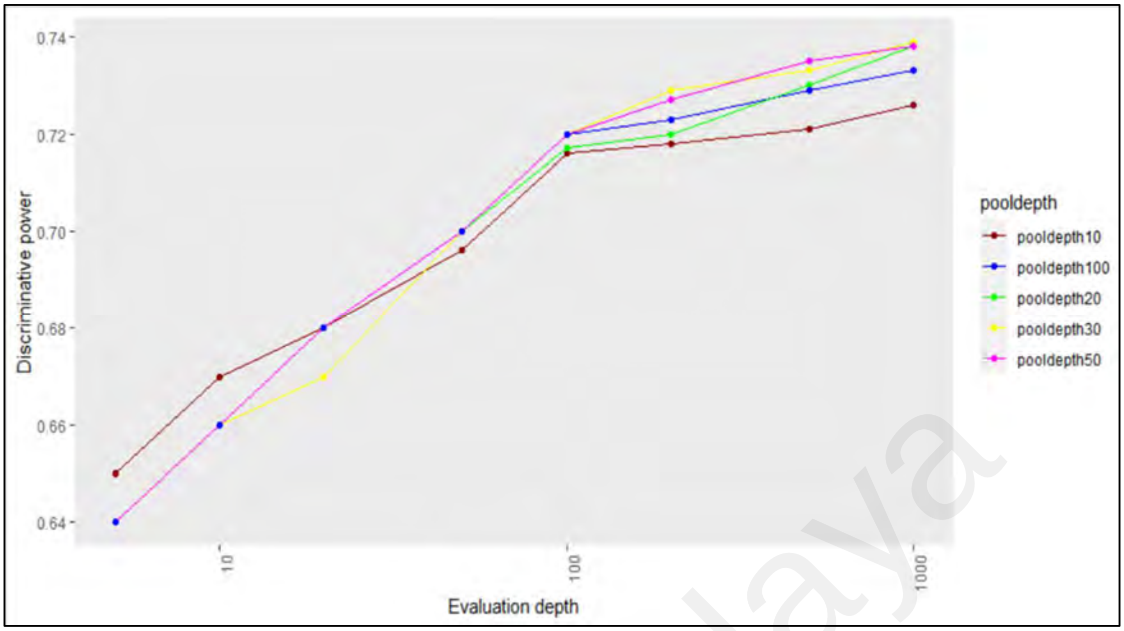


Figure 5.6: T-test: Mean average precision metric stability based on discriminative power - TREC-8 collection

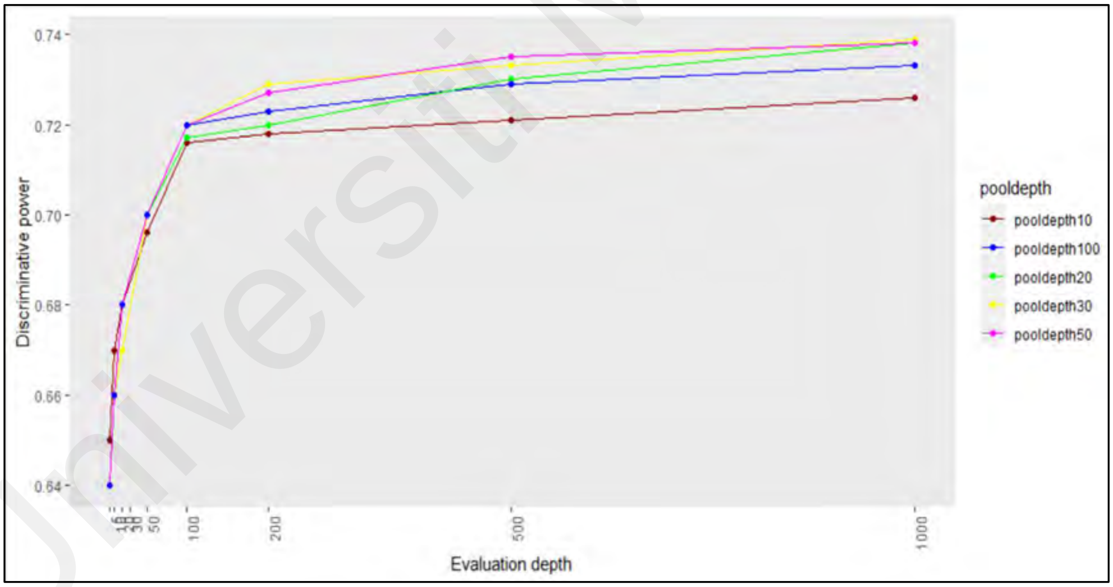


Figure 5.7: T-test: Mean average precision metric stability based on discriminative power (detailed evaluation depth) - TREC-8 collection

The results show that discriminative power increases steadily if the pool depth is greater than the evaluation depth and also when the pool depth is equal to the evaluation depth.

It is mostly because greater pool depth allows relevant documents retrieved to exceed the total number of relevant documents for more topics. But when the evaluation depth is greater than the pool depth and the number of relevant documents retrieved is greater than the pool depth, results show up in different ways significantly at α=0.05. This shows that pooling beyond evaluation depth affects the discriminative power. This is because deeper pooling increases the value of a total number of relevant documents without increasing the ability of runs to retrieve more relevant documents. However, the impact and importance of evaluating the deeper pool depth and also the importance of the evaluation depth need to be out. Table 5.4 shows Kendall's correlation for carious evaluation depth and pool depth using paired, two-tailed t-tests based on TREC-8 data collection shown in Table 5.4. Ranking based on paired test values is more similar to the same pool depth and evaluation depth. However, the correlation value is different with different evaluation depths and pool depths. However, as per the results with the same depths, the correlation values are higher compared to the different depths. When the comparison was done with the different depths, the pool depth of 10 and evaluation depth of 100 had a closer correlation with pool depth of 100 and evaluation with 100 than the pool depth of 10 and evaluation depth of 10. The same variation can be found extended with pool depths of 10 and 100 with various evaluation depths of 100 and 1000.

These results suggest that evaluating beyond pool depth still can retrieve more reliable results based on the proposed methodology also it has proven that evaluation depth has an important impact on the results compared to the pooled depth. It also shows that evaluation depth and pool depth are more important in the selecting of metrics to evaluate the system performance.

Table 5.4: Kendall's correlation between system pairs with various metrics with different evaluation depth and pool depth based on TREC-8 dataset collection

Pool depth	Eval depth	Metric	P@10 E@10			P@10 E@100			P@100 E@100			P@10 E@1000			P@100 E@1000	
			ndcg	Rbp	MAP	ndcg	Rbp	MAP	ndcg	Rbp	MAP	ndcg	Rbp	MAP	ndcg	Rbp
10	10	MAP	0.9	0.9	0.75	0.74	0.69	0.75	0.75	0.74	0.67	0.68	0.64	0.75	0.75	0.73
		ndcg		0.98	0.75	0.75	0.71	0.75	0.73	0.76	0.68	0.69	0.65	0.76	0.73	0.76
		Rbp			0.73	0.75	0.69	0.74	0.75	0.75	0.69	0.69	0.66	0.74	0.75	0.75
10	100	MAP				0.88	0.83	0.86	0.85	0.82	0.73	0.71	0.72	0.72	0.73	0.7
		ndcg					0.87	0.8	0.87	0.84	0.71	0.72	0.73	0.71	0.72	0.71
		Rbp						0.76	0.82	0.84	0.71	0.72	0.72	0.71	0.72	0.72
100	100	MAP							0.83	0.83	0.74	0.73	0.72	0.86	0.84	0.83
		ndcg								0.88	0.73	0.72	0.72	0.87	0.85	0.84

Based on the evaluation measures, it has shown that the proposed methodologies can performed better. In all these experiments, methodologies based on clustering have been done with k-means. Considering k-means clustering mainly because of the baseline work, ICIR has used clustering techniques with k-means and with k-5. The methodology has been implemented with agglomerative hierarchical clustering using average linkage also to know the variation or the similarity in the performance. Here the performance is evaluated with MAP value. The methodology has been compared with k-means and agglomerative hierarchical clustering with three TREC collections such as TREC-8, TREC-9, and TREC-10 shown in Figure 5.8, Figure 5.9 and Figure 5.10.

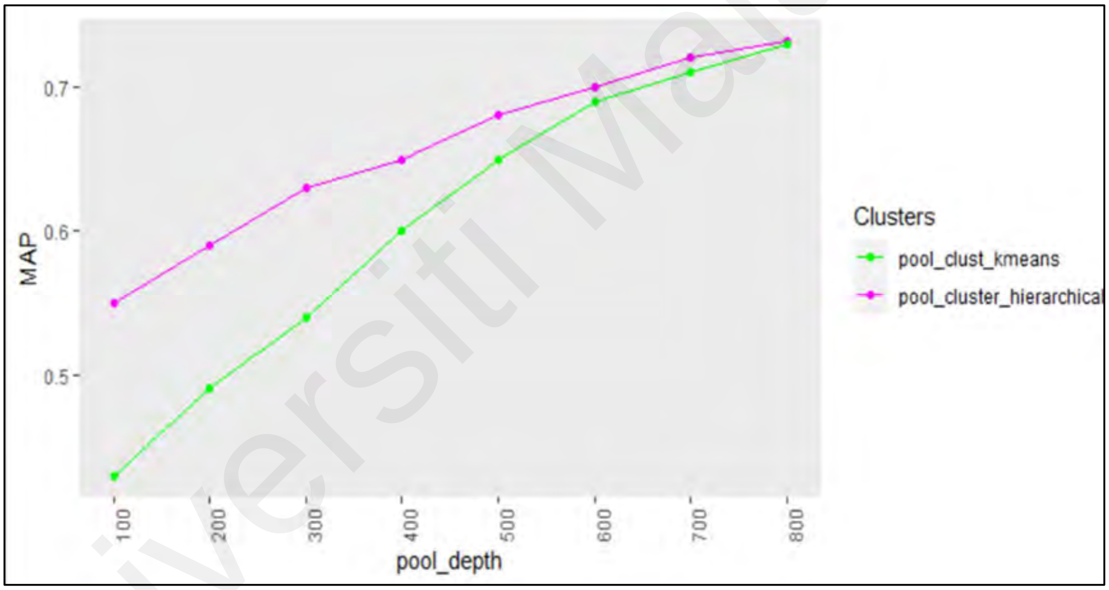


Figure 5.8: Comparison of clustering techniques on the proposed methodology based on TREC-8

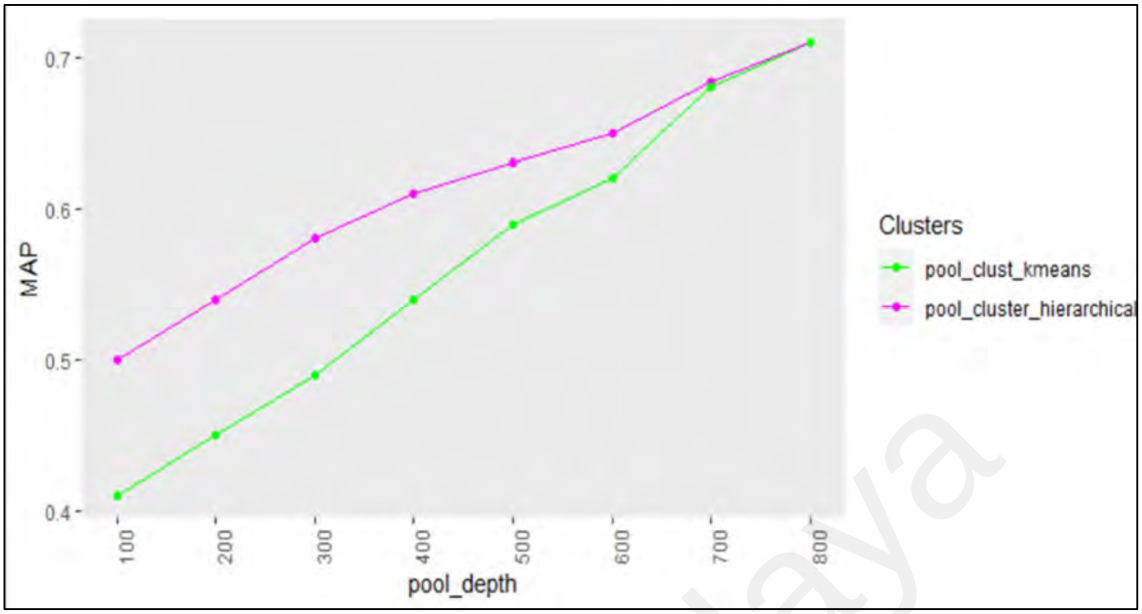


Figure 5.9: Comparison of clustering techniques on the proposed methodology based on TREC-9

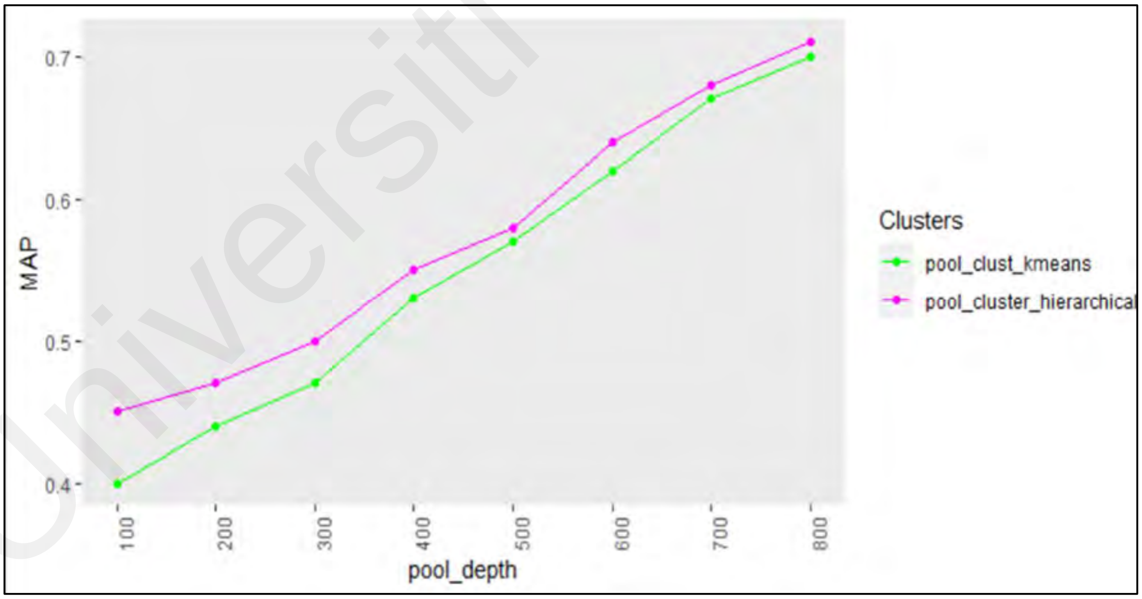


Figure 5.10: Comparison of clustering techniques on the proposed methodology based on TREC-10

The results show that the hierarchical clustering performed well on the proposed methodology compared to the k-means clustering. Also, it has shown that even with lesser pool depth, hierarchical clustering was able to retrieve a greater number of relevant documents. As the pool depth increases, the retrieval efficiency is almost the same with both clustering techniques. However as the number of clusters is not fixed in hierarchical clustering, the visualization of the dendrogram is time-consuming. So, in the remaining experiments, the k-means clustering has continued for evaluation purposes.

5.1.2 Significance and Contributions

The main aim of this objective was to increase the quality of the judgment sets in the information retrieval evaluation process. The proposed methodologies based on pooling and document similarity with clustering and classification techniques help to achieve a greater number of relevant documents in the judgment list and through that it helps to increase the quality of the judgment sets. Also, compared to the baseline works, the proposed methodologies were able to retrieve a greater number of relevant documents with lesser pool depth. It helps indirectly to improve the accuracy of the evaluation process by improving the system effectiveness score.

Also, through the evaluation measures, it has been noticed that the systems perform better based on different pool depths and evaluation depths. It has been noticed that the system performs better when the pool depth is greater than the evaluation depth and also when the pool depth is equal to the evaluation depth. It also has been noticed that when the number of relevant documents in the qrels is greater than the pool depth and when the evaluation depth becomes greater than the pool depth, the performance of the system varies significantly. It is due to the methodology implemented without considering increasing the ability of runs to retrieve more relevant documents within that pool depth.

Also, it has been noticed that in the clustering-based methodology, hierarchical clustering performed better compared to k-means clustering. It has shown that with lesser pool depth, the systems were able to retrieve a greater number of relevant documents based on hierarchical clustering.

5.2 Increasing the quality of relevant judgments by considering topics and participating systems from the test collections

This objective aims to find out how the topics and participating systems can effectively work on increasing the quality of the judgment sets. The objective has split up into two categories such as topics and participated systems.

5.2.1 Results and Discussion

5.2.1.1 Effect of Topic size to improve the quality of the pooled list

The experiment based on topic size has been done as a continuation of the proposed methodologies. The proposed methodology has been done with pooling and document similarity techniques using clustering and classification techniques. The topic-based experiment was done with pooling and cluster-based document similarity. The impact of topics in the proposed methodology has been evaluated here and how effectively these topics can perform on the proposed methodology have evaluated here. The overall aim is to achieve or maintain the quality of the relevant judgment sets with reduced topic size and at the same time choose the best topics with minimal topic hardness. For this experiment, the TREC-8 test collection was used with 129 systems and 50 topics and the

TREC-10 test collection has used with 97 systems with 50 topics. The MAP value of the TREC-8 test collection for each topic is shown in Figure 5.11. The x-axis shows the topic numbers and the y-axis shows the difference in the mean average precision from the median value for 50 topics from the TREC-8 test collection.

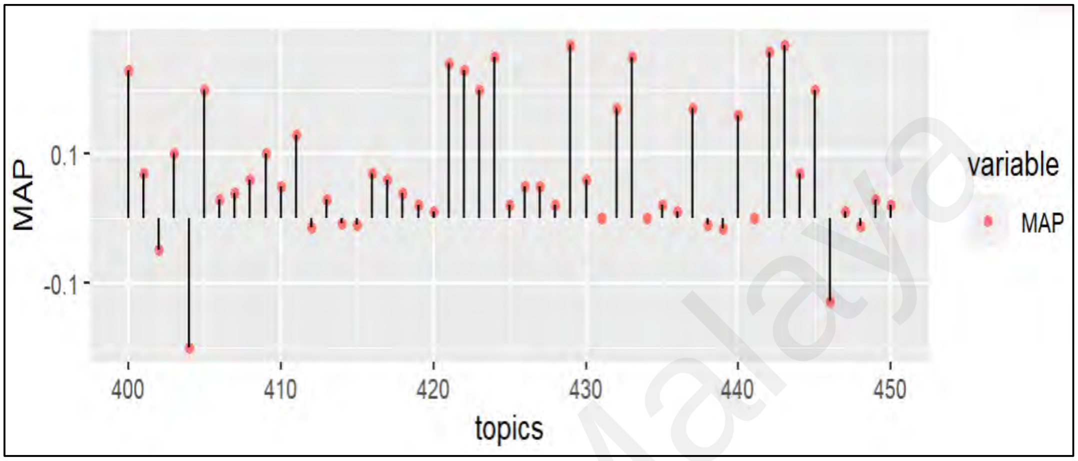


Figure 5.11: Every 50 topics mean average precision difference from the median from TREC-8

For this experiment, topics were categorized into easy and hard based on the average of the average precision score, AAP. If the value is less than 0.17, it considers are hard topic and greater than 0.17, considers as easy topics. Based on TREC-8 dataset, out of 129 systems, 124 systems were considered 70% of the topics were hard and 30% of topics were easy. Same way, based on the TREC-10 test collection, out of 97 systems, 77 systems were considered. Based on the methodology proposed in Section 4.1, actual MAP was calculated with all systems and topics and compared with easy topics and hard topics. Figure 5.12 shows the Kendall correlation of easy topics and hard topics MAP with the actual MAP value based on the TREC-8 test collection. Figure 5.13 shows the Kendall correlation with the TREC-10 test collection.

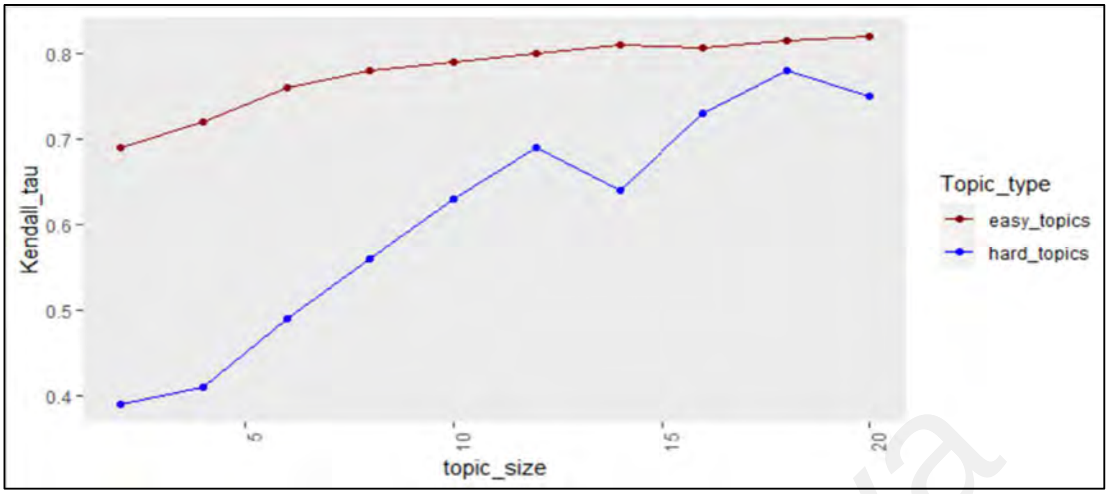
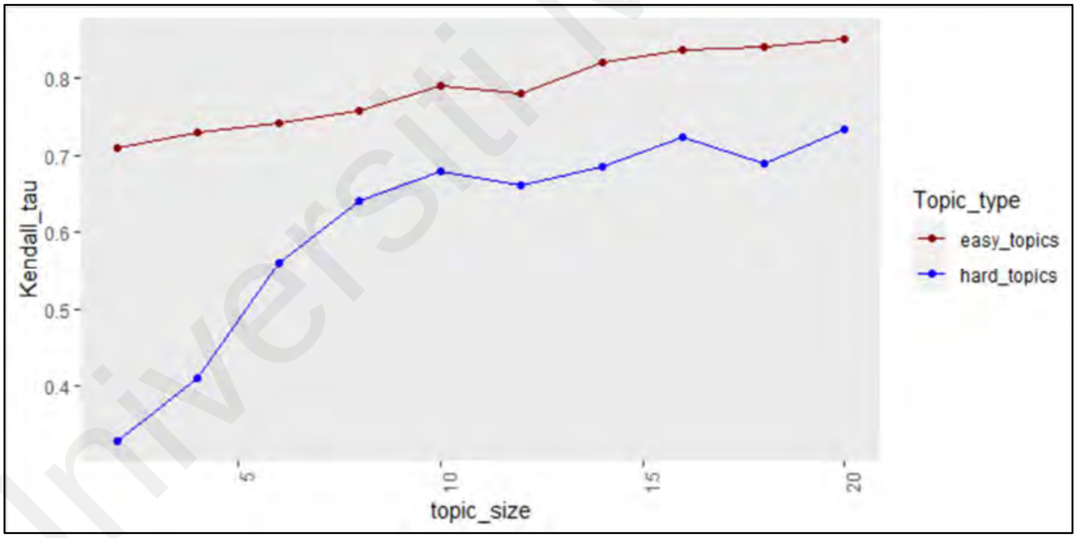


Figure 5.12: Kendall correlation of easy and hard topics MAP with actual MAP based on TREC-8 data collection

Figure 5.13: Kendall correlation of easy and hard topics MAP with actual MAP



based on TREC-10 data collection

Topics are split up into easy and hard topics based on average of average precision. Some random sets of easy topics and hard topics have been considered with different topic sizes. These data have been plotted on the x-axis and compared with the actual MAP value. Kendall's correlation of these values is shown in the y-axis. The results show that easy topics are highly correlated to actual MAP and it can be achieved even with a lesser number of topic sizes. Hard topics are less correlated to the actual MAP with the lesser number of topics. But it can be closer to the actual MAP if can consider a large set of topic sizes. However, larger sets of topics have an impact on the higher computational cost.

These results have proven that topics and topic sizes have a greater impact on the quality of the judgment sets. By only considering easy topics, the evaluation can be done and will be almost similar results to the actual whole topic results. With a smaller number of easy topics itself, can maintain the effectiveness measurement of the information retrieval systems. It is actually cost-effective with fewer resources.

5.2.1.2 Effect of Participated systems to improve the quality of the pooled list

The effectiveness of an information retrieval system can be measured using various evaluation measures such as average precision, mean average precision, f-measure, ndcg, and rbp. The variation in the results of their effectiveness score contributes to the reliability of the retrieval systems. Based on these scores can find out the good contributing systems and less contributing systems. The systems that assign better rank for the relevant documents by suppressing the irrelevant ones are called good contributing systems. The systems that assign a lower rank for the documents that are actually relevant or do not consider those documents in the judgment sets are called less contributing systems. The documents that are chosen from these types of less contributing systems or the irrelevant documents into the pooled list and later considered for evaluation purposes will affect the quality of the relevant judgment sets (Iwayama,2000) (Djenouri et al., 2018) (Rahman et al., 2020).

Based on the methodology proposed in Section 4.2 induced AP evaluation measures have been used instead of AP. If the number of documents in the judgment set reduces, the average precision value also reduces and due to that ranking of the systems also changes. This is because all the unjudged documents are considered as irrelevant in average precision. If the number of documents in the judgment set is reduced, the number of relevant documents retrieved before a relevant document, hence precision is also reduced for that relevant document. Through that average precision also reduced (Buckley and Voorhees,2004). It affects the ranking of the systems and indirectly affects the accuracy of the evaluation process. Figure 5.14 shows the mean average precision of 10% of sample randomly chosen systems performance with the actual mean average precision with correlation coefficient value of r=0.87 of the TREC-8 test collection. Actual MAP indicates the MAP value of all the topics of all the systems.

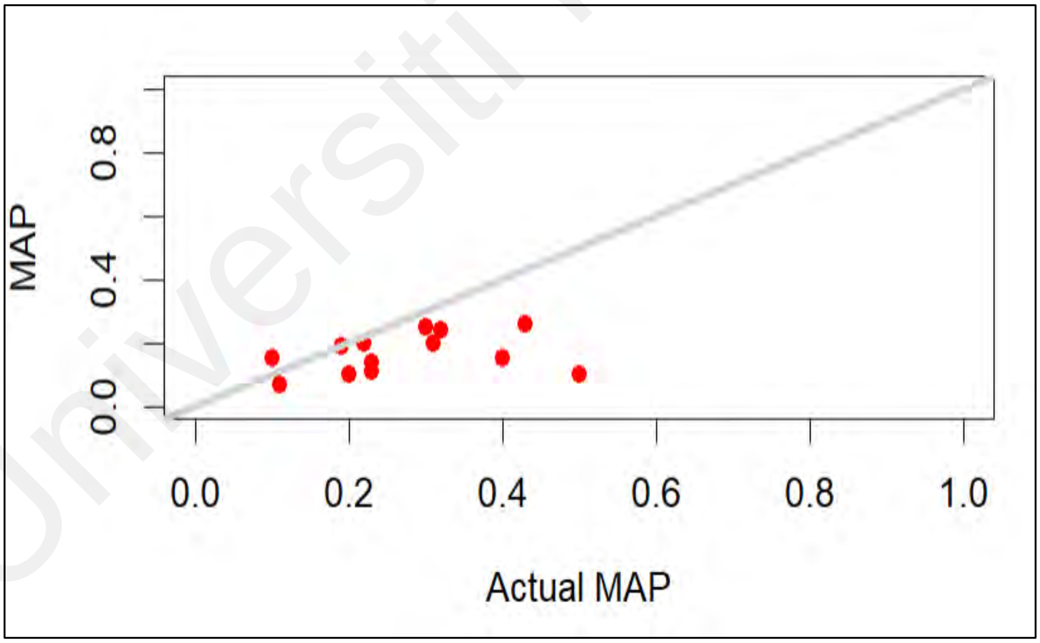


Figure 5.14: MAP vs Actual MAP for 10% of random retrieval systems from TREC-8

Another version of average precision called induced average precision (indAP) proposed by (Yilmaz and Aslam,2006) did not consider unjudged documents. Induced Average Precision calculated the relevant document score the same way as average precision with a change in that. First itself the unjudged documents were removed from the list and with the remaining documents only evaluation was done. Once the unjudged documents are removed, later the calculation is exactly the same way as average precision. The result of the performance of the same system which is shown in Figure 5.14 has been evaluated again with induced AP have shown in Figure 5.15. In this result, it has shown that the results are better compared to average precision. Based on this induced AP value, TREC-8 data collection systems were split up into good participating and less participating systems. This calculation is based on the number of relevant documents for each topic in each system.

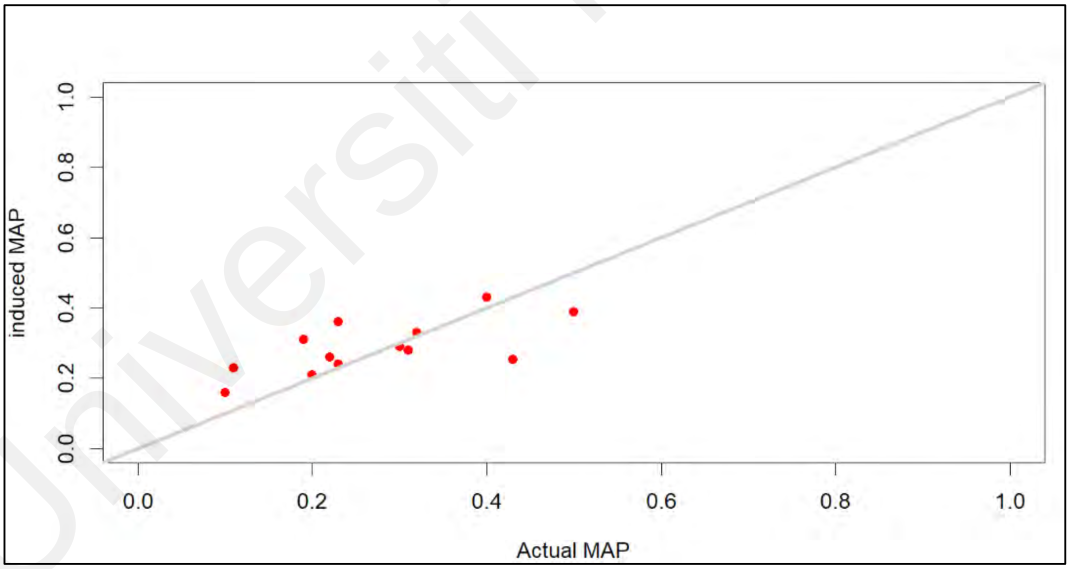


Figure 5.15: IndAP vs Actual MAP for 10% of random retrieval systems from TREC-8

Once the systems are split up into good and less contributed systems, documents from the good participated systems are considered in the pooled list. Top-k documents from the good participated systems are considered into the pooled list and documents from the less contributed systems are considered into the unjudged list. These documents were clustered and classified based on their similarity using the TF-IDF technique. And the methodology is continuous as proposed in Chapter 3. For TREC-8 data collection out of 129 systems, 124 systems were considered. Among those, 84 systems are considered as good contributing systems, and 40 systems are considered as less contributing systems. The same experiment was done with the TREC-10 test collection.

The experiment was done with this methodology and found out how many relevant documents were retrieved into the judgment set. Figure 5.16 shows the number of relevant documents retrieved based on the number of judgments. The results were compared with an earlier proposed methodology which was proposed in Section 5.1. Figure 5.17 shows the results of the proposed methodology based on the TREC-10 test collection.

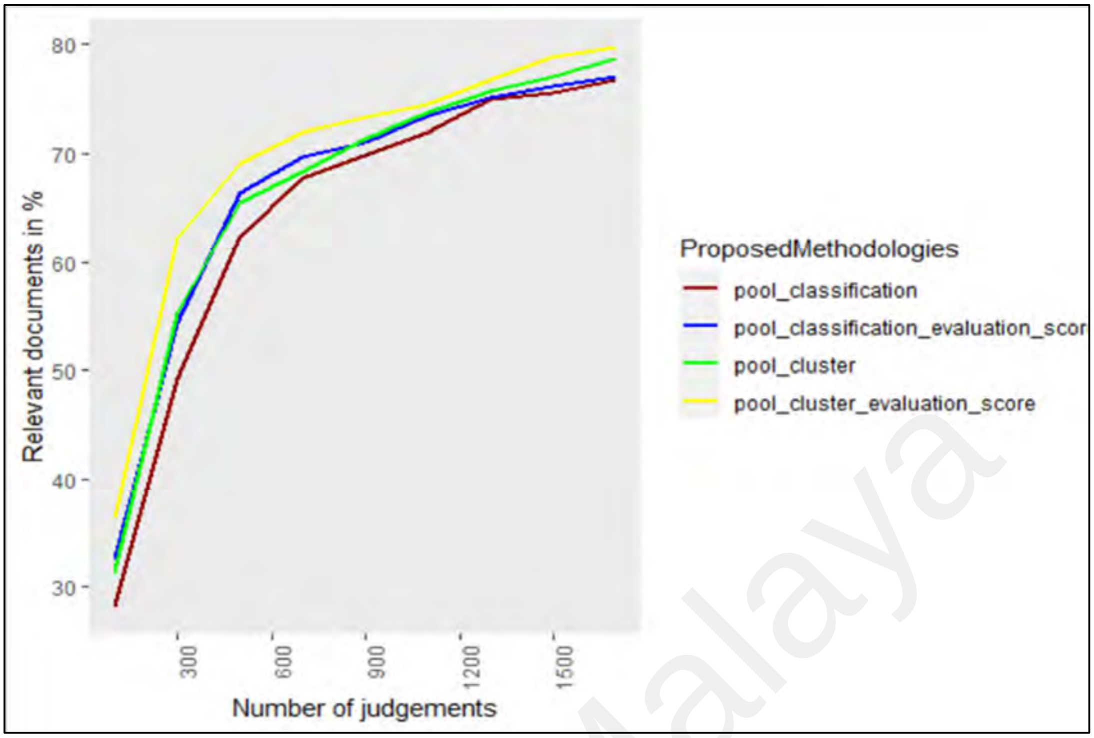


Figure 5.16: Relevant documents retrieved using proposed methodologies (in %) using TREC-8 test collection

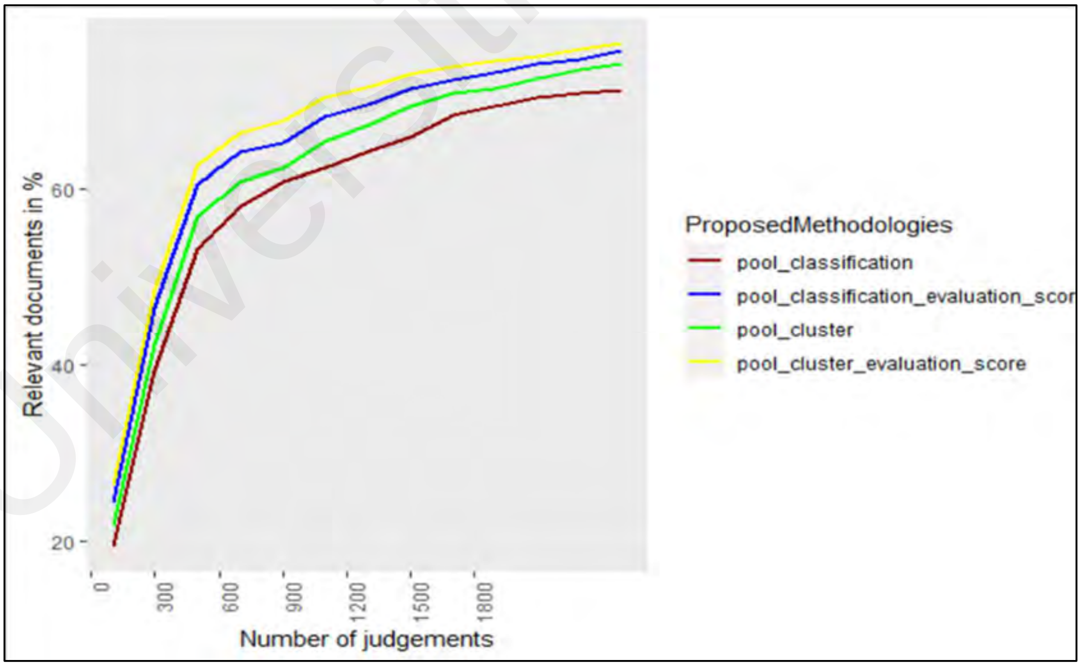


Figure 5.17: Relevant documents retrieved using proposed methodologies (in %) using TREC-10 test collection

Based on the results of TREC-8 and TREC-10 data collection from Figure 5.16 and Figure 5.17, it has proven that the proposed methodologies were able to retrieve a greater number of relevant documents into the judgment set compared to the proposed one in Section 5.1. Also, it has shown that with lesser pool depth itself, the systems were able to retrieve many relevant documents into the judgment set. It is cost-effective based on the evaluation process.

The performance of these methodologies based on the evaluation score in order to retrieve the number of relevant documents into the judgment sets is shown in Table 5.5. The comparison of these proposed methodologies with evaluation scores and the proposed ones in Section 5.1 has been done. The average precision of topics overall topics was calculated based on the pool depth of 100 and evaluation depth of 1000. Later, mean average precision is calculated based on the mean of the average precision of the topics over all the topics were done. The experiments have been done with both TREC-8 and TREC-10 test collections. The results were ordered according to the ascending order of their performance scores.

Table 5.5: Comparison of Mean average precision score of the enhanced proposed methodologies with previous ones

Methodology	TREC-8 (Adhoc)	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

5.2.2 Significance and Contributions

Test collections have a greater impact on increasing the quality of the judgment sets. Topics and participation systems' involvement in the enhancing quality of the judgment sets were evaluated here. For topics, a number of topics, topic size, and quality of the topics were considered. The main aim of the topics was to maintain the quality of the judgment sets with reduced topic size. For that topics were classified into easy and difficult topics. Based on the results it has proven that if considering only easy topics, can maintain the quality of the judgment sets. Even hard topics also can be considered for the judgment purpose, but the topic size need to increase. However with a smaller number of effective topics, easy topics can maintain the reliability of the effectiveness measurement of information retrieval systems.

By considering the participated systems' performance, the quality of the judgment sets can be increased. So the system performance was considered based on the evaluation measures and based on these scores systems were categorized into good contributing systems and less contributing systems. An enhancement of the proposed methodologies has been done based on these categorizations of the systems. The results show that a greater number of relevant documents were able to achieve the relevance judgment set based on the enhanced proposed methodologies. Also, it has shown that with lesser depth itself, systems were able to achieve more relevant documents. Also, it has shown that as the judgment document size increases, the MAP value gets closer to all the methodologies.

As per the evaluation of the quality of the judgment sets, it has shown that test collections have a greater impact on increasing the number of relevant documents in the judgment sets. Lesser topic sizes and enhanced proposed methodologies by considering systems performance with lesser depth help to increase the quality of the judgment sets and through that, it reduces the computational cost of the evaluation process.

5.3 Evaluating the effectiveness of the proposed evaluation methodology in terms of the incompleteness of the judgment sets and biasness in ranking

5.3.1 Results and Discussions

The pooling technique in the Cranfield paradigm has biasness in the judgment sets. Biasness is calculated based on how many irrelevant documents have moved into the judgment sets and ranked higher than the relevant documents. This creates the biasness or incompleteness in the judgment sets and systems are not able to evaluate correctly, which affects the performance of the systems. This research mainly focuses on the biasness in ranking of the documents in the judgment sets. The biasness in the ranking can be calculated with evaluation measures such as average precision, mean average precision, recall, etc. However, the error rate of different measures is marked differently (Voorhees, 2019).

Bpref measures have been introduced to evaluate the incompleteness in the judgment set (Buckley and Voorhees,2004). Later this bpref evaluated against MAP over judged documents and computed scores by removing unjudged documents from the ranking instead of assuming that they were not relevant. Later, the results have proven that MAP performs better compared to bpref in terms of defining different run sets and also the similarity of runs with Kendall’s correlation scores (Sakai and Kando, 2008). Other evaluation measures such as induced AP, subcollection AP, and inferred AP have been always proposed to overcome this incompleteness (Yilmaz and Aslam,2008), but it is highly correlated in terms of system rankings when the judgment set is complete. These measures can be used in the assumption that uniform random samples of relevant judgments are known, so these above AP values can perform better. But practically, relevance judgment cannot be with uniform random samples. So our experiments are continuing with bpref measures. Later, many studies have been done with the incompleteness and biasness in the judgment sets using different evaluation measures described in (Valcarce et al., 2020). Various ranking evaluation measures based on incompleteness in the judgment sets and estimation of these metrics with incompleteness of judgments have been done (Kirnap et al., 2021).

Based on the algorithm proposed in Section 4.3 and the proposed methodologies, ranking biasness in the judgment sets have reduced with the help of score adjustment. New scores have been given to similar documents in the unjudged document list based on the similarity from the pooled list and moved those documents into the pooled list. The aim of the research itself is to increase the number of relevant documents in the judgment sets and at the same time, assign better ranks for the documents that are relevant over the irrelevant ones.

For this experiment, the TREC-8 dataset was used with 528k documents (1.9GB) 50 topics (401-450), and 124 runs. For this experiment, various qrels sizes have been used to measure the effectiveness changes as the qrels value varies. 6 qrels set with various sizes have been used for the experiment. Based on Figure 5.17, the results show that as the the qrels size increases, MAP and P@10 value increase while at the same time, as the qrels size increases bpref value decreases. It happens due to the incompleteness in the relevance judgments as the document size increases. To maintain consistency, score adjustments have been made. Based on the score adjustment, the Mean absolute error rate has been calculated based on the Rbp@10 with p=0.8. Rbp assigns decaying weight to each rank and the adjusted score values with reduced error rates are shown in Table 5.6.

In this MAE value is calculated based on the difference in the true value (original value) and the adjusted score value. The results show that with the lesser pool depth the difference is somehow higher compared to the deeper pool depth. As the pool depth increases, the biasness difference is somehow lesser.

Table 5.6: Bias evaluation from systems

Pool depth	Raw value	Adjusted value
1	0.441	0.304
5	0.389	0.296
10	0.291	0.238
20	0.235	0.198
50	0.162	0.14

Based on the score adjustments, system incompleteness has been calculated with baseline work and the proposed methodology with score adjustments. Here for comparison, the cluster-based ICIR has been compared with the proposed methodology with pooling and clustering technique. The incompleteness is calculated with different qrels sets with different sizes. Kendall τ correlations between system ranking produced with 100% qrels and system ranking produced with different qrels are shown in Figure 5.17.

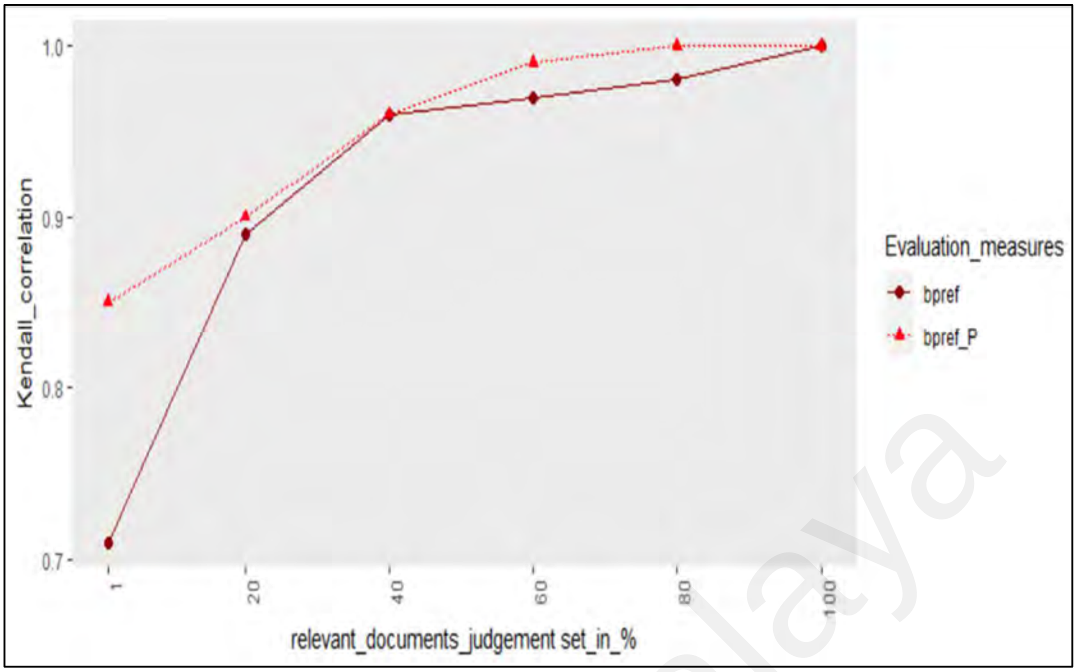


Figure 5.17: Kendall's correlation based on different judgments in the TREC-8 collection

The results show that the bpref measure based on the proposed methodology is flatter than the proposed one. So, it's proven that consistency has been achieved even with the increased qrels. Consistent scores are an important part of the incomplete collection. If bpref values are consistent at different sets of qrels with various sizes, shows that score and methodology are meaningful. The performance of the bpref compared to other measures is also evaluated and shown in Figure 5.18.

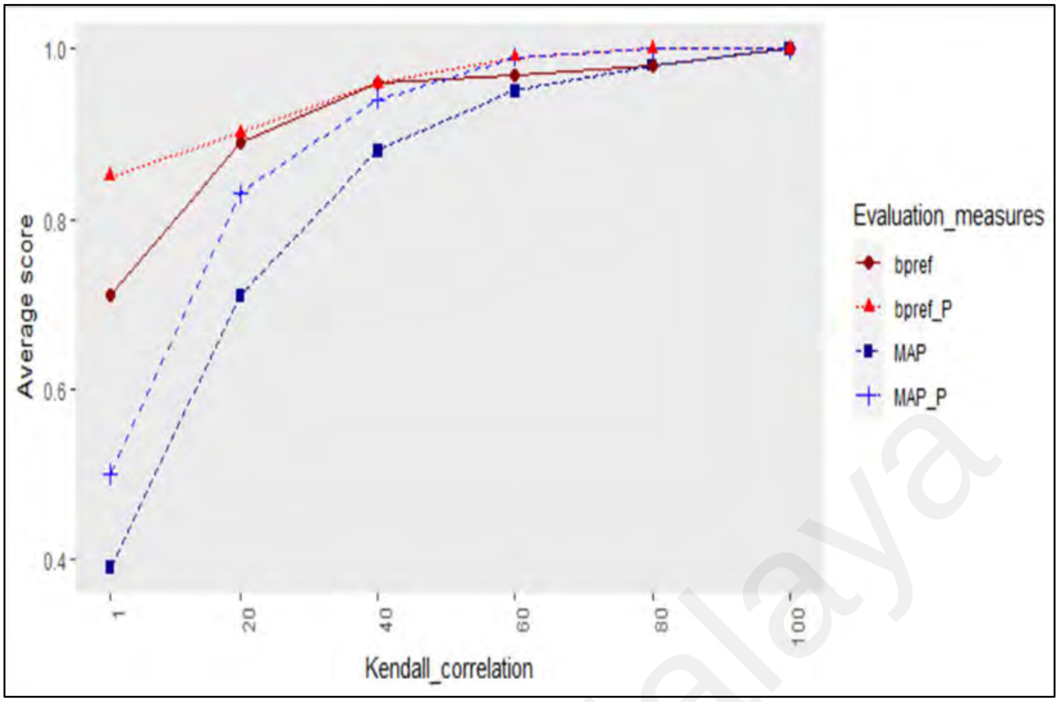


Figure 5.18: Changes in Kendall's correlation of measures based on different judgments in TREC-8 collection

The comparison of the baseline works (bpref and MAP) with proposed methodologies (bpref_P and MAP_P) shows that the bpref measure is more flatter than all the other measures. It shows that bpref measures continue to rank all the systems in the same preference order by using complete judgments for a higher level of incompleteness. This proves that proposed methodologies have shown better effectiveness results in the biasness of ranking of documents and also ranking in the systems.

5.3.2 Significance and Contributions

Biasness in ranking and the incompleteness of the judgments in relevance judgment sets is one of the main concerns of information retrieval evaluation researchers. The limitation in the consistency of the incompleteness in judgment sets as the number of documents in the judgment set was an issue. This limitation has been reduced somehow by the proposed methodologies. With the help of score adjustments and document selection for the relevance judgments, incompleteness has been reduced. Also, the proposed methodologies were able to adjust the order in which the documents are added to the judgment set. With the help of bpref measure, it has been proven that incompleteness has reduced, and it is almost flatter as the relevant judgment set size increases. Also, it continuously ranks all systems in the same preference order as when using the complete judgments for a higher level of incompleteness.

CHAPTER 6: CONCLUSION

The main aim of the research is to improve the quality of the judgment sets by increasing the number of relevant documents in the judgment sets and through that increase the quality of the judgment sets. If the quality of the judgment sets increases, the accuracy of the information retrieval evaluation process also increases. In this thesis, the evaluation process is done with system-based evaluation. The main issues considered in this thesis are partial relevance judgments and biasness in the ranking. The contributions of this thesis are, Firstly, proposing an experimental methodology to improve the accuracy of the information retrieval evaluation process by improving the number of relevant documents in the judgment sets. Secondly, evaluate how this proposed methodology can improve the quality of the judgment sets with the impact of the test collections by considering topics and participated systems cost-effectively. Thirdly, measured the effectiveness of the proposed methodology in terms of biasness in ranking. The results and discussions have continued in the following chapter and concluded it with its contributions. Finally concluded this thesis with some future works.

6.1 Thesis Contributions

6.1.1 Improve the accuracy of the Information Retrieval Evaluation process by increasing the number of relevant documents

One of the main aims of the information retrieval evaluation process is to improve the quality of the judgment sets. Pooling, topics consideration, human accessors consideration, and document similarity methodologies have been considered for this purpose. Pooling and document similarity based on the clustering and classification techniques have been considered for our evaluation process. Pooling is well well-known and popular traditional methodology. But considering document similarity techniques have some limitations in the quality of the judgment sets. Partial relevance judgment means not considering or retrieving all the relevant documents into the judgment sets is one of the main concerns of this research.

The first problem addressed in this experimentation is considering document similarity through a classifier and clustering globally can achieve the number of relevant documents, but the quality of the documents based on their relevancy is lesser compared to the traditional approaches.

The second problem addressed is the variation in the system rankings during the evaluation process when considering different evaluation depths and pool depths.

The third problem addressed is how the contribution of this thesis works helps to retrieve more relevant documents and increase the judgment sets compared to the baseline works.

The objectives of this experimentation address the above problems.

The first objective is to propose an experimental methodology to improve the accuracy of the evaluation process by increasing the number of relevant documents in the judgment sets.

The second objective is to measure the performance of the system using various evaluation metrics by assigning various evaluation depths and pool depths.

The third objective is to measure the performance of the proposed methodology by comparing it with the baseline works.

The fourth objective was to explore the effectiveness of the proposed methodologies using different clustering techniques.

Summary

The results show that the proposed methodologies based on pooling and document similarity with clustering and classification techniques help to achieve a greater number of relevant documents in the judgment list and through that it helped to increase the quality of the judgment sets. Also, compared to the baseline works, the proposed methodologies were able to retrieve a greater number of relevant documents with lesser pool depth. It helps indirectly to improve the accuracy of the evaluation process by improving the system effectiveness score.

Also, through the evaluation measures, it has been noticed that the systems perform better based on different pool depths and evaluation depths. The system performs better when the pool depth is greater than the evaluation depth and also when the pool depth is equal to the evaluation depth. Also, when the number of relevant documents in the qrels is greater than the pool depth and when the evaluation depth becomes greater than the pool depth, the performance of the system varies significantly. Also, it has been noticed that hierarchical clustering techniques perform better compared to k-means clustering for this methodology.

6.1.2 Increase the quality of relevance judgments by considering topics and participating systems in the test collections

Test collections have a greater impact on the quality of the judgment sets. This research has considered topics and participated systems. The main aim here is to consider how effectively can use the topics and participated systems in a cost-effective way and at the same time maintain the quality of the judgment sets. Here cost effectiveness is considered based on the topic sizes and participated systems efficiency. Less topic sizes and consideration of good participated systems reduced the computational cost.

The first problem considered here is to reduce the computational cost of the proposed methodology by considering topics in an effective way and at the same time maintain the quality of the judgment sets.

The second problem considered here is to reduce the computational cost of the proposed methodology by considering participated systems in an effective way and at the same time maintaining the quality of the judgment sets.

Thirdly, how these proposed topic sizes and contributed systems perform effectively compared to the baseline methodologies

The objectives of this experimentation address the above problems.

The first objective is to reduce the computational cost of the proposed methodologies by considering reduced topic sizes and at the same time maintain the quality of the judgment sets.

The second objective is to reduce the computational cost of the proposed methodologies by considering good contributing systems documents efficiently and increasing the quality of the judgment sets with reduced pool depth.

The third objective is to measure the effectiveness of the proposed methodologies based on the topics and participating systems.

Summary

The results show that considering only easy topics can maintain the quality of the judgment sets. Even hard topics can also be considered for judgment purposes, but the topic size needs to increase. However, with a smaller number of effective topics, easy topics can maintain the reliability of the effectiveness measurement of information retrieval systems. With the enhancement of the proposed methodologies by considering good participated systems documents, the results show that a greater number of relevant documents were able to achieve the relevance judgment set. Also, it has shown that with lesser depth itself, systems were able to achieve more relevant documents. Also, it has shown that as the judgment document size increases, the MAP value gets closer to all the methodologies. These all results show that computational cost gets lesser with reduced topic size and good contributed systems documents.

6.1.3 To measure the effectiveness of the proposed evaluation methodology in terms of incompleteness of the judgment sets and biasness in ranking

The Cranfield paradigm assumes that all the documents that are retrieved are completely relevant and judgment sets are complete. For smaller datasets, this concept might be true. But for larger datasets like TREC, this will not always. Many irrelevant documents might be retrieved by the systems into the relevant judgment sets and assign higher ranks to these documents over relevant documents. This affects the quality of the judgment sets and through that affects the accuracy of the evaluation process.

The first problem is the incompleteness or biased judgments. This means the judgments that are biased against the systems or systems that are not contributing enough relevant documents to the pooled document list.

The second problem is the systems are assigning higher ranks to the documents that are irrelevant than the relevant documents.

The third problem is whenever the relevance judgment set size increases, the incompleteness or the biasness also increases. The consistency is not maintained well.

The objectives of this experimentation address the above problems.

The first objective is to measure the effectiveness of the proposed methodologies in the matter of biasness in relevance judgment sets by considering score adjustments to the documents which are not considered in the pooled list.

The second objective is to measure the overall effectiveness of the proposed methodologies in order to improve the accuracy of the evaluation process.

Summary

The results show that with the help of score adjustments and document selection for the relevance judgments, incompleteness has been reduced. Also, the proposed methodologies were able to adjust the order in which the documents are added to the judgment set. With the help of bpref measure, it has been proven that incompleteness has reduced, and it is almost flatter as the relevant judgment set size increases. Also, it continuously ranks all systems in the same preference order as when using the complete judgments for a higher level of incompleteness.

6.2 Limitations and Future Works

This section highlights some of the limitations of the existing methodologies and interesting studies that can be extended from the work that is mentioned in this thesis.

The main aim of this thesis is to improve the quality of the relevance judgment sets. For that, this thesis mainly focused only on increasing the number of relevant documents in the judgment sets. Here pooling and document similarity methodologies were used.

Pooling of the documents and later the document similarity of clustered or classified unjudged document list and the pooled document list have done and if found a similarity, then these documents were moved to the judgment sets by assigning a new score. This methodology helped to increase the number of relevant documents in the judgment sets and through that increased the quality of the judgment sets. But the execution time or the computational time is higher. This happens mainly because of the large test collection considered for the evaluation process and also, similarity checking was done with the whole test collection. As an enhancement to reduce the execution time, some similarity measures based on clustering and classification need to be considered in the vector space-based model, where distance metrics can be applied to choose nearby similar documents (Eminagaoglu,2022) and word embedding. The word embedding can be created based on a document and helps to identify the most relevant document based on a topic (Brundha and Meera,2022). This might help to identify the most other relevant documents within a vector distance in less time and reduce the computational time in considering all the documents.

The topic size reduction and consideration of good contributing systems documents help to increase the quality of the judgment sets. Also, it helped to reduce the computational cost. But still, for large document collections like TREC, pre-trained methods (Fan et al., 2022) could evaluate the systems at document level scores and cut-off ranks. Also, the role of fair ranking of systems based on relevancy (Balagopalan et al., 2023) might help to reduce the biasness in the judgment sets.

Finally, Information retrieval is still an ongoing process which helps to retrieve the relevant data based on the user's query on the World Wide Web. It is achieved mainly through retrieval systems performance on query formation, accessing real-time document corpus, and performance on retrieval algorithms. Search engines or retrieval systems have a major role in satisfying real user’s requests on the Web. For that web search engines must perform smarter with the retrieval algorithms on the fast-growing Web. The proposed methodologies in this thesis have helped to measure the quality of the retrieval effectiveness of the systems and have also proven that it has improved the results. Improvision of these methodologies in the system-oriented evaluation is an important aspect of the advanced user's queries and enhancing Web in the current era.

LIST OF PUBLICATIONS

- Joseph, M. H., & Ravana, S. D. (2022, December). Generation of High-Quality Relevant Judgments through Document Similarity and Document Pooling for the Evaluation of Information Retrieval Systems. In *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (pp. 261-265). IEEE.

- Joseph, M. H., & Ravana, S. D. (2024). Reliable Information Retrieval Systems Performance Evaluation: A Review. *IEEE Access*. Improve the Accuracy of the Information Retrieval Evaluation process by considering unjudged document lists from the relevant judgment sets

- Improve the Accuracy of Information Retrieval Evaluation process by considering unjudged document list from the relevance judgment sets

  Journal: Information Research-Q4

  Status: (Accepted)

REFERENCES

A.Lipani, D. E. Losada, G. Zuccon and M. Lupu, "Fixed-Cost Pooling Strategies," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1503-1522, 1 April 2021

A.Moffat, W. Webber, J. Zobel, Strategic system comparisons via targeted relevance judgments, in: Proc. 30th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, NY, USA, 2007, pp. 375–382.

A. Moffat, J. Zobel, Rank-biased precision for measurement of retrieval effectiveness, ACM Trans. Inf. Syst. 27 (1) (2008) 2:1–2:27, doi: 10.1145/1416950. 1416952 .

Aliwy, A. H., Aljanabi, K., & Alameen, H. A. (2022, January). Arabic text clustering technique to improve information retrieval. In *AIP Conference Proceedings* (Vol. 2386, No. 1). AIP Publishing.

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Information processing & management*, *48*(6), 1053-1066

Andrade, C. (2019). The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. *Indian journal of psychological medicine*, *41*(3), 210-215.

Arabzadeh, N., Vtyurina, A., Yan, X., & Clarke, C. L. (2021). Shallow pooling for sparse labels. *arXiv preprint arXiv:2109.00062*

Arora, M., Kanjilal, U., & Varshney, D. (2016). Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, *12*(2), 224-236.

Aslam, J. A., Yilmaz, E., & Pavlu, V. (2005, August). The maximum entropy method for analysing retrieval measures. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 27-34).

Ateyah, S., & Al-Augby, S. (2023, March). Proposed information retrieval systems using LDA topic modelling for answer finding of COVID 19 pandemic: A brief survey of approaches and techniques. In *AIP Conference Proceedings* (Vol. 2591, No. 1). AIP Publishing.

Aydın, A., Arslan, A., & Dinçer, B. T. (2024). A set of novel HTML document quality features for Web information retrieval: Including applications to learning to rank for information retrieval. *Expert Systems with Applications*, *246*, 123177.

B. Carterette, Robust test collections for retrieval evaluation, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '07, ACM, New York, NY, USA, 2007, pp. 55–62 .

B. T. Bartell, G. W. Cottrell, and R. K. Belew. Automatic combination of multiple ranked retrieval systems. In Proceedings ACM-SIGIR'94, pages 173–181. Springer-Verlag, 1994.

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

Basu, D. (2011). Randomization analysis of experimental data: the Fisher randomization test. *Selected Works of Debabrata Basu*, 305-325.

Bellogín, A., Castells, P., & Cantador, I. (2017). Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, *20*, 606-634.

Berto, A., Mizzaro, S., & Robertson, S. (2013, September). On using fewer topics in information retrieval evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval* (pp. 30-37).

Bellot, P., Doucet, A., Geva, S., Gurajada, S., Kamps, J., Kazai, G., ... & Wang, Q. (2013, September). Overview of INEX 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 269-281). Berlin, Heidelberg: Springer Berlin Heidelberg.

Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). Statistics for experimenters. In *Wiley series in probability and statistics*. Hoboken, NJ: Wiley.

Braschler, M. (2000, September). CLEF 2000—overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 89-101). Berlin, Heidelberg: Springer Berlin Heidelberg.

Brundha, J., & Meera, K. N. (2022, April). Vector Model Based Information Retrieval System with Word Embedding Transformation. In *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)* (pp. 01-04). IEEE.

Buckley, C., & Voorhees, E. M. (2004, July). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 25-32).

Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2006, August). Bias and the limits of pooling. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 619-620).

Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2007). Bias and the limits of pooling for large collections. *Information retrieval*, *10*, 491-508.

Buckley, C., & Voorhees, E. M. (2017, August). Evaluating evaluation measure stability. In *ACM SIGIR Forum* (Vol. 51, No. 2, pp. 235-242). New York, NY, USA: ACM

Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I. (2007, July). Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th* *annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 63-70).

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, *44*(1), 1-50.

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & Allan, J. (2008, July). 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).

Carterette, B., Pavlu, V., Fang, H., & Kanoulas, E. (2009a). Million query track 2009 overview. In Proceedings of TREC.

Carterette, B., Kanoulas, E., Pavlu, V., & Fang, H. (2010, July). Reusable test collections through experimental design. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 547-554).

Carterette, B. A. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 1-34.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008, July). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 659-666).

Clarke, C. L., Vtyurina, A., & Smucker, M. D. (2021). Assessing Top-Preferences. *ACM Transactions on Information Systems (TOIS)*, *39*(3), 1-21.

Clarke, C. L., Diaz, F., & Arabzadeh, N. (2023, February). Preference-Based Offline Evaluation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (pp. 1248-1251).

Cleverdon, C. (1967, June). The Cranfield tests on index language devices. In *Aslib proceedings* (Vol. 19, No. 6, pp. 173-194). MCB UP Ltd.

Cormack, G. V., Palmer, C. R., & Clarke, C. L. (1998, August). 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).

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

Crane, M., Culpepper, J. S., Lin, J., Mackenzie, J., & Trotman, A. (2017, February). A comparison of document-at-a-time and score-at-a-time query evaluation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 201-210).

Culpepper, J. S., Faggioli, G., Ferro, N., & Kurland, O. (2021). Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems (TOIS)*, *40*(1), 1-36.

Dalianis, H., & Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical Text Mining: secondary use of electronic patient records*, 45-53.

De Winter, J. C. (2019). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, *18*(1), 10.

Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2007). CLEF 2006: Ad hoc track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers 7* (pp. 21-34). Springer Berlin Heidelberg.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020, October). Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 275-284).

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

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.

Djoerd Hiemstra, "Information Retrieval Models", published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, November 2009, Ltd., ISBN-13: 978-0470027622

E. M. Voorhees and C. Buckley, "The effect of topic set size on retrieval experimental error," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'02, 2002, pp. 316–323.

Eguchi, K., Kuriyama, K., & Kando, N. (2002, May). Sensitivity of IR systems Evaluation to Topic Difficulty. In *LREC*.

Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Evaluation of Cross-Language Information Retrieval Systems. Proceedings of CLEF 2001, number 2406 in Lecture Notes in Computer Science, pages 355–370, 2002

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.

Fan, Y., Xie, X., Cai, Y., Chen, J., Ma, X., Li, X., ... & Guo, J. (2022). Pre-training methods in information retrieval. *Foundations and Trends® in Information Retrieval*, *16*(3), 178-317.

Ferrante, M., Ferro, N., & Fuhr, N. (2021). Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. *IEEE Access*, *9*, 136182-136216.

Ferro, N. (2017). Reproducibility challenges in information retrieval evaluation. *Journal of Data and Information Quality (JDIQ)*, *8*(2), 1-4.

Ferro, N., Kim, Y., & Sanderson, M. (2019). Using collection shards to study retrieval performance effect sizes. *ACM Transactions on Information Systems (TOIS)*, *37*(3), 1-40

Gienapp, L., Stein, B., Hagen, M., & Potthast, M. (2020, October). Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 2033-2036).

Guiver, J., Mizzaro, S., & Robertson, S. (2009). A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, *27*(4), 1-26.

Frei, H. P., & Schäuble, P. (1991). Determining the effectiveness of retrieval algorithms. *Information Processing & Management*, *27*(2-3), 153-164.

Harman, D. (1993, July). Overview of the first TREC conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 36-47).

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, *31*(3), 271-289.

Harman, D. K. (1995). *Overview of the third text retrieval conference (TREC-3)* (No. 500). DIANE Publishing.

Harman, D. K. (1996). Overview of the fourth text retrieval conference (TREC-4).

Harman, D. (2011). *Information retrieval evaluation*. Morgan & Claypool Publishers.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

Hawking, D., Voorhees, E., Craswell, N., & Bailey, P. (1999, November). Overview of the trec-8 web track. In *TREC*.

Hawking, D. (2000, November). Overview of the TREC-9 Web Track. In *Trec*.

Hawking, D., & Craswell, N. (2002). Overview of the TREC-2001 web track. *Nist Special Publication Sp*, (250), 61-67.

Hofstetter, S., Lin, S. C., Yang, J. H., Lin, J., & Hanbury, A. (2021, July). Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 113-122).

Hu, X., Bandhakavi, S., & Zhai, C. (2003, July). Error analysis of difficult TREC topics. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 407-408).

Hui, K., Berberich, K., & Mele, I. (2017, October). Dealing with Incomplete Judgments in Cascade Measures. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 83-90).

Hull, D. (1993, July). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 329-338).

Iwayama, M. (2000, July). Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 10-16).

J. Aslam, M. Montague , Models for metasearch, in: Proc. of the 24th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, NY, USA, in: SIGIR '01, 2001, pp. 276–284 .

J. Aslam, V. Pavlu , E. Yilmaz , A statistical method for system evaluation using incomplete judgments, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR '06, ACM, New York, NY, USA, 2006, pp. 541–548 .

J.A . Shaw , E.A . Fox , Combination of multiple searches, in: The Second Text Retrieval Conference (TREC-2), 1994, pp. 243–252 .

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422-446.

Jayasinghe, G. K., Webber, W., Sanderson, M., & Culpepper, J. S. (2014, November). Improving test collection pools with machine learning. In *Proceedings of the 19th Australasian Document Computing Symposium* (pp. 2-9).

Jones, K. S., & Willett, P. (Eds.). (1997). *Readings in information retrieval*. Morgan Kaufmann.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.

Kando, N. (2004). Evaluation of information access technologies at the NTCIR workshop. In *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers 4* (pp. 29-43). Springer Berlin Heidelberg.

Kando, N. (2007, May). Overview of the Sixth NTCIR Workshop. In *NTCIR*.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information processing & management*, *41*(5), 1019-1033.

Kırnap, Ö., Diaz, F., Biega, A., Ekstrand, M., Carterette, B., & Yilmaz, E. (2021, April). Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference 2021* (pp. 1065-1075).

Kishida, K. (2005). *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments* (p. 19p). Tokyo, Japan: National Institute of Informatics.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the web. *ACM computing surveys (CSUR)*, *32*(2), 144-173.

Korencic, D., Ristov, S., Repar, J., & Snajder, J. (2021). A topic coverage approach to evaluation of topic models. *IEEE Access*, *9*, 123280-123312

Kutlu, M., Elsayed, T., & Lease, M. (2018). Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging. *Information Processing & Management*, *54*(1), 37-59.

Kwok, K. L. (2005, August). An attempt to identify weakest and strongest queries. In *ACM SIGIR Workshop on Predicting Query Difficulty*.

Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994

Li, D., & Kanoulas, E. (2017, November). Active sampling for large-scale information retrieval evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp.49-58)

Li, H. (2022). *Learning to rank for information retrieval and natural language processing*. Springer Nature.

Liang, S., Markov, I., Ren, Z., & de Rijke, M. (2018, April). Manifold learning for rank aggregation. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1735-1744).

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461

Lin, J., Ma, X., Lin, S. C., Yang, J. H., Pradeep, R., & Nogueira, R. (2021, July). Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2356-2362).

Lipani, A., Lupu, M., & Hanbury, A. (2015, August). Splitting water: Precision and anti-precision to reduce pool bias. In *proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 103-112).

Losada, D. E., Parapar, J., & Barreiro, Á. (2016, April). Feeling lucky? Multi-armed bandits for ordering judgements in pooling-based evaluation. In *proceedings of the 31st annual ACM symposium on applied computing* (pp. 1027-1034).

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.

Losada, D. E., Parapar, J., & Barreiro, A. (2018, June). Cost-effective construction of Information retrieval test collections. In *Proceedings of the 5th Spanish Conference on Information Retrieval* (pp. 1-2).

Losada, D. E., Parapar, J., & Barreiro, A. (2019). When to stop making relevance judgments? A study of stopping methods for building information retrieval test collections. *Journal of the Association for Information Science and Technology*, *70*(1), 49-60.

Lu, X., Moffat, A., & Culpepper, J. S. (2016). The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal*, *19*(4), 416-445.

M. Crane, A. Trotman, and R. O'Keefe. 2013. Maintaining discriminatory power in quantized indexes. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13). 1221–1224.

M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval 4, 4 (2010), 247–375.

Mackenzie, J., Trotman, A., & Lin, J. (2023). Efficient document-at-a-time and score-at-a-time query evaluation for learned sparse representations. *ACM Transactions on Information Systems*, *41*(4), 1-28.

Maddalena, E., Roitero, K., Demartini, G., & Mizzaro, S. (2017, October). Considering assessor agreement in IR evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 75-82).

Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2017). On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, *35*(3), 1-32.

Magdy, W., & Jones, G. J. (2010). Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgements. In *Multilingual and Multimodal Information Access Evaluation: International Conference of the Cross-Language Evaluation Forum, CLEF 2010, Padua, Italy, September 20-23, 2010. Proceedings 1* (pp. 82-93). Springer Berlin Heidelberg.

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

Mandl, T. (2009). Easy tasks dominate information retrieval evaluation results. *Datenbanksysteme in Business, Technologie und Web (BTW)–13. Fachtagung des GI-Fachbereichs" Datenbanken und Informationssysteme"(DBIS)*.

Majumder, P., Mitra, M., Pal, D., Bandyopadhyay, A., Maiti, S., Mitra, S., ... & Pal, S. (2008, July). Text collections for FIRE. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 699-700).

Markovskiy, E., Raiber, F., Sabach, S., & Kurland, O. (2022, July). 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).

Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web*, *11*(2), 255-335.

Melucci, M., & Baeza-Yates, R. (2011). Chapter 4. The User in Interactive Information Retrieval Evaluation. Advanced topics in information retrieval. Berlin: Springer.

Mhawi, D. N., Oleiwi, H. W., Saeed, N. H., & Al-Taie, H. L. (2022). An efficient information retrieval system using evolutionary algorithms. *network*, *2*(4), 583-605.

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, *22*(4), 407.

Mizzaro, S., & Robertson, S. (2007, July). Hits hits TREC: exploring IR evaluation results with network analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 479-486).

Mizzaro, S. (2008, March). The good, the bad, the difficult, and the easy: something wrong with information retrieval evaluation?. In *European Conference on Information Retrieval* (pp. 642-646). Springer, Berlin, Heidelberg.

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.

Moffat, A., Webber, W., & Zobel, J. (2007, July). Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 375-382)

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, *27*(1), 1-27.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86-97.

Nguyen, T., MacAvaney, S., & Yates, A. (2023, March). A Unified Framework for Learned Sparse Retrieval. In *European Conference on Information Retrieval* (pp. 101-116). Cham: Springer Nature Switzerland.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modeling for qualitative studies. *Journal of Information Science*, *43*(1), 88-102.

Nowrozi, Y., Maleki, M., & Zarei, E. (2022). Information Retrieval in the Web Environment (Case Study: Iranian Library Software). *Knowledge Retrieval and Semantic Systems*, *9*(30), 67-92.

Otero, D., Parapar, J., & Barreiro, Á. (2023). Relevance feedback for building pooled test collections. *Journal of Information Science*, 01655515231171085.

Pang, W. T., Rajagopal, P., Wang, M., Zhang, S., & Ravana, S. D. (2019, December). Exploring Topic Difficulty in Information Retrieval Systems Evaluation. In *Journal of Physics: Conference Series* (Vol. 1339, No. 1, p. 012019). IOP Publishing.

Parapar, J., Losada, D. E., Presedo-Quindimil, M. A., & Barreiro, A. (2020). Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, *71*(1), 98-113.

Park, L. A., & Zhang, Y. (2007). On the distribution of user persistence for rank-biased precision. In *Proceedings of the 12th Australasian document computing symposium* (pp. 17-24). New York, NY: ACM.

Pavlu, V., Rajput, S., Golbus, P. B., & Aslam, J. A. (2012). Ir system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12* (pp. 393–402). New York, NY: ACM.

R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", *IJCER*, sep 2012, Vol. 2 Issue. 5 , PP: 1443-1444

Raghavan V, Bollmann P, Jung GS (1989) A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans Info Syst* 7(3): 205–229

Rahman, M. M., Kutlu, M., Elsayed, T., & Lease, M. (2020, September). Efficient test collection construction via active learning. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (pp. 177-184)

Rajagopal, P., Ravana, S. D., & Ismail, M. A. (2014). Relevance Judgments Exclusive of Human Assessors in Large Scale Information Retrieval Evaluation Experimentation. *Malaysian Journal of Computer Science*, *27*(2), 80–94

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).

Rajagopal, P., & Ravana, S. D. (2019). Effort-based information retrieval evaluation with varied evaluation depth and topic sizes. In H. Ketamo, P. Nueno, & U. Kumar (Eds.), *ICBIM 2019 - 2019 The 3rd International Conference on Business and Information Management* (pp. 143-147). Association for Computing Machinery (ACM).

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 (IJIRR)*, *12*(1), 1-18.

Rasmussen, E. (2003). Evaluation in information retrieval. *The MIR/MDL evaluation project white paper collection edition*, *3*, 45-49.

Rashidi, L., Zobel, J., & Moffat, A. (2023). The Impact of Judgment Variability on the Consistency of Offline Effectiveness Measures. *ACM Transactions on Information Systems*, *42*(1), 1-31.

Ravana, S. D., Rajagopal, P., & Balakrishnan, V. (2015). Ranking retrieval systems using pseudo relevance judgments. *Aslib Journal of Information Management*.

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010, July). Extending average precision to graded relevance judgments. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 603-610).

Roshdi, A., & Roohparvar, A. (2015). Information retrieval techniques and applications. *International Journal of Computer Networks and Communications Security*, *3*(9), 373-377.

Roitero, K., Maddalena, E., & Mizzaro, S. (2017). Do easy topics predict effectiveness better than difficult topics?. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39* (pp. 605-611). Springer International Publishing.

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.

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).

Roitero, K., Barbera, D. L., Soprano, M., Demartini, G., Mizzaro, S., & Sakai, T. (2023). How many crowd workers do i need? on statistical power when crowdsourcing relevance judgments. *ACM Transactions on Information Systems*.

Rozin, B., Pereira-Ferrero, V. H., Lopes, L. T., & Pedronette, D. C. G. (2021). A rank-based framework through manifold learning for improved clustering tasks. *Information Sciences*, *580*, 202-220.

S. Liang and M. de Rijke. Burst-aware data fusion for microblog search. *Information Processing & Management*, pages 89–113, 2015.

S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer, "TREC: Topic Engineering Exercise," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2014, pp. 1147–1150.

Sanderson, M., & Zobel, J. (2005, August). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 162-169).

Sanderson, M. (2010). Test collection-based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247-375.

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information processing & management*, *43*(2), 531-548.

Sakai, T. (2007, July). Alternatives to bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 71-78).

Sakai, T., & Kando, N. (2007). A Further Note on Alternatives to Bpref. *ディジタル図書館*, (33), 52-59.

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, *11*, 447-470.

Sakai, T., Tao, S., Chen, N., Li, Y., Maistro, M., Chu, Z., & Ferro, N. (2023). On the Ordering of Pooled Web Pages, Gold Assessments, and Bronze Assessments. *ACM Transactions on Information Systems*.

Sasirekha, K., & Baby, P. (2013). Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, *83*(3), 83.

Schnabel, T., Swaminathan, A., Frazier, P. I., & Joachims, T. (2016, September). Unbiased comparative evaluation of ranking functions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (pp. 109-118).

Smucker, M. D., Allan, J., & Carterette, B. (2007, November). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 623-632).

Smucker, M. D., & Jethani, C. P. (2012, August). Time to judge relevance as an indicator of assessor error. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1153-1154).

Sakai, T. (2007a). Alternatives to Bpref. *In ACM SIGIR 2007 Proceedings* (pp. 71–78).

Sakai, T. (2016). Topic set size design. *Information Retrieval Journal*, *19*, 256-283.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender systems handbook*, 257-297.

Shraga, R., Roitman, H., Feigenblat, G., & Canim, M. (2020, April). Ad hoc table retrieval using intrinsic and extrinsic similarities. In *Proceedings of The Web Conference 2020* (pp. 2479-2485).

Sormunen, E. (2002). Liberal relevance criteria of TREC—Counting on negligible documents? In M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 324–330). ACM, New York.

Sparck Jones, K., & Van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of documentation*, *32*(1), 59-75.

K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an “ideal” information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

T. Sakai. 2016. Evaluating evaluation metrics based on the bootstrap. In SIGIR 2006, pages 525–532.

T. Sakai. 2016. A Simple and Effective Approach to Score Standardization. In ACM ICTIR. 95–104.

Taha, K. (2023). Semi-supervised and un-supervised clustering: A review and experimental evaluation. *Information Systems*, 102178.

Tonon, A., Demartini, G., & Cudré-Mauroux, P. (2015). Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval Journal*, *18*(5),

Urbano, J., Lima, H., & Hanjalic, A. (2019, July). A new perspective on score standardization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1061-1064).

Valcarce, D., Bellogín, A., Parapar, J., & Castells, P. (2018, September). On the robustness and discriminative power of information retrieval metrics for top-N recommendation. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 260-268).

Valcarce, D., Bellogín, A., Parapar, J., & Castells, P. (2020). Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal*, *23*, 411-448.

Voorhees, E. M., & Harman, D. (2000). Overview of the sixth text retrieval conference (TREC-6). *Information Processing & Management*, *36*(1), 3-35.

Voorhees, E. M. (2000, September). Report on trec-9. In *ACM SIGIR Forum* (Vol. 34, No. 2, pp. 1-8). New York, NY, USA: ACM.

Voorhees, E. M. (2001, September). The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages* (pp. 355-370). Berlin, Heidelberg: Springer Berlin Heidelberg.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 63). Cambridge: MIT press.

Voorhees, E. M. (2007). TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM*, *50*(11), 51-54.

Voorhees, E. M. (2009, July). Topic set size redux. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 806-807).

Voorhees, E. M. (2019). The evolution of Cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, 45-69.

Voorhees, E. M., Craswell, N., & Lin, J. (2022, July). Too Many Relevants: Whither Cranfield Test Collections?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2970-2980).

Webber, W., Moffat, A., & Zobel, J. (2008, July). Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 51-58).

Webber, W., Moffat, A., Zobel, J., & Sakai, T. (2008, July). Precision-at-ten considered redundant. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 695-696).

Webber, W., & Park, L. A. (2009, July). Score adjustment for correction of pooling bias. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 444-451).

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

Wilkerson, S. (2008). Application of the Paired t-test. *XULAneXUS*, *5*(1), 7.

Wu, H. (2016). *Improving efficiency and flexibility of information retrieval systems*. University of Delaware.

Yilmaz, E., & Aslam, J. A. (2006, November). Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 102-111).

Yilmaz, E., & Aslam, J. A. (2008). Estimating average precision when judgments are incomplete. *Knowledge and Information Systems*, *16*(2), 173-211.

Zampieri, F., Roitero, K., Culpepper, J. S., Kurland, O., & Mizzaro, S. (2019, July). On topic difficulty in IR evaluation: the effect of systems, corpora, and system components. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 909-912)

Zhu, D., Nimmagadda, S. L., Wong, K. W., & Reiners, T. (2022). Relevance Judgment Convergence Degree--A Measure of Inconsistency among Assessors for Information Retrieval. *arXiv preprint arXiv:2208.04057*

Zobel, J. (1998, August). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 307-314).

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, *62*(1), 75-86.

Zuva, K., & Zuva, T. (2012). Evaluation of information retrieval systems. *International journal of computer science & information technology*, *4*(3), 35.