DCU School of Computing

Practicum Paper

Students' Name(s):	Anjali Kumari, Agnideep Mukherjee
Student Number(s):	22261032, 22260149
Program:	Master of Science in Computing
Practicum Title:	Covid-19 Information Retrieval
Module code:	CA685
Supervisor:	Alan Smeaton
Project Due Date:	31st July 2023

Declaration

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Anjali Kumari Date: 31/07/2023

Name: Agnideep Mukherjee Date: 31/07/2023

DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Citation-Driven Improvements in Information Retrieval: A TREC COVID Exploration

1st Anjali Kumari #22261032
School of Computing, Dublin City University
Dublin, Ireland
Email: anjali.kumari2@mail.dcu.ie

2nd Agnideep Mukherjee #22260149
School of Computing, Dublin City University
Dublin, Ireland
Email: agnideep.mukherjee2@mail.dcu.ie

***Abstract*—Information retrieval (IR) technology is the foundation of Web-based search engines and is crucial to biomedical research due to its influence on the development of the software that supports literature searches. It is a branch of computer science which deals with how to process documents with free text so they may be quickly retrieved using keywords from a user's search. The most successful models used for IR are probabilistic. This paper provides an overview of structure, finding, and exploration of information retrieval within the complex information landscape of the TREC COVID benchmark. By utilising the widely used Okapi BM25 information retrieval system, we conducted extensive information searching and analysis. The main objective of our study is to determine if adding citation information from documents to the corpus could improve the baseline score in information retrieval. Our findings showed promising results as including citation details led to noticeable enhancements in the baseline score. This improvement not only made the retrieval process more effective but also emphasised the important role that citation data can play in refining overall retrieval performance.**

I. Introduction

The COVID-19 pandemic brought both a public health catastrophe and an information crisis. More inaccurate or incomplete information can spread on the worldwide internet quicker than a virus can in the real world, leading to a rise in the demand for virus-related information. The information retrieval community responded by looking at how we may better meet these demands through the TREC-COVID challenge [8]. TREC-COVID is community evaluation to build a pandemic document or information retrieval (IR) system in order to support clinicians and clinical research.

The Text Retrieval Conference (TREC), which was first held in 1992 as a component of the TIPSTER Text was co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defence. By providing the infrastructure required for extensive examination of text retrieval strategies, it served the purpose to support research within the information retrieval field. [7]

In TREC and in other IR tasks, the data which is dealt with are mostly academic papers. In the basic anatomy of a published paper, retrieval in general mostly revolves around the usage of the title, abstract or body of the text. However, in an attempt to get extra accuracy and to find additional information in a paper, the use of citations is often overlooked.

In the realm of academic writing, citations play a pivotal role in ensuring the credibility, authenticity, and reliability of research papers. A citation in a paper refers to an information source used to support or reinforce an argument in an academic article. Scholarly articles, books, websites, interviews, and other publications are examples of possible sources for citations. These citations also might hold the key to increasing the retrieval effectiveness of a given document [6].

Following our thorough investigation of the many ways that various research teams have taken to obtain good retrieval accuracy in the TREC-COVID IR benchmark, we discovered that they are largely based on the use of ensemble methods, advanced IR toolkits, and neural networks. As a result, we determined that enriching the corpus of text documents in the TREC-COVID challenge by using citation analysis is a viable option and worth investigating, which is what we do in this paper [1], [6], [12], [13].

This paper's main objective is to take an off-the-shelf information retrieval system, Okapi BM25, and replicate information searching in a complex information landscape, specifically the TREC COVID benchmark. Alongside, we shall delve deep into our research questin which is *"can we elevate our baseline score for BM25 retrieval by enriching the text corpus with invaluable citation information?"* Our evaluation will be conducted using the TREC-Eval framework, ensuring rigorous assessment and insightful outcomes.

II. Literature Review

In the TREC-COVID overview paper [7], an overview of the entire TREC-COVID challenge was presented. The Allen Institute for AI provided the COVID-19 Open Research Dataset (CORD-19) and is available at https://github.com/allenai/cord19. About 150,000 scholarly publications were made available by CORD-19, both in embedded and parsed pdf.json formats.

The TREC-COVID competition is typically organised into rounds, and each round consists of different tasks designed to evaluate the performance of participating IR systems. The goal of each participating team is to generate runs that perform well according to the evaluation metrics for the given task and round, and to use their runs to demonstrate the effectiveness of their retrieval methods and approaches [8]. In the context of TREC-COVID benchmark, a "run" refers to a specific system submission by a participating team for a particular task and round. Each run consists of a list of retrieved documents or answers generated by the participant's system for a given set of queries, along with any additional metadata or information that the team wishes to provide about the system or the retrieval process. In each round, the participating systems' results are evaluated based on various metrics such as accuracy, average precision, and mean reciprocal rank, which measure how well the systems perform in retrieving and finding the relevant documents. These evaluations provide valuable insights into the effectiveness of different IR methods for COVID-19 literature and help advance the development of more effective methods for searching and retrieving relevant information.

TREC-COVID received 556 submissions with 92 participating research teams over the course of five evaluation rounds. The complete test collection includes 69,318 manual evaluations on 50 crucial COVID-19 topics [3]. We now look into some of the top participants and the methods they have used to get high retrieval scores.

A. Participants

Participants in the TREC-COVID 19 track used a variety of information retrieval techniques in order to make their submissions, including traditional Boolean and vector-space models, as well as more recent techniques such as deep learning and neural network-based models. The performance of the systems was evaluated using standard metrics such as mean average precision (MAP), normalised discounted cumulative gain (nDCG), binary preference (bpref), and precision at 20 (P@20) [1].

In order to understand how some of the teams dealt with this challenge, we picked 4 participating teams out of 20 on the basis of their nDCG@20 scores in the final round (round 5) published on the TREC covid archive https://ir.nist.gov/covidSubmit/archive.html. We picked 3 top performing teams — unique-ptr, covidex and Elhuyar-NLP-team —– and also a team named xj4wang which was ranked 15. Our goal is to demonstrate how the teams used different methodologies, although the concept of corpus enrichment using citation information is still unexplored among participating groups.

1) Team: unique-ptr: This team has used the ensemble approach named as 'Hierarchical rank fusion ensemble' to bring out the best results in the final round. In order to obtain this the result, the team had firstly used a combination of different retrieval mechanisms to obtain a comprehensive document set to increase recall. Secondly, to obtain high precision the team has used multiple BERT based rankers on the obtained document set.

In order to get optimum performance the team decided to use reciprocal rank fusion (RRF), which organises documents using a scoring algorithm, where the score of the document is the sum of the reciprocal ranks of the runs. But due to usage of different systems in this ensemble, the number of runs from each of the systems may vary. Thus, a system having more runs can incorrectly bias the result. To mitigate this problem, Hierarchical Reciprocal Rank Fusion (h-RRF) was used where the runs from the different system were divided among the sub-pools and then recursively the scoring algorithm was applied. However due to inequality in the performances of the various system, each of the system had to be weighted as well.

The different systems used in the h-RRF were as follows.

1) Lexical Retrieval System: The team had used two well-known open source search engines to design lexical retrieval runs, Terrier and Anserini. Terrier was selected on the basis of the ease of query language, thorough documentation and easy to configure common retrieval algorithms [1]. However, Anserini was used with the help of the runs published by the 'covidex' team, made available for general use by the participants [13];
2) Relevance Feedback Systems: The team had designed two relevance feedback runs on the basis of the Terrier system with the abstract indexed plus one with the query+question field expanded by 300 terms from the 10 highest ranked relevant documents and another with the query+question field expanded by 1,000 terms from 30 highest rank documents [1].
3) Semantic Retrieval System: This was designed by conglomerating 3 sub models, a Neural Retrieval Model based on BERT, a Synthetic Question Generator and a Hybrid Retrieval System. The Neural Retrieval model is based on the class of dual encoder models which encodes pair of items in dense sub-spaces. To train a dual encoder model a mammoth amount of training data is required. Due to the shortage of this in the biomedical domain, Synthetic Question Generator was used. This uses an encoder-decoder generation model which generates questions relative to a certain passage. It was used to generate approximately 166 million question and abstract pairs to train the model. As the author said although the duel encoder models excels in finding the semantic similarity, however it performs below par in lexical matching. To deal with this the duel encoder model was merged with the strength of the BM25 model in term matching. BM25 will be explained in detail later in this paper.
4) BERT Rankers: The team fine-tuned the BERT re-ranking model based on the data present in the MS-MACRO dataset [1].

For this participant, the ensemble of the systems were used to create a robust model which had a nDCG@20 and MAP scores of 0.8496 and 0.4718 respectively which was the best run among all the submitted results in round 5 of the TREC-COVID benchmark. This is referred from the data published on TREC Covid archive at https://ir.nist.gov/covidSubmit/archive.html

2) Team: Covidex: This team had taken a 3-step strategy to meet the goals of the Information retrieval in the TREC COVID benchmark. Firstly, the team thinks that despite of the advances of Neural Networks in text ranking, the basic keyword search algorithm like Bag-of-words still remains sturdy and robust. On top of the Anserini IR toolkit, the team have released a user-friendly open source key-word search system that the community can use. Secondly, the team has explored the usage of a sequence-to-sequence transformer model for text ranking with classification-based feedback by using the existing relevance judgements. This system was also made open source for the community. Lastly, the team merged the previous two components into Covidex, which is a search engine and browsing interface deployed at covidex.ai [13].

The team thought that instead of making a monolithic ranker it would be wiser to create a multi-stage architecture which begins with an initial retrieval stage by using bag-of-words queries against an inverted index and in the next stage it would be re-ranked and refined. The team has used the Anserini IR toolkit for the initial retrieval stage which the team had been developing for several years. Anserini provides abstraction for document collections and also can incorporate various different corpus and formats like XML documents in tarballs, JSON objects in text files, etc. The ranking quality is then further improved by the use of re-rankers like 'monoT5'. Anserini returns a query and a set of candidate documents which is in turn fed to the re-ranking model. The model is fine-tuned to return either 'True' or 'False' depending upon the relevance of the document to the query. After that the team had re-ranked the candidate document based on the probabilities assigned to the true token. The fine-tuning of the monoT5 model has been done at the rate of 0.01 with 10k iterations with class-balanced batches of size 128 [13].

The team has also built Covidex as an operational search engine which is built using the Python FastAPI framework. Searching, re-ranking and text highlighting are performed on the incoming API data. The search is performed by Anserini, and results are re-ranked by monoT5 and PyGaggle. The front-end is built on REACT JS [13]. Using the built model, team Covidex has performed well and ranked 2nd among the top teams in round 5. Their nDCG@20 and MAP scores were 0.8311 and 0.3922 respectively which tells us how robust their multi-stage architecture has worked. This is referred from the data published on the TREC Covid archive at https://ir.nist.gov/covidSubmit/archive.html.

3) Team: Elhuyar-NLP-team: In this paper [11], the team members have proposed 7 different techniques and models to improve the ranking of documents.

1) Run 1: The first run, BM25-bs, is the baseline run, which uses the BM25 algorithm to rank the documents based on their relevance to the query;
2) Run 2: The second run, BM25-dfr-lmd-rrf, is a fusion of the BM25 algorithm, DFR (Divergence from Randomness) model, and LMD (Language Model with Dirichlet smoothing) with the RRF (Rank Reciprocal Fusion) technique. This run combines different retrieval models to improve the ranking of the documents;
3) Run 3: The third run, trf-brx-rrf, uses the fusion of BERT, RoBERTa, and XLNet, which are neural language models trained independently from the ranked list of run 2. This run uses deep learning techniques to improve the ranking of the documents;
4) Run 4: The fourth run, ir-bdl-trf-brx-lm, re-ranks the results based on LambdaMART, a machine learning algorithm that uses decision trees to improve the ranking of the documents;
5) Run 5: The fifth run, ir-bdl-trf-brx-rrf, combines the results of run 2 and run 3 using RRF to further improve the ranking of the documents;
6) Run 6: The sixth run, bdl-brx-logit, takes the classic IR features, along with neural language model features and simply uses logistic regression to re-rank the results. This run combines different features to improve the ranking of the documents;
7) Run 7: The seventh run, ir-trf-logit-rrf, merges the results of run 3 and run 5 using RRF to further improve the ranking of the documents. This run combines different retrieval models and machine learning algorithms to improve the ranking of the documents.

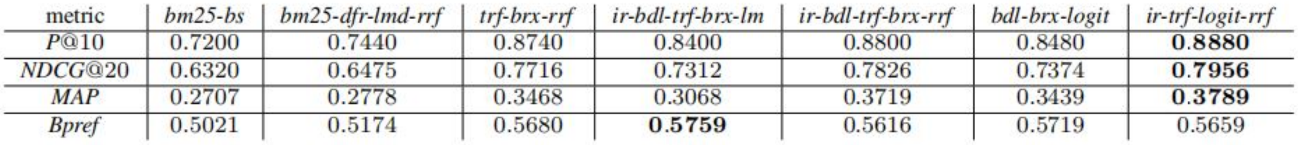
Overall, the different runs in this information retrieval system use a combination of different techniques, models, and algorithms to improve the ranking of the documents based on their relevance to the query. Figure II-A3 shows all the nDCG@20 and MAP scores for the different models.

4) Team: xj4wang: The paper [12] discusses this team's participation in the TREC 2020 COVID track using a continuous active learning approach. The authors used a Continuous Active Learning (CAL) method to improve the effectiveness of their information retrieval system. This approach is used to gather the relevant information related to a particular subject from the huge repository of Electronically stored information. Based on the feedback of the already judged documents by the user it refines the relevancy of it's search. This process is already tried and tested in technology assisted review for electronic discovery in legal matters [12].

The team underwent 4 steps in their implementation of CAL.

1) They produced synthetic documents which are hypothetically relevant. In order to create this the team had concatenated the narrative, question and the query components of the topic file provided by the TREC COVID [12].
2) The team further used Sofia-ML, a Machine Learning algorithm to fetch the relevancy of the document search [12].
3) Furthermore, to provide a relevance feedback mechanism to Sophia ML, the team used manual human annotations. The top results of the Sophia ML were sent to human annotators using a text based user interface. The score is 0 if not relevant, 1 if partially relevant and 2 if totally relevant. This corresponds to the result provided by biomedical experts after the end of each round in TREC COVID [12].
4) Steps 2 and 3 are repeated until the number of non-relevant documents get diminished under a certain threshold.

As the team suggested, there is one major setback in usage of CAL. In CAL a huge number of documents must be reviewed when the number of relevant documents is large which is not feasible in the long run. Thus, to overcome this issue, the team has leveraged the use of the Scalable



metric	*bm25-bs*	*bm25-dfr-lmd-rrf*	*trf-brx-rrf*	*ir-bdl-trf-brx-lm*	*ir-bdl-trf-brx-rrf*	*bdl-brx-logit*	*ir-trf-logit-rrf*
P@10	0.7200	0.7440	0.8740	0.8400	0.8800	0.8480	**0.8880**
NDCG@20	0.6320	0.6475	0.7716	0.7312	0.7826	0.7374	**0.7956**
MAP	0.2707	0.2778	0.3468	0.3068	0.3719	0.3439	**0.3789**
Bpref	0.5021	0.5174	0.5680	**0.5759**	0.5616	0.5719	0.5659

Figure 1. nDCG@20 and MAP scores for the different models used by Team: Elhuyar-NLP-team in TREC-COVID

Continuous Active Learning (S-CAL). In S-CAL, the corpus is segmented into batches and the annotator needs to assess only a sub strata of relevant sample from each batch [12]. To assist the human assessor, the team had used key-term highlighting to highlight the top 5 highest scoring words in a document provided by Sophia-ML [12].

The study demonstrates the potential of continuous active learning for COVID-19 information retrieval and provides a useful tool for researchers and practitioners working in this area. The major limitation is the need for manual annotation, which can be a painstaking task when the number of documents is high and the team size is small. Overall the team had performed decent in the results published on TREC Covid archive athttps://ir.nist.gov/covidSubmit/archive.html. From this we can see that the nDCG@20 and MAP scores were 0.6663 and 0.2448 respectively. Though it is significantly less than the top performers, yet we got to understand the pros and cons of CAL in the field of information retrieval [12].

B. Motivation

Since the 1960s, citation analysis has been a significant area of study in information science, and as a result, citation data has been employed in information retrieval previously [6]. Subject indexing and citation indexing have traditionally been the two basic strategies for searching the literature. In subject indexing, certain papers are sought after using descriptors of their literature [2], [6]. In citation indexing, the citations between documents are recorded and documents are located by following these links. Citation indexing is used to establish an inter-document relationship, where the similarity between documents A and B is measured by the number of documents that cite both A and B. Dunlop et al. have used similar citation indexing for the retrieval of non-textual documents like image, sound and video files [2].

On the contrary, O'Connor et al. [4] have ingeniously developed a rudimentary IR prototype for chemistry journals. The team painstakingly enriched the indexing by incorporating abstract information from the cited papers. O'Connor reached the insightful conclusion that citing statements can be beneficial; however, the identification of citation information proved to be a daunting task, requiring human intervention and mostly relying on knowledge of sentence boundaries [4], [5].

Ritchie's research aligns closely with our own efforts. Her innovative approach involves formulating ideas to create diverse sizes of word windows encircling in-line citations, thereby augmenting an existing corpus with supplementary information to achieve enhanced retrieval effectiveness. For evaluation, Ritchie employs popular off-the-shelf IR toolkits such as Okapi BM25, Indri, and KL Divergence, systematically assessing the impact of integrating the citation-wrapping sentences on baseline scores [5], [6].

Remarkably, none of the existing research has harnessed the direct potential of citation titles to enhance the retrievability of a pre-existing corpus. We firmly believe that citation titles, being the very sources referenced by authors to substantiate their research, harbour a key to infusing invaluable significance into a corpus, thereby reaping substantial benefits for information retrieval.

III. Our Research

A. Implementation

1) Dataset: The COVID-19 Open Research Dataset (CORD-19) was taken from https://github.com/allenai/cord19 and used in the work in this paper. We have used the metadata.csv as the starting point wherein the $cord\_uid$ is a string-valued field that assigns a unique identifier to each CORD-19 paper, $title$ is a string-valued field for the paper title, $abstract$ is a string-valued field for the paper's abstract, pdf_json_files is a List[string]-valued field containing paths from the root of the current data dump version to the parses of the paper PDFs into JSON format. Multiple paths are semicolon-separated. Further, we have used the collection of 84,000 JSON files that contain full-text parses of a subset of CORD-19 papers.

2) Data Exploration and Data Pre-Processing: At first we visualised and pre-processed the above mentioned data. Figure2 depicts the distribution of documents within the TREC Covid dataset based on the number of citations in each document. The spectrum spans from the lowest citation count of 0 to the highest at an impressive 2440 citations from within a single document. Unremarkably, around 90% of the data falls within the range of 0 to 150 citations, highlighting the predominant concentration within this citation count bracket.

For the pre-processing of the data, we cleaned the data first by removing duplicate documents and proceeded with only unique documents. After that from each of the document we removed the stopwords and special characters, performed stemming with the help of the $nltk$ library in Python. We have also performed data normalisation in the citation as the number citations in each document were not balanced. We normalised our citation's title by keeping the limit from 0 to 150 combination of words from each of the titles for each

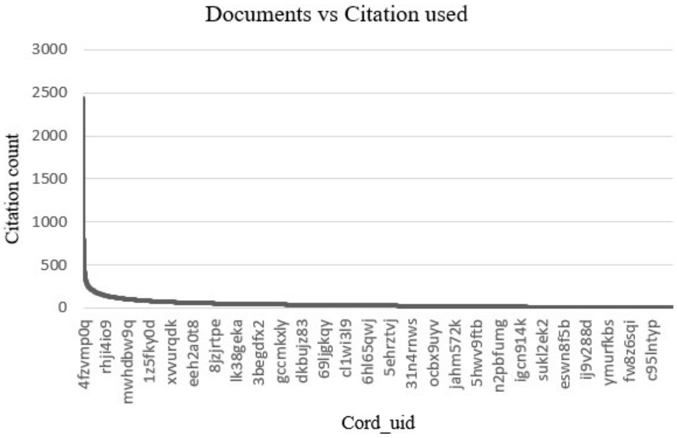


Figure 2. Distribution of document on citation count

document. Finally tokenisation was performed on the cleaned data.

3) Okapi BM25: For the "bag of words" document retrieval, BM25 is possibly the most well-known scoring function [10]. This probabilistic framework for IR has evolved from the binary independence relevance model to incorporate within-document term frequency data and document length normalisation [9]. Despite the widespread usage of learning-to-rank techniques and neural ranking models today, these techniques are often used as candidate documents obtained using a straight forward term-matching technique utilising conventional inverted indexes. This is frequently done using BM25, thus even now, search applications still heavily rely on this seemingly obsolete scoring formula. A set of documents are ranked by BM25 according to how often certain search terms appear in them.

Given a query Q, with keywords $q_1...q_n$ the BM25 score of a document D is given as:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}$$

Figure 3. Formula for BM25

where $f(q_i, D)$ is the number of times that q_i occurs in the document D, $|D|$ is the length of the document D in words, and $avgdl$ is the average document length in the text collection from which documents are drawn, k_1 and b are free parameters frequently selected in the absence of a sophisticated optimisation. IDF(q_i) is the inverse document frequency weight of the query term q_i, it normally computes as shown in Figure 4.
where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

We have used the above 2 equations to determine document scores by trying different combinations of values for $k1$ and b of each document. The best scores waere achieved by the combination of k1= 1.5 and b= 0.5.

$$\text{IDF}(q_i) = \ln\left(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}+1\right)$$

Figure 4. Formula for IDF

B. Our Results

context	Evaluation Metrics			
	ndcg20	P20	bpref	map
nc-50-0	0.1873	0.1721	0.1918	0.0434
nc-150-0	0.1843	0.1700	0.1880	0.0430
nc-50-50	0.1818	0.154	0.191	0.0409
unnormalized	0.1755	0.145	0.1898	0.0369
nc-100-100	0.1749	0.1330	0.1888	0.0366
No citation	0.1715	0.1671	0.1697	0.0388
nc_150_150	0.1676	0.131	0.1838	0.0339
nc_300_300	0.1589	0.127	0.1833	0.0262

Figure 5. retrieval scores

The table in Figure 5 presents a comprehensive evaluation of information retrieval scores under various citation contexts, providing valuable insights into the system's performance. The "No Citation" row stands as the benchmark score, representing our system's baseline performance in the absence of any contextual information.

As we observed in Figure 2, the citation counts for documents in the collection exhibited a significant range, spanning from a remarkable 2440 to 0. To harness the full potential of citations in our corpus, we made the decision to normalise the includsion of this citation data in document representations. To represent these normalised citations, we adopted the notation "nc-x-y" where "x" denotes the upper bound and "y" signifies the lower bound.

For instance, consider "x=10". In this scenario, if the total number of words in a citation exceeds 10, we only consider the top 10 most frequent words present in the citations. On the other hand, when "y=10" if the total number of words in a citation falls below 10, we replicate the citation words until the total count reaches 10.

By employing this normalisation approach, we aimed to level the playing field and ensure equitable treatment of citations throughout the corpus.

C. Evaluation

To calculate the percentage increase and decrease of various parameter combinations for including citation information vs.

context	Change in ndcg20	Change in P20	Change in bpref	Change in map
nc-50-0	9.23%	2.99%	13.02%	11.86%
nc-150-0	7.46%	1.74%	10.76%	10.31%
nc-50-50	6.01%	**-7.84%**	12.58%	5.41%
unnormalized	2.33%	**-13.25%**	11.86%	**-4.89%**
nc-100-100	1.98%	**-20.44%**	11.27%	**-5.67%**
nc_150_150	**-2.28%**	**-21.58%**	8.32%	**-12.63%**
nc_300_300	**-7.36%**	**-23.96%**	7.97%	**-32.48%**

Figure 6. Our scores relative to the benchmark

the benchmark score (No citation), we can use the following formula:

$$Change = ((NewVal - Benchmark)/Benchmark) * 100$$

The percentage changes for each row with respect to the benchmark is shown in Figure 6.

The first notable outcome from these results is that they confirm that adding citation terms improves retrieval effectiveness, however there are few anomalies.

A striking observation emerges as we delve into the data —- every instance of adding citation information has consistently led to an increase in bpref compared to its benchmark score. From this, we can infer that the inclusion of citations provides additional information that helps the retrieval system better identify and rank relevant documents. When citations are added, the retrieval system has more context and knowledge about the relationships between documents, which can lead to improved rankings of relevant documents in the result set.

Looking at the table, we notice that the MAP and p@20 scores vary for different citation contexts. Some citation contexts, such as "nc-150-0," "nc-50-50," and "nc-50-0" show higher MAP scores than the benchmark score, suggesting an improvement in the overall precision performance of the retrieval system when citation information is incorporated. However, an intriguing observation arises for contexts with a broader window of words, where the MAP and P@20 scores have diminished. This suggests that larger citation contexts may pose challenges, potentially due to the complexities arising from an extensive number of citations. However, there is a sweet spot which can be used to push the benchmark performance in terms of precision.

Notably, we observe that introducing a lower boundary while normalising the citation data adversely impacts the retrieval scores. The top-performing corpus reveals no implementation of lower boundaries. A striking comparison between nc-50-0 and nc-50-50 elucidates that nc-50-0 outperforms its counterpart in all aspects. This suggests that duplicating the citation information to enhance its weight backfires, as it may accumulate redundant data, consequently compromising the corpus's quality.

Taking a panoramic view and focusing on the ndcg20 scores, our conclusion emerges that the inclusion of citation information indeed elevates the overall system performance, yet the key lies in smart normalisation. When the citation window is excessively large, the impact turns adverse. However, with an optimal window size carefully cultivated, we witness a remarkable 9% increase in ndcg and a 12% boost in MAP. Moreover, it becomes evident that increasing the lower bound proves counterproductive, as it tends to hamper the system's performance. Through these insightful observations, we gain valuable guidance in refining our citation strategies, steering us towards the improvement of information retrieval.

IV. Conclusions

Through our practicum, we successfully pursued the main objective of exploring information retrieval within the intricate information retrieval landscape, exemplified by the TREC COVID benchmark. Leveraging the widely used off-the-shelf information retrieval system Okapi BM25, we conducted comprehensive information searching and analysis. The central focus of our investigation revolved around the research question: "Can we enhance the baseline score by enriching the corpus with added citation information?"

Our findings showcased promising outcomes, as enriching the corpus with citation information yielded noticeable improvements in the baseline score. The inclusion of citation details proved to be a valuable augmentation, shedding light on the significant impact of citation context in information retrieval tasks. This enhancement not only amplified retrieval effectiveness but also exemplified the crucial role that citation data can play in refining the overall retrieval process.

While this paper lays a solid foundation for understanding the potential of citation information in improving information retrieval performance, numerous avenues for future research and development emerge from our work. Some key areas of future exploration include:

1) **Fine-tuning models**: Investigating advanced techniques and machine learning models to optimise the incorporation of citation data into the retrieval process. This could involve neural networks or other state-of-the-art models to harness the full potential of context-rich information;

2) **Scalability and efficiency**: Exploring methods to handle larger and more diverse citation datasets efficiently to ensure practical applicability in real-world information retrieval scenarios;
3) **Incorporating Advanced IR Techniques**: Integrating other advanced information retrieval techniques, such as query expansion, relevance feedback, or context-aware retrieval, to further refine the retrieval results;
4) **Citation Parsing and Extraction**: Developing improved algorithms for accurate citation parsing and extraction, ensuring the quality and reliability of the enriched corpus.

Acknowledgements

Heartfelt gratitude is extended to Professor Alan Smeaton for his invaluable and thorough guidance throughout our journey. We also express our sincere appreciation to Professor Gareth Jones for generously sharing enlightening contents and study materials on Information Retrieval, which proved immensely beneficial in the development of our project. Special thanks are due to our esteemed Institution for providing us with a robust infrastructure, abundant resources, and powerful computer systems, enabling us to conduct successful research and execute complex programs seamlessly. Their unwavering support paved the way for our scholarly endeavors and contributed significantly to the realization of our research goals.

References

[1] Michael Bendersky, Honglei Zhuang, Ji Ma, Shuguang Han, Keith Hall, and Ryan McDonald. RRF102: Meeting the TREC-COVID challenge with a 100+ runs ensemble. *arXiv preprint arXiv:2010.00200*, 2020.

[2] M.D. Dunlop and C.J. van Rijsbergen. Hypermedia and free text retrieval. *Information Processing & Management*, 29(3):287–298, 1993.

[3] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ Digital Medicine*, 4(1):68, 2021.

[4] John O'Connor. Citing statements: Computer recognition and use to improve retrieval. *Information Processing & Management*, 18(3):125–131, 1982.

[5] Anna Ritchie. *Citation Context Analysis for Information Retrieval*. PhD thesis, University of Cambridge, Computer Laboratory, March 2009.

[6] Anna Ritchie, Stephen Robertson, and Simone Teufel. Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 213–222, New York, NY, USA, 2008. Association for Computing Machinery.

[7] K. Roberts, T. Alam, and S. et al Bedrick. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *Journal of biomedical informatics*, 121, 2021.

[8] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 07 2020.

[9] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[10] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. 1994.

[11] Douglas Teodoro, Sohrab Ferdowsi, Nikolay Borissov, Elham Kashani, David Vicente Alvarez, Jenny Copara, Racha Gouareb, Nona Naderi, and Poorya Amini. Information retrieval in an infodemic: the case of covid-19 publications. *Journal of Medical Internet Research*, 23(9):e30161, 2021.

[12] Xue Jun Wang, Maura R. Grossman, and Seung Gyu Hyun. Participation in TREC 2020 COVID track using continuous active learning. *arXiv preprint arXiv:2011.01453*, 2020.

[13] Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, et al. Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset. *arXiv preprint arXiv:2007.07846*, 2020.