Journal of Biomedical Informatics 117 (2021) 103745

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

A comparative analysis of system features used in the TREC-COVID information retrieval challenge

Jimmy S. Chen [a,*], William R. Hersh [b]

[a] School of Medicine, Oregon Health & Science University, Portland, OR, USA
[b] Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

ARTICLE INFO

Keywords:
COVID-19
TREC-COVID
Information retrieval
Search systems
Taxonomy

ABSTRACT

The COVID-19 pandemic has resulted in a rapidly growing quantity of scientific publications from journal articles, preprints, and other sources. The TREC-COVID Challenge was created to evaluate information retrieval (IR) methods and systems for this quickly expanding corpus. Using the COVID-19 Open Research Dataset (CORD-19), several dozen research teams participated in over 5 rounds of the TREC-COVID Challenge. While previous work has compared IR techniques used on other test collections, there are no studies that have analyzed the methods used by participants in the TREC-COVID Challenge. We manually reviewed team run reports from Rounds 2 and 5, extracted features from the documented methodologies, and used a univariate and multivariate regression-based analysis to identify features associated with higher retrieval performance. We observed that fine-tuning datasets with relevance judgments, MS-MARCO, and CORD-19 document vectors was associated with improved performance in Round 2 but not in Round 5. Though the relatively decreased heterogeneity of runs in Round 5 may explain the lack of significance in that round, fine-tuning has been found to improve search performance in previous challenge evaluations by improving a system's ability to map relevant queries and phrases to documents. Furthermore, term expansion was associated with improvement in system performance, and the use of the narrative field in the TREC-COVID topics was associated with decreased system performance in both rounds. These findings emphasize the need for clear queries in search. While our study has some limitations in its generalizability and scope of techniques analyzed, we identified some IR techniques that may be useful in building search systems for COVID-19 using the TREC-COVID test collections.

1. Introduction

Since the World Health Organization declared the Coronavirus Disease 2019 (COVID-19) a public health emergency [1], there has been explosive growth in scientific knowledge about this novel virus. Consequently, the use of preprints and fast-track publication policies has resulted in a significant increase in the number of COVID-19 related publications over a short period of time [2,3]. Information retrieval (IR, also known as search) systems are the tool usually employed to manage access to large corpora of literature [4]. The efficacy of IR systems is often assessed in challenge evaluations that provide reusable test collections, such as those led by the National Institutes of Standards and Technology (NIST) in the Text REtrieval Conference (TREC) [5].

To address the need for system evaluation in this rapidly changing information environment, NIST sponsored the TREC-COVID Challenge [6]. Similar to prior IR challenge evaluations, test collections of documents, topics for searching, and relevance judgments were developed [7]. Given the rapidly evolving climate of information in a global pandemic, the structure of the TREC-COVID Challenge differed from typical TREC track in two key ways [8]. First, unlike the static data collections used in prior challenges, the document collections were derived from snapshots of the COVID-19 Open Research Dataset (CORD-19), an approximately weekly updated dataset of manuscripts consisting of coronavirus-related research gathered from various sources including journal articles, PubMed references, arXiv, medRxiv, and bioRxiv [3]. Each iteration of the CORD-19 dataset contained up-to-date articles with document IDs, bibliographic metadata, as well as each article's title, abstract, and full-text, which was available in most of the articles [3]. Second, compared to prior challenges where teams were allowed multiple weeks to months to develop and fine-tune retrieval systems, the TREC-COVID challenge operated on a compressed schedule, with only 1–3 weeks per round over 5 consecutive rounds [8]. This is in part due to

* Corresponding author at: 3181 SW Sam Jackson Park Rd., Oregon Health & Science University, Portland, OR 97239, USA.
E-mail address: chenjim@ohsu.edu (J.S. Chen).

https://doi.org/10.1016/j.jbi.2021.103745
Received 16 October 2020; Received in revised form 2 December 2020; Accepted 5 March 2021
Available online 6 April 2021

the rapidly iterative nature of the CORD-19 dataset. Each team was allowed to submit 3–8 runs per round; a run consisted of ranked documents perceived by the IR system to be relevant to each topic. Between rounds, there was approximately 10 days for TREC evaluators to manually assess the relevance of documents from each iteration of the CORD-19 dataset, thus generating relevance judgments, which were then used to score IR systems and provide feedback for future runs [7,8]. Unlike prior challenges, each round was a superset of prior rounds: new documents and topics were added to the prior corpus and task list, though only relevance for newly added documents to each corpus were evaluated after each round [7,9]. Ultimately, this structure was designed to allow for iterative improvements to methodologies consistent with a dynamic dataset with the eventual goal of building a reusable test collection for future research [8].

There exists prior work retrospectively examining feature performances associated with retrieval performance in the medical domain. In a user study by Hersh et al, knowledge expertise between various medical trainees, presumed to be likely correlated with effective query formulation and searching strategies, was also associated with relevant manuscript retrieval [10]. Subsequent work by Repakalli et al used multivariate analysis to examine features of IR systems associated with retrieval performance in the TREC Genomics track [11]. More recently, Roberts et al performed a review of several systems developed in the TREC Clinical Decision Support Track in 2014 and highlighted some features of high-performing systems [12]. However, there is a gap in knowledge characterizing methods used and systems developed by participants in the TREC-COVID challenge for a dynamic document corpus.

To address this gap in knowledge, the purpose of this study was to compare performance in different approaches used in the TREC-COVID Challenge by: (1) developing a taxonomy to characterize IR techniques and system characteristics used, and (2) applying this taxonomy to identify features of IR systems associated with higher performance. Using run reports from Round 2 and Round 5, we designed a taxonomy and evaluated its features using a univariate and multivariate regression analysis. We performed a multivariate regression analysis to explore relationships between several independent features and their associations with performance. In this study, we assessed how certain methodologies were associated with higher retrieval performance and discussed the implications and limitations of our analysis.

2. Methods

2.1. Dataset and features

The TREC-COVID challenge [6] occurred over 5 rounds in 2020 on the rapidly growing CORD-19 dataset [3], with 30 initial topics in Round 1 and 5 new topics added each subsequent round. Each topic consisted of three fields: (1) a short *query* statement that a user might enter, (2) a longer *question* field more thoroughly expressing the information need of the topic, and (3) a *narrative* field describing what would constitute a relevant document. After Round 1, relevance judgments, consisting of IDs of manuscripts assessed by human assessors as relevant, partially relevant, or irrelevant, were made available for previously unassessed manuscripts after each round. Table 1 summarizes the CORD-19 corpora, topics, number of teams, runs, and judgments present in each round of the TREC-COVID challenge.

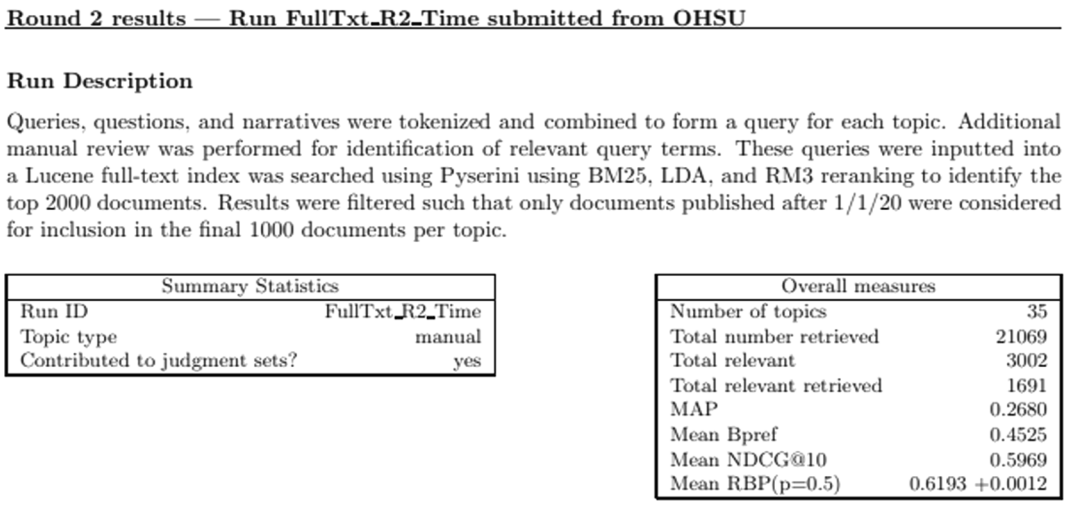
Reports for all submitted runs, including those from Rounds 2 and 5, were made publicly available from the TREC-COVID Challenge (https://ir.nist.gov/covidSubmit/archive.html) and were manually reviewed by author JC for features characteristic of IR systems. These features were validated by author WH, with disagreements resolved by discussion among both authors. Of note, though a leaderboard was included each round, no actual methodology beyond links to the same reports reviewed were provided. We chose to review reports from Rounds 2 and 5 because we wanted to compare methodologies used in two different rounds where feedback methods from topics from previous rounds were available. Each run report was written as a textual description of the methodology used to produce the run in whatever detail the submitting team provided. An example run report is shown (Fig. 1).

The following features were extracted for each run in the reports from Round 2 and Round 5:

- Text used (i.e., title and abstract only, paragraph-based indices, or full-text).
- Type of query (i.e., any combination of the query, question, and narrative from the TREC-COVID topic fields).
- Any query pre-processing (i.e., stemming, removing stop words).
- Query term expansion (addition of terms not originally provided in each topic).
- Manual review methods (i.e., human interventions including the use of human assessors in Continuous Active Learning) [13].
- Any weighted ranking system used (i.e. non-neural scoring functions such as BM25 [14] and term frequency–inverse document frequency, or TF-IDF [15]).
- Any ranking model that used a neural architecture (including deep transformer models such as BERT [16], SciBERT [17], T5 [18] as well as DeepRank [19], a neural network that attempts to simulate humans in relevance judgments).
- Other techniques (machine learning models such as SVM, logistic regression, custom scoring functions, otherwise known as term proximity scores. Custom search methods were also included in this category including ReQ-ReC [20], a double-loop retrieval system)
- Dataset used to fine tune any system (i.e., MS-MARCO, a large dataset of annotated documents based off 100,000 Bing queries [21], MED-MARCO, a subset of MS-Marco containing queries and documents exclusively in the medical domain [24], CORD-19 dataset transformed into document vectors, and relevance judgments from previous rounds).
- Fusion of multiple runs into a single run (including use of reciprocal rank fusion [22], COMB fusion methods [23]).
- Re-ranking implemented, defined as whether a second system (most commonly a neural network) was used to refine an initial scoring system.
- Pseudo-relevance feedback, or system-generated relevance feedback based on an initial query.
- How/if human-generated relevance feedback, or relevance judgements, from the previous round(s) were used.
- Runs filtered by date. Removing documents published before 2020 (or when the pandemic began to gain widespread notice) had been

Table 1
Overview of the TREC-COVID challenge. Over 5 rounds, research teams implemented information retrieval (IR) systems to search the growing CORD-19 dataset. After each round, new topics and relevance judgments of manuscripts from previous iterations of the CORD-19 dataset were released for use in subsequent rounds.

Round	CORD-19 Date	Documents	Docs Changed	Topics - new/total	Teams	Runs	Cumulative Judgments
1	4/10/2020	51,103	N/A	30/30	56	143	8691
2	5/1/2020	59,851	20	5/35	51	136	20,728
3	5/19/2020	128,492	2017	5/40	31	79	33,068
4	6/19/2020	157,817	104	5/45	27	72	46,203
5	7/16/2020	191,175	1137	5/50	28	126	69,318



Round 2 results — Run FullTxt_R2_Time submitted from OHSU

Run Description

Queries, questions, and narratives were tokenized and combined to form a query for each topic. Additional manual review was performed for identification of relevant query terms. These queries were inputted into a Lucene full-text index was searched using Pyserini using BM25, LDA, and RM3 reranking to identify the top 2000 documents. Results were filtered such that only documents published after 1/1/20 were considered for inclusion in the final 1000 documents per topic.

Summary Statistics	
Run ID	FullTxt_R2_Time
Topic type	manual
Contributed to judgment sets?	yes

Overall measures	
Number of topics	35
Total number retrieved	21069
Total relevant	3002
Total relevant retrieved	1691
MAP	0.2680
Mean Bpref	0.4525
Mean NDCG@10	0.5969
Mean RBP(p=0.5)	0.6193 +0.0012

Fig. 1. Example Run Report from Round 2. During submission of a run, participants were encouraged to provide a methodological description of each submitted run. This run description, along with the run ID, topic types, and performance metrics, were reported in a publicly available repository of archived results (https://ir.nist.gov/covidSubmit/archive.html). These run reports were manually reviewed for features. Runs with less than 1 identifiable feature or unusually poorly performing runs were excluded from our analysis.

previously suggested by McAvaney et al. in their post-hoc analysis of their neural re-ranking system as a possible method to improve performance [24].

These features were selected to be as independently predictive as possible of retrieval performance (i.e., to minimize collinearity) and to encompass a broad set of commonly used techniques in ad-hoc retrieval. Of note, TREC challenges typically do not occur over multiple rounds; thus, the addition of relevance judgments was a novel addition to the TREC-COVID challenge. Since the length of reports varied at the researcher's discretion, many reports likely had some number of missing features. To minimize the impact of these null features, we assumed that runs that did not provide information about the type of text or query used in their system likely searched on the full-text using the query subfield from each topic. Features extracted from reports were either characterized as a binary feature (used or not used in the system) or a categorical feature (a description of feature for the system, i.e., BM25 as the weighted system used). Categorical features were later one-hot encoded, or converted into binary features over multiple columns, prior to input into our regression analysis. The extracted features and their encoding is shown in Table 2.

We included all runs that contained more than 1 extracted feature to ensure a reasonably large enough and useful dataset for analysis. We excluded unusually poorly performing runs that likely represented poor system or method implementations. The exclusion threshold for these runs was defined as an average performance of less than 0.2 across all 5 performance metrics used to evaluate run performance in the TREC-COVID Challenge. Performance metrics reported by NIST and used in our analysis included: precision at K documents (P@K), normalized discounted cumulative gain at K documents (NDCG@K), rank-based precision with depth = 5 (RBP (p = 0.5)), binary preference (bpref), and mean average precision (MAP). In Round 2, the depth of documents for P@K and NDCG@K were 5 and 10 respectively, while in Round 5, the depth of documents for P@K and NDCG@K were both 20. These changes were made out of concern for inflated performance when evaluating precision on a small number of documents [6]. For each run, these performance metrics were computed as the mean performance across all topics in the round.

2.2. Univariate and multivariate analysis

All data analysis and pre-processing was performed using R (version 4.0.2) [25] using the glmnet package [26]. For each round, 5 univariate linear regressions were created using all extracted features as the independent variables, and each of the 5 performance metrics (NDCG@K, P@K, RBP, bpref, and MAP) as the dependent variable. Coefficients and standard errors were calculated for each feature, and p-values were extracted for each feature coefficient, with significance defined as $p < 0.05$. Features that met the threshold for significance in the univariate regression were subsequently input into a multivariate linear regression. Overall, positive coefficients were interpreted to be associated with higher performance. Therefore, features which remained significant after both univariate and multivariate regression were likely associated with high performance in the TREC-COVID challenge.

Table 2
Taxonomy Features Extracted from Run Reports. Features extracted from run reports from Round 2 and Round 5 are listed. Features were either extracted as a binary feature or a categorical feature (in which a short description specifying the feature was provided). Prior to input into univariate and multivariate analysis, categorical features were one-hot encoded, or converted into binary variables over multiple columns.

Feature Names	Input Type	No. of Binary Values > 0 (%)	
		Round 2	Round 5
Title + Abstract Index Used	Binary	42 (38.2%)	78 (70.3%)
Paragraph Index Used	Binary	29 (26.4%)	27 (24.3%)
Full-Text Index Used	Binary	92 (83.4%)	70 (63.1%)
Filtered Dataset by Time	Binary	8 (7.3%)	5 (4.5%)
Query	Binary	109 (99.1%)	111 (100.0%)
Question	Binary	56 (50.9%)	60 (54.1%)
Narrative	Binary	28 (25.5%)	32 (28.8%)
Input Query Preprocessing	Binary	29 (26.4%)	42 (37.8%)
Term Expansion	Binary	37 (33.6%)	23 (20.7%)
Manual Review	Binary	21 (19.1%)	4 (3.6%)
Weighted System	Categorical[a]	N/A	N/A
Neural	Binary	49 (44.5%)	67 (60.4%)
Dataset for Fine-Tuning	Categorical[a]	N/A	N/A
Other Technique	Categorical[a]	N/A	N/A
Re-ranking	Binary	44 (40.0%)	68 (61.3%)
Fusion Technique Used	Binary	31 (28.2%)	32 (28.8%)
Pseudo-relevance feedback	Binary	44 (40.0%)	24 (21.6%)
Feedback from Judged Manuscripts	Binary	23 (20.9%)	59 (53.2%)

[a] Categorical features were features that were extracted as a description rather than a binary variable.

3. Results

Round 2 consisted of 136 runs submitted by 51 teams (with a permitted maximum of 3 runs submitted per team), and Round 5 consisted of 126 runs submitted by 28 teams (with a permitted maximum of 8 runs submitted per team). The topics in Round 2 included 30 previous (i.e., from Round 1) topics with relevance judgments, and 5 new topics. The topics in Round 5 included 45 previous (i.e., from Rounds 1–4) topics and 5 new topics. Overall, 110 runs from 42 teams met inclusion criteria in Round 2 and 111 runs from 23 teams met inclusion criteria in Round 5. The proportion of manual (defined as involving human intervention), feedback (defined as using judgments from prior rounds), and automatic (defined as neither feedback nor manual) runs varied between runs. In Round 2, the majority of the runs were categorized as automatic runs; in round 5, the majority of the runs were characterized as feedback runs. These findings are summarized in Table 3.

Significant features for the 5 univariate regressions each for Round 2 and Round 5 are shown in Fig. 2 and varied depending on the performance metric used. In Round 2, query term expansion (n = 37 runs), fine-tuning of ranking systems on MS-MARCO (n = 18 runs), Round 1 judgments (n = 9 runs), or document vectors formed by the CORD-19 dataset (n = 9 runs) were associated with higher performance across most, if not all, performance metrics. Use of ReQ-Rec (n = 3 runs submitted by 1 team), and narrative text in the query (n = 28 runs) were associated with decreased performance across the majority of performance metrics. In Round 5, use of the question text in the query (n = 32 runs) and TF-IDF vectors were associated with increased performance (n = 14 runs), whereas the use of neural networks, narrative text in the query (n = 67 runs), and proximity score (n = 2 runs) were associated with decreased performance across all performance metrics.

Significant features from multivariate regressions on the 5 different performance metrics in Rounds 2 and 5 are shown in Fig. 3. After features found to be significant on univariate regression were input into a multivariate regression, the following features remained significantly associated with increased performance in Round 2 with the majority of performance metrics: term expansion (n = 37), ranking system fine-tuning on CORD-19 vectors (n = 9), MS-MARCO (n = 18), and Round 1 judgments (n = 9). Using ReQ-Rec (n = 3) remained significantly associated with decreased performance. In Round 5, using the question text to formulate the query (n = 60) and TF-IDF vector weighting (n = 14) were associated with increased performance, while a custom proximity score (n = 2) as a scoring function was associated with decreased performance. As seen in Round 2, using feedback in Round 5 (n = 59) was associated with increased performance when runs were evaluated on RBP.

4. Discussion

This study aimed to develop a taxonomy of features to evaluate techniques associated with higher performance in runs submitted to Rounds 2 and 5 of the TREC-COVID Challenge. The key findings were: (1) fine-tuning ranking systems using relevance judgments resulted in significant improvement in performance, particularly in Round 2, and (2) query formulation is an important component of successful search.

Our first key finding was that fine-tuning ranking systems using relevance judgments resulted in significant improvement in performance, particularly in Round 2. Unlike previous TREC challenges, rapid turnout of relevance judgments over multiple rounds resulted in opportunities to fine-tune ranking systems for improved performance. Many of the runs labelled as feedback runs (n = 41 in Round 2 and n = 65 in round 5) employed fine-tuning, though a small portion specifically used the relevance judgments specifically in fine-tuning their ranking systems. Other teams who fine-tuned on similar datasets, including the vectorized CORD-19 dataset (represented as Dataset.for-FineTuning_CORD-19) and MS-MARCO also achieved comparable levels of improvement when compared to systems that did not use fine-tuning on these specific datasets. The noted improvement of fine-tuning on an annotated dataset has been reported in other TREC challenges, most notably the usage of MS-MARCO by Nogueira et al to refine a neural network system that vastly outperformed other runs in the TREC CAR challenge [27]. Interestingly, the benefits of fine-tuning systems did not persist into Round 5 (with the exception of evaluation on RBP) despite more prevalent use of neural systems and feedback runs.

Since the TREC-COVID Challenge brought together a mix of research teams with varying experience in IR challenge evaluations, along with the short time between rounds (1–3 weeks), the absence of significance with fine-tuning on previous round judgments may be explained by implementation differences between teams, as many teams implemented variations of the popular sequence of an initial weighted system (most commonly BM25) followed by a neural re-ranker (i.e., BERT with or without fine-tuning on MS-MARCO or previous relevance judgments) [28,29]. However, since we one-hot encoded other techniques, our linear regressions may have overrepresented individual techniques that few teams used (including ReQ-ReC in Round 2, and proximity scoring in Round 5). Future work may be needed to validate the performance of other techniques compared with the standard weighted and neural pipelines. Furthermore, since we chose to set neural networks as a binary variable, there may be opportunities to explore how different architectures influence performance in the TREC-COVID challenge.

Our second key finding was that query formulation was an important component of successful search. While most teams used the query and question fields in formulating an input query, several teams (n = 28 and 32 in Rounds 2 and 5 respectively) chose to use the narrative portion of the topic, which was associated with decreased performance in both rounds. Because the narrative contained freehand descriptions qualifying each topic, these descriptive fields were noisy. For example, topics 33 and 34 contained the phrase "excluding…," with subsequent wording describing what not to search. Furthermore, vocabulary used in the topics designed by the TREC-COVID organizers were not consistently used in manuscripts included in the CORD-19 dataset (i.e. differences in how COVID-19 was named: SARS-CoV 2, coronavirus) and may have adversely affected search performance for those who did not expand their queries to include such terms.

In fact, many of the successful runs from teams that used baseline runs from Anserini [30] employed a query preprocessing tool (which will subsequently be referred to as "Udel") produced by University of Delaware. The "Udel" method used SciSpacy [31] to lemmatize and remove non-stop words from the combined query, question, and narrative fields for each topic. Runs generated by Anserini comparing standard addition of various topic fields with and without the "Udel" method consistently showed improvements in retrieved relevant documents no matter what topics were used to construct the query and which indices were used [32]. This approach was taken further either manually by certain teams (i.e., OHSU) or automatically, as seen in approaches in initial iterations of Covidex [33], a consistently high-performing neural re-ranking methodology that was an early adopter of the "Udel" preprocessing method. In fact, adapting the queries to better represent document representations, or minimize query-document mismatch, has long been researched and includes work using relevance judgments [34]

Table 3
Distribution of Included Runs in Rounds 2 and 5. The number of included runs and the number of participating teams is included in this table. Runs were either sub-categorized as feedback (defined as using judgments from prior rounds), manual (involving human intervention), and automatic (neither manual nor feedback) runs.

	Round 2	Round 5
Number of Runs Included	110	111
Feedback	41	65
Automatic	49	43
Manual	20	3
Number of Teams	42	23

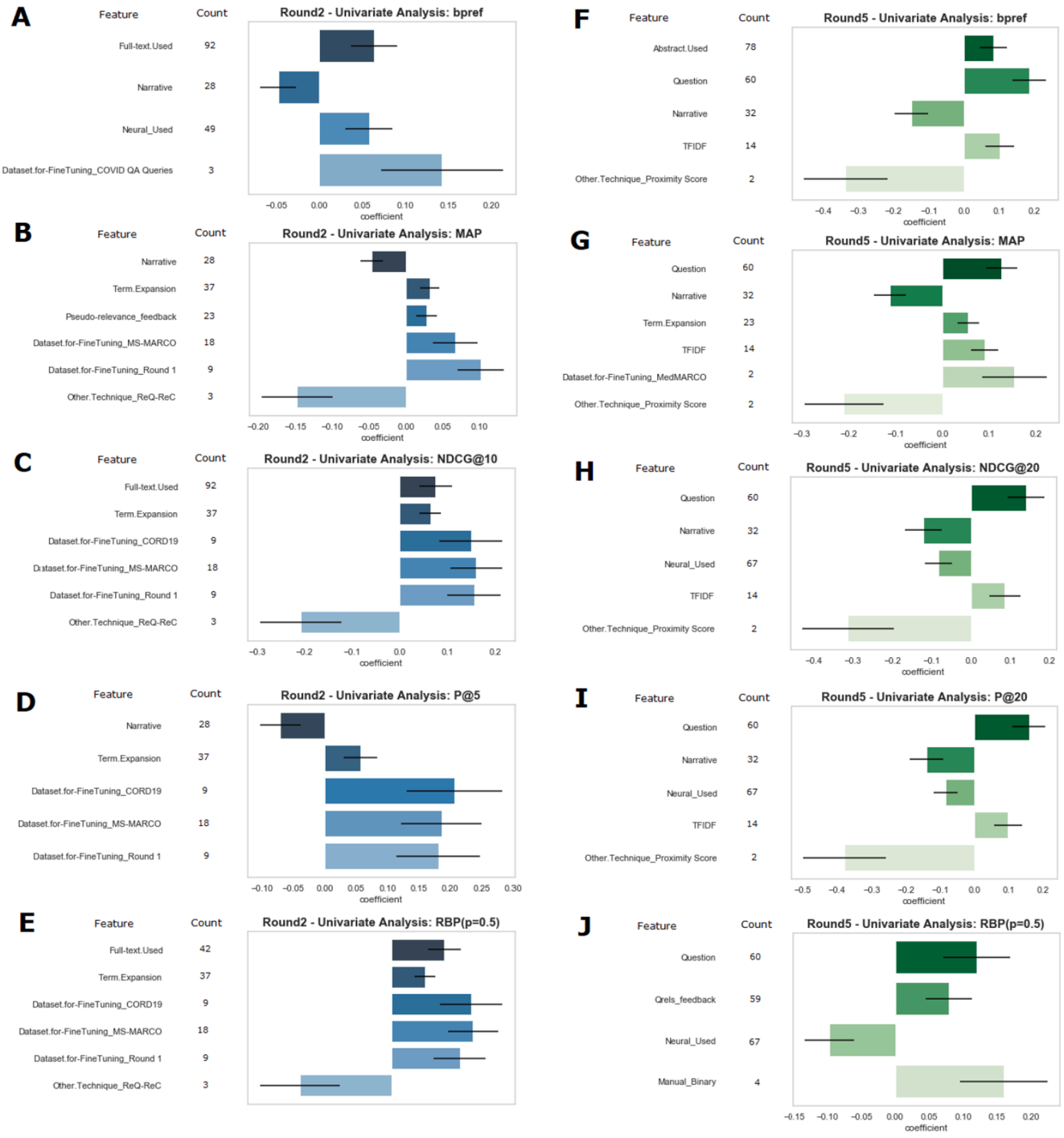


Fig. 2. Significant Features after Univariate Regression Analysis in Rounds 2 and 5. Univariate analysis was performed on features extracted from reports from Rounds 2 and 5. Features that were significant after input into a univariate linear regression are shown for the following performance metrics from Rounds 2 and 5 respectively: binary preference (**A** and **F**), mean average precision (**B** and **G**), normalized distributive cumulative gain (**C** and **H**), precision @ k documents (**D** and **I**), and rank-based precision (**E** and **J**). The count, or number of times that the feature occurred in our extracted dataset, is displayed adjacent to the feature name. These significant features were subsequently input into a multivariate regression to determine which features were independently associated with performance.

and query expansion [35]. Novel methods have focused on reverse: adjusting documentation representations to better represent queries - for example, Doc2Query [36] was employed most commonly in Round 5 by one team, though this technique was not shown to be significantly associated with high performance in our study. However, the team that incorporated this technique submitted runs that were widely variable in performance, and may have used other features not found to be significant in our taxonomy. The importance of defining relevant terms in queries has also been reflected in human user studies, in which previous work by Hersh et. al has demonstrated the importance of search ability and domain knowledge of the user in a biomedical search task [10].

5. Limitations

This study had several limitations that future work could address. First, the instructions for describing methodologies in the run reports varied in detail. As such, the data used for this study were only as complete as what was provided in the reports. This not only presented a

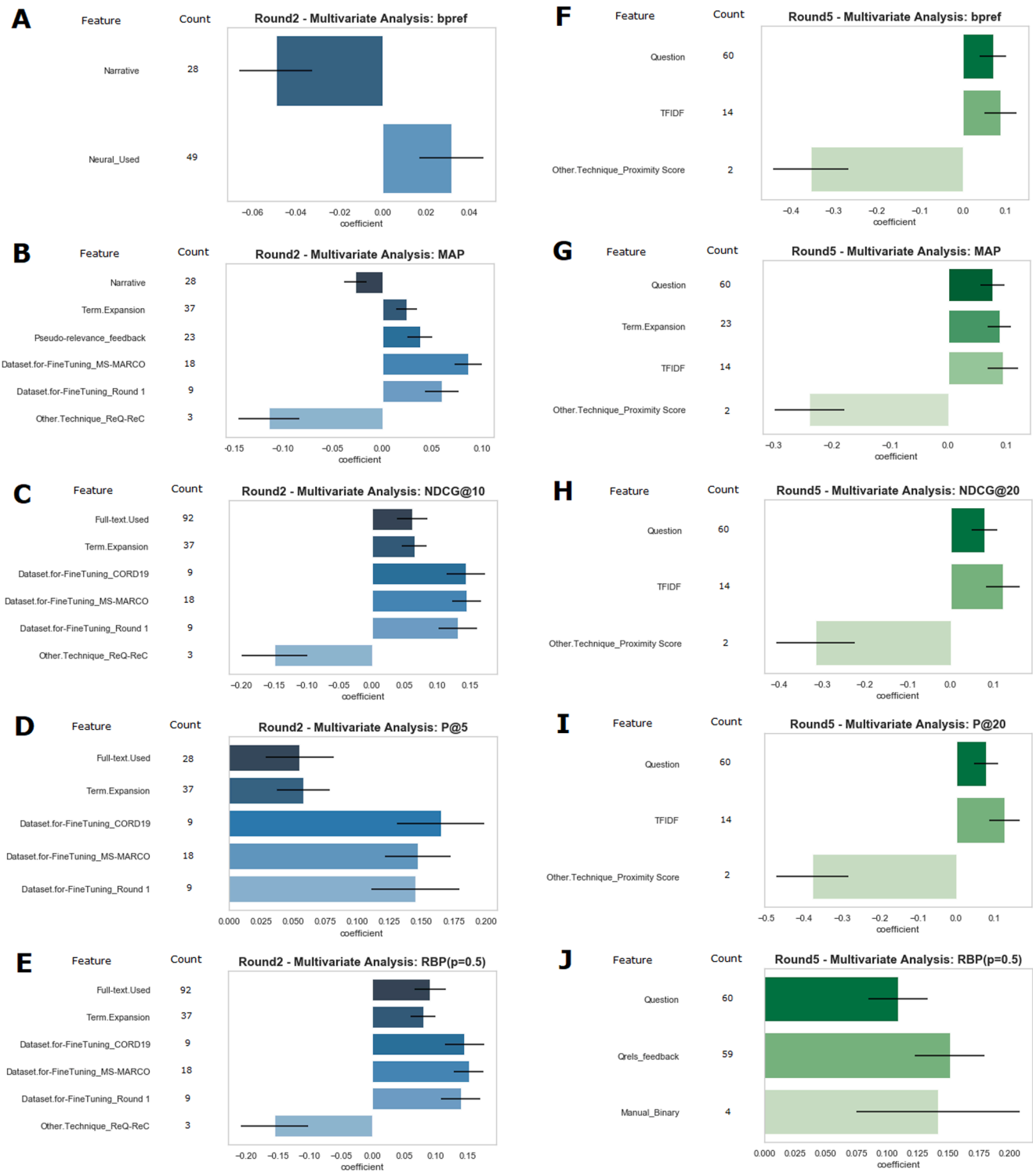


Fig. 3. Significant Features after Multivariate Regression Analysis in Rounds 2 and 5. Features that were found to be significant in univariate regression were further input into a second, multivariate regression. Significant features were reported for the following performance metrics in Rounds 2 and 5 respectively: binary preference (**A** and **F**), mean average precision (**B** and **G**), normalized distributive cumulative gain (**C** and **H**), precision @ k documents (**D** and **I**), and rank-based precision (**E** and **J**). Depending on the coefficients, these features were concluded to be significantly associated with increased or decreased performance in the TREC-COVID challenge.

challenge to building our taxonomy, but also meant that important features may not have been (and likely were not) reported. In the future, teams should document methodologies that promote reproducibility or publish their results in reports as is done in the regular TREC challenges.

Second, it was difficult to capture run-specific differences between runs submitted by the same team, as team-specific features were often not provided. This had important implications in runs submitted in Round 5, where teams were allowed to submit up to 8 runs. While many runs submitted from the same team were largely similar (and often performed similarly), our methodology was not well-suited to capture nuances such as hyperparameter tuning that were likely small adjustments to otherwise similar methods and pipelines. This may have been

pronounced in Round 5, where there appeared to be a convergence in methodology; that is, many teams created neural re-ranking pipelines using similar models but vastly different hyperparameters. However, we sought to characterize runs broadly, rather than capture each individual technique and adjustment in each run, since features built around individual techniques were both subject to bias and difficult to account for. However, to find a balance between granularity vs. breadth of techniques, we attempted to take into account differences between runs (even from the same team) using a one-hot encoded column of other techniques that we thought were unique enough to warrant specific inclusion. Future directions for this work may include identifying how to best capture adjustments between runs using similar techniques that result in different performances.

Third, our study was retrospective and limited in scope. While our key findings regarding the performance improvements derived from input query preprocessing (whether that is a combination of query, question, and/or narrative fields) and relevance judgments are well-documented in IR, it is unclear to what extent these findings are generalizable to other test collections. Specifically, our study aimed to categorize features associated with high retrieval performance on the CORD-19 dataset and may have overfitted to certain techniques, particularly with teams that used unique methodologies (i.e. associated a feature with significantly low or high performance despite a low number of teams employing this feature, such as ReQ-ReC [20] or Proximity score). These techniques have demonstrated success in prior search tasks, but may have performed poorly in the TREC-COVID challenge in part due to limited use and challenges with implementation, which depend on team experience. While we feel our findings have broad implications in ad-hoc retrieval, future work will be needed to validate our findings and prospectively evaluate less-commonly used techniques across different developers and users.

6. Conclusion

Using multivariate regression analysis, we developed and evaluated a taxonomy of features IR systems associated with high performance in the TREC-COVID Challenge. While our multivariate analysis demonstrates the utility of relevance feedback and the need for well-defined queries, it remains unclear which broad methodologies are associated with high performance in the TREC-COVID test collection. While our study has limitations in generating specific, prospective generalizations about IR systems and techniques, our work broadly showcases general techniques that may be useful in building search systems for COVID-19, and serves as a springboard for future work on TREC-COVID and related test collections.

7. Financial support

None.

CRediT authorship contribution statement

Jimmy Chen: Conceptualization, Data curation, Methodology, Software, Formal analysis, Resources, Investigation, Writing - original draft, Writing - review & editing, Visualization. **William R. Hersh:** Conceptualization, Methodology, Resources, Supervision, Validation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (accessed September 8, 2020).
[2] A. Palayew, O. Norgaard, K. Safreed-Harmon, T.H. Andersen, L.N. Rasmussen, J. V. Lazarus, Pandemic publishing poses a new COVID-19 challenge, Nat. Hum. Behav. 4 (7) (2020) 666–669, https://doi.org/10.1038/s41562-020-0911-0.
[3] L.L. Wang, K. Lo, Y. Chandrasekhar, et al., CORD-19: The COVID-19 Open Research Dataset, ArXiv200410706 Cs. http://arxiv.org/abs/2004.10706, Published online July 10, 2020 (accessed September 8, 2020).
[4] W. Hersh, Information retrieval: a biomedical and health perspective, fourth ed., 2020. doi: http://dx.doi.10.1007/978-3-030-47686-1.
[5] E.M. Voorhees, D. Harman, TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, Cambridge, MA (Digital Libraries and Electronic Publishing series), 2005.
[6] TREC-COVID Home. https://ir.nist.gov/covidSubmit/ (accessed October 13, 2020).
[7] K. Roberts, T. Alam, S. Bedrick, et al., TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19, J. Am. Med. Inform. Assoc. (2020), https://doi.org/10.1093/jamia/ocaa091.
[8] E. Voorhees, T. Alam, S. Bedrick, et al., TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection, ArXiv200504474 Cs. http://arxiv.org/abs/2005.04474, Published online May 9, 2020 (accessed September 8, 2020).
[9] K. Roberts, T. Alam, S. Bedrick, et al., Searching for answers in a pandemic: an overview of TREC-COVID submitted to journal of biomedical informatics COVID-19 special issue, J. Biomed. Inform. COVID-19 Special Issue (2020).
[10] W.R. Hersh, M.K. Crabtree, D.H. Hickam, et al., Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions, J. Am. Med. Inform. Assoc. 9 (3) (2002) 283–293, https://doi.org/10.1197/jamia.m0996.
[11] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, W. Hersh, State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track, Inf. Retr. J. 19 (1) (2016) 113–148, https://doi.org/10.1007/s10791-015-9259-x.
[12] H.K. Rekapalli, A.M. Cohen, W.R. Hersh, A comparative analysis of retrieval features used in the TREC 2006 Genomics Track passage retrieval task, in: AMIA Annu Symp Proc AMIA Symp., 2007, pp. 620–624.
[13] G.V. Cormack, M.R. Grossman, Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review, ArXiv150406868 Cs. http://arxiv.org/abs/1504.06868, Published online April 26, 2015 (accessed October 14, 2020).
[14] M.M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, P. Williams, Okapi at TREC-5, in: The Fifth Text REtrieval Conference (TREC-5). The Fifth Text REtrieval Conference (TREC–5), NIST, Gaithersburg, MD, (1997) 143–165, https://www.microsoft.com/en-us/research/publication/okapi-at-trec-5/ (accessed October 13, 2020).
[15] Data mining, in: A. Rajaraman, J.D. Ullman (Eds.), Mining of Massive Datasets. Cambridge University Press, 2011, 1–17. doi: http://dx.doi.10.1017/CBO9781139058452.002.
[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv181004805 Cs. http://arxiv.org/abs/1810.04805, Published online May 24, 2019 (accessed October 14, 2020).
[17] I. Beltagy, K. Lo, A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text, ArXiv190310676 Cs. http://arxiv.org/abs/1903.10676, Published online September 10, 2019 (accessed October 14, 2020).
[18] R. Tang, R. Nogueira, E. Zhang, et al., Rapidly Bootstrapping a Question Answering Dataset for COVID-19, ArXiv200411339 Cs. http://arxiv.org/abs/2004.11339, Published online April 23, 2020 (accessed May 4, 2020).
[19] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, X. Cheng, DeepRank: a new deep architecture for relevance ranking in information retrieval, in: Proc 2017 ACM Conf Inf Knowl Manag., 2017, pp. 257–266.
[20] C. Li, Y. Wang, P. Resnick, Q. Mei, ReQ-ReC: High recall retrieval with query pooling and interactive classification, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14. Association for Computing Machinery, 2014, pp. 163–172.
[21] P. Bajaj, D. Campos, N. Craswell, et al., MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, ArXiv161109268 Cs. http://arxiv.org/abs/1611.09268, Published online October 31, 2018 (accessed October 11, 2020).
[22] G.V. Cormack, C.L.A. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09. Association for Computing Machinery, 2009, pp. 758–759.
[23] J.A. Shaw, E.A. Fox, Combination of multiple searches, in: The Second Text Retrieval Conference, TREC-2, (1994) 243–252.
[24] S. MacAvaney, A. Cohan, N. Goharian, SLEDGE: A Simple Yet Effective Baseline for Coronavirus Scientific Knowledge Search, ArXiv200502365 Cs. http://arxiv.org/abs/2005.02365, Published online May 6, 2020 (accessed May 7, 2020).
[25] R Core Team. R, A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2020 https://www.R-project.org/.
[26] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22.

[27] R. Nogueira, K. Cho, Passage Re-ranking with BERT, ArXiv190104085 Cs. http://arxiv.org/abs/1901.04085, Published online April 14, 2020 (accessed May 4, 2020).
[28] B. Mitra, N. Craswell, An introduction to neural information retrieval, Found. Trends Inf. Retr. 13 (2018) 1–126.
[29] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, W.B. Croft, Neural Ranking Models with Weak Supervision, ArXiv170408803 Cs. http://arxiv.org/abs/1704.08803, Published online May 29, 2017 (accessed October 13, 2020).
[30] P. Yang, H. Fang, Lin J. Anserini, Enabling the use of Lucene for information retrieval research, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17. Association for Computing Machinery, 2017, pp. 1253–1256.
[31] M. Neumann, D. King, I. Beltagy, A.W. ScispaCy, Fast and robust models for biomedical natural language processing, in: Proc 18th BioNLP Workshop Shar Task, 2019, pp. 319–327.
[32] A. Castorini, Lucene toolkit for replicable information retrieval research. GitHub. https://github.com/castorini/anserini (accessed October 13, 2020).
[33] E. Zhang, N. Gupta, R. Tang, et al., Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset, ArXiv200707846 Cs. http://arxiv.org/abs/2007.07846, Published online July 14, 2020 (accessed October 11, 2020).
[34] J.J. Rocchio, Relevance feedback in information retrieval, in: G. Salton (Ed.), The Smart Retrieval System - Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, 1971, pp. 313–323.
[35] E.M. Voorhees, Query expansion using lexical-semantic relations, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '94, Springer-Verlag, 1994, pp. 61–69.
[36] R. Nogueira, W. Yang, J. Lin, K. Cho, Document Expansion by Query Prediction, ArXiv190408375 Cs. http://arxiv.org/abs/1904.08375, Published online September 24, 2019 (accessed September 20, 2020).