https://doi.org/10.1007/s42979-020-0069-x

ORIGINAL RESEARCH

Feature Weighting in Finding Feedback Documents for Query Expansion in Biomedical Document Retrieval

Jainisha Sankhavara[1]

Received: 2 October 2019 / Accepted: 20 January 2020 / Published online: 3 March 2020
© Springer Nature Singapore Pte Ltd 2020

Abstract
Finding good feedback documents for query expansion is a well-known problem in the field of information retrieval. This paper describes a novel approach for finding relevant documents for feedback in query expansion for biomedical document retrieval. The proposed approach relies on a small amount of human intervention to find good feedback documents and tries to learn the relation between query and documents in terms of usefullness of document for query expansion. This proposed approach uses an NLP-based feature weighting technique with classification and clustering method on the documents and identifies relevant documents for feedback. The documents are represented using term frequency and inverse document frequency (TF–IDF) features and these features are weighted according to the type of query and type of the terms. The experiments performed on CDS 2014, 2015 and 2016 datasets show that the feature weighting in finding feedback documents for query expansion approach gives good results as compared to the results of pseudo-relevance feedback, relevance feedback and the results of TF–IDF features without weighting.

Keywords Biomedical document retrieval · Feedback document discovery · Query expansion

Introduction

The huge amount of biomedical literature, available nowadays, makes searching as well as information extraction difficult. Biomedical information retrieval is a new field of research that can be helpful to solve the real problems in the field of biology and medicine. Automated medical systems such as biomedical document retrieval, question answering system, biomedical document summarization system, medical data visualization, medical history extractors would require a good amount of information retrieval and extraction research as well as natural language processing for biomedical data.

Biomedical document retrieval systems focus on finding relevant documents for user's query. These systems do query document matching based on terms present in them. Such a document retrieval system usually suffers from term mismatch problem which is due to multiple synonyms available in biomedical terminology. Also the term abbreviations and term inconsistency obstructs the retrieval system in finding true relevant documents. To overcome the problems of term mismatching, query reformulation techniques are used in the retrieval systems. Query reformulation is a process to reform the user query in a way to get better matching of relevant documents. Query reformulation process includes adding terms and/or removing terms and/or re-weighting the terms. When new terms are added to the query with some weights, it is known as query expansion.

Query expansion is a type of query reformulation which expands the query with the other related terms. Automatic query expansion techniques use some feedback documents from which the expansion technique selects terms for expansion of the query. It has been seen in the literature that automatic query expansion improves the system performance as compared to no expansion to the queries [6]. There are mainly two techniques to automatic query expansion, that is, relevance feedback (RF) and pseudo-relevance feedback (PRF). Relevance feedback-based techniques use only relevant documents as feedback documents from the

This article is part of the topical collection "Forum for Information Retrieval Evaluation" guest edited by Mandar Mitra and Prasenjit Majumder.

✉ Jainisha Sankhavara
jainishasankhavara@gmail.com

[1] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India

top retrieved documents while pseudo-relevance feedback-based techniques use all the top retrieved documents as feedback documents. Thus, relevance feedback techniques require human judgement to get relevant documents from top retrieved documents and this includes the cost in the system to get human judgements. The pseudo-relevance feedback-based techniques do not require any human judgement to identify relevant feedback documents. It assumes that all the top retrieved documents are relevant and use them as feedback documents. It has been seen that RF-based techniques outperform PRF-based techniques for biomedical document retrieval [14] but includes the cost of human judgements. But PRF-based techniques are fully automated and they do not require any external expensive inputs in the retrieval process and still giving a good improvement.

Query expansion methods largely rely on feedback documents and feedback terms. The feedback documents used by pseudo-relevance feedback-based query expansion methods might not be all relevant. Pseudo-relevance feedback documents contain both, actually relevant documents as well as not relevant documents. Actual relevant documents contain relevant terms to the query and those terms are useful to fetch more relevant documents when added to the query but non-relevant documents contain noisy terms which are not relevant to the query and choosing those noisy terms for expansion of the query will not help to find more relevant documents but it will fetch irrelevant documents and degrade the retrieval system performance instead. Thus, choosing proper feedback documents and feedback terms is crucial in the query expansion process.

When relevance feedback is costly for a large number of top retrieved documents, the feedback document discovery process tries to find good feedback documents automatically using human relevance judgement for a small set of documents from top retrieved documents instead of using human judgement for a large set of top retrieved documents. The feedback document discovery process tries to learn to identify good feedback documents using a little human intervention.

In this paper, we focus on feature weighting in the process of finding good feedback documents. Using the knowledge of some of the documents being relevant, the system tries to learn a classifier model to separate relevant documents from non-relevant documents for future feedback. This classification module considers queries and documents in the form of term frequency and inverse document frequency (TF–IDF) features of the terms contained in them. Here, we introduce a feature weighting scheme for term features in the document classification process based on the semantic type of the term.

The remainder of this paper is organized as follows: the next section includes related work followed by which feature weighting scheme for feedback document discovery process is explained. Experiments and results are presented before the final section. Conclusion is given in the final section.

Related Work

Query reformulation based on relevance feedback was studied by Salton and Buckley [13]. A recent survey on query expansion (QE) for information retrieval (IR) highlights the current progress, emerging research directions, potential new research areas and novel classification of state-of-the-art approaches in the field of query expansion [3]. Various feedback models [1, 5] show effectiveness of query reformulation in various fields. More recently, query expansion using local and global context analyses is studied by Xu and Croft [19]. For clinical decision support retrieval, an approach for pseudo-relevance feedback based on proximity information has been proposed by Pan et al. [10].

An analysis of state-of-the-art in biomedical literature retrieval shows that the pseudo-relevant feedback-based query expansion techniques are working best for biomedical information retrieval systems [12]. Query expansion method using UMLS metathesaurus for biomedical IR system was proposed by Aronson and Rindflesch [2]. Query expansion methods using medical ontologies have been proposed in the literature [7, 8]. Query expansion using external collections in biomedical IR was carried out by Oh and Jung [9]. The fusion of automatic and manual feedback for query expansion in biomedical information retrieval experiments shows that manual feedback helps to improve the performance of biomedical IR systems [16]. Stokes et al. [17] discussed the success factors for effective query expansion with respect to various sources of term expansion such as corpus-based co-occurrence statistics, pseudo-relevance feedback methods, and domain-specific and domain-independent ontologies. The cluster-based external expansion model proposed by Oh and Jung [9] incorporates the structure of external collections at estimating document models for feedback in query expansion.

Our approach for query expansion is based on feature weighting scheme for feedback document identification based on concept types and document features. This approach is a combination of human relevance feedback and blind relevance feedback. It tries to learn relevance from available human judgements and then uses it for automatic query expansion. This learning method is a domain-specific method that can be applied to other domains also.

Feature Weighting in Finding Feedback Documents for Query Expansion

Feedback document discovery-based query expansion approach for biomedical document retrieval was first described by Sankhavara and Majumder [15]. The documents are represented by TF–IDF of the terms present in the documents that means the words are features of the documents with weights as TF–IDF. The two algorithms based on classification and clustering are used to find pseudo-judgements and to predict the feedback documents. Feedback document discovery algorithm uses only predicted relevant documents as feedback documents in query expansion.

For classification and clustering in feedback document discovery, we propose an NLP-based feature weighting technique. We have used Clinical Named Entity Recognition system (CliNER) [4] which is an open-source natural language processing system for named entity recognition in the clinical text of electronic health records. CliNER is implemented as a sequence classification task, where every token is predicted using inside–outside-beginning (IOB) tagging style [11] as either problem, test, treatment, or none. We have trained it on i2b2 2010 dataset [18] which includes discharge summaries from Partners Health-Care, from Beth Israel Deaconess Medical Center and from University of Pittsburgh Medical Center. These discharge summaries are manually annotated for concept, assertion, and relation information. The model trained on i2b2 dataset is now used on the documents to identify medical entities of type ‘problem’, ‘test’ and ‘treatment’ from them. The TF–IDF word feature documents are now also weighted by the type of words. The two proposed approaches for feature weighting on these entities are as follows:

FW1: The first approach does feature weighting of medical concepts based on the type of the query. There are three types of queries in the dataset: ‘diagnosis’, ‘test’ and ‘treatment’. For queries of a type, only features of the entities of the same type are weighted. The feature of term t is determined as follows:

$$f(t) = \begin{cases} w \times \mathrm{TF} \times \mathrm{IDF} & \text{if } q \text{ is of type diagnosis,} \\ & t \text{ is a problem type term} \\ w \times \mathrm{TF} \times \mathrm{IDF} & \text{if } q \text{ is of type test,} \\ & t \text{ is a test type term} \\ w \times \mathrm{TF} \times \mathrm{IDF} & \text{if } q \text{ is of type treatment,} \\ & t \text{ is a treatment type term} \\ \mathrm{TF} \times \mathrm{IDF} & \text{otherwise} \end{cases}.$$

For ‘diagnosis’ type of queries, only ‘problem’ type of entities is weighted by weight w, for ‘test’ type of queries, only ‘test’ type of entities is weighted by weight w and for ‘treatment’ type of queries, only ‘treatment’ type of entities is weighted by weight w, thus giving importance to the query type similar entities while learning.

FW2: The second approach does feature weighting of medical concepts irrespective of the type of query. For all the queries of type ‘diagnosis’, ‘test’ and ‘treatment’, all the entities of types ‘problem’, ‘test’, and ‘treatment’ are weighted by weight w. The feature of term t is determined as follows:

$$f(t) = \begin{cases} w \times \mathrm{TF} \times \mathrm{IDF} & \text{if } t \text{ is either problem or} \\ & \text{test or treatment term} \\ \mathrm{TF} \times \mathrm{IDF} & \text{otherwise} \end{cases}.$$

Experiments and Results

The experiments are performed on TREC Clinical Decision Support (CDS) track[1] dataset. Three-year data have been used in the experiments, i.e. CDS 2014, CDS 2015 and CDS 2016 dataset. The dataset contains narrations of patients’ medical case reports as queries. The document collection is an open-access subset of PubMed Central articles. Each year’s dataset contains 30 queries for which relevant documents from the document collections need to be extracted. These queries are of three types: ’diagnosis’, ’test’ and ’treatment’. The queries describe the patient’s medical history, symptoms, condition, test results and other related medical information. The retrieved documents should suggest the diagnosis of the patient or test to perform or treatment to the patient. The data statistics are given in Table 1.

The query expansion considers the top N retrieved documents for feedback. Here, we have considered the top 250 documents, from which the set of top 50 documents are used as training, i.e. human judgements for top 50 documents are used in training and the rest of 200 documents are taken for testing data. The relevance is predicted for those 200 documents and only relevant predicted documents are then used for feedback. The number of training documents is empirically chosen as 50. These training documents are marked as either relevant or partially relevant or not relevant. Therefore, there will be three-class classification of test documents. In case of fewer training documents, there were cases when some classes are empty that means there were no documents in one of the three classes in the training data which may lead to misclassification. Therefore, an adequate amount of documents needs to be considered for training.

The result of relevance feedback using the top 50 documents is the baseline for other results. All the computed results are compared with the baseline.

For both the feature weighting techniques, the weight w is considered as 3 in the experiments. The comparison of results of two feature weighting techniques with the results of original queries without expansion, expansion with relevance feedback and expansion with feedback document discovery without feature weighting, i.e. only using TF–IDF for CDS 2014 dataset is given in Table 2 in terms of MAP (mean average precision) and infNDCG (normalized

[1] http://www.trec-cds.org/.

Table 1 CDS DATA statistics

Dataset	CDS 2014	CDS 2015	CDS 2016
#Documents	733,138	733,138	1,255,259
Collection size	47.2 GB	47.2 GB	87.8 GB
#Total terms	1,600,536,286	1,600,536,286	2,954,366,841
#Uniq. terms	3,689,317	3,689,317	4,564,612
#Topics	30	30	30
#Rel. docs/topic	112	150	182
Query forms	Description, summary	Description, summary	Note, description, summary
Avg. length of description (in words)	75.8	80.4	119.9
Avg. length of summary (in words)	24.6	20.4	33.3
Avg. length of note (in words)	–	–	239.4
Avg. doc length (in words)	2183	2183	2353

Table 2 MAP and infNDCG results on CDS 2014

CDS 2014	MAP			infNDCG		
	TF–IDF	FW1	FW2	TF–IDF	FW1	FW2
Original queries	0.1071			0.1836		
Queries + PRF$_{50}$	0.1502			0.2301		
Queries + RF$_{50}$	0.2768 (84%)			0.4186 (82%)		
Nearest neighbors	0.2761	0.2754	0.2747	0.4177	0.4161	0.4140
Nearest neighbors + *k*-means	0.2794	0.2778	0.2777	0.4220	0.4168	0.4195
Neural net	0.2790	0.2784	0.2787	0.4235	0.4243	0.4240
Neural net + *k*-means	0.2790 (86%)	0.2788	**0.2807 (87%)**	0.4218 (83%)	0.4225	**0.4269 (86%)**

Percentage improvement with respect to PRF are shown in brackets

The highest results are shown in bold

Table 3 MAP and infNDCG results on CDS 2015

CDS 2015	MAP			infNDCG		
	TF–IDF	FW1	FW2	TF–IDF	FW1	FW2
Original queries	0.1147			0.2115		
Queries + PRF$_{50}$	0.1693			0.2658		
Queries + RF$_{50}$	0.2283 (35%)			0.3478 (31%)		
Nearest neighbors	0.2234	0.2212	0.2234	0.3518	0.3480	0.3518
Nearest neighbors + *k*-means	0.2244	0.2214	0.2299	0.3541	0.3519	0.3506
Neural net	0.2295	0.2297	0.2284	0.3528	0.3514	0.3492
Neural net + *k*-means	0.2299 (36%)	**0.2302 (36%)**	0.2301	0.3529 (33%)	0.3525	**0.3526 (33%)**

Percentage improvement with respect to PRF are shown in brackets

The highest results are shown in bold

discounted cumulative gain) score. The same result comparisons for CDS 2015 and CDS 2016 dataset are given in Tables 3 and 4, respectively. The results of feature weighting techniques show improvement over original queries as well as relevance feedback.

Figures 1 and 2 show querywise difference in MAP and infNDCG, respectively, between Neural net + *k*-means using FW2 and queries + RF$_{50}$. Out of 30 queries of CDS 2014, the performance score of infNDCG degraded for 2 queries but improved for 7 queries.

Table 4 MAP and infNDCG results on CDS 2016

CDS 2016	MAP			infNDCG		
	TF–IDF	FW1	FW2	TF–IDF	FW1	FW2
Original queries	0.062			0.1710		
Queries + PRF$_{50}$	0.0800			0.2021		
Queries + RF$_{50}$	0.1456 (82%)			0.3094 (53%)		
Nearest neighbors	0.1456	0.1463	0.1458	0.3113	0.3124	0.3113
Nearest neighbors + *k*-means	0.1459	0.1470	0.1467	0.3127 (55%)	**0.3158 (56%)**	0.3139
Neural net	0.1460	0.1467	0.1463	0.3073	0.3136	0.3143
Neural net + *k*-means	0.1466 (83%)	**0.1471 (84%)**	0.1458	0.3100	0.3132	0.3124

Percentage improvement with respect to PRF are shown in brackets

The highest results are shown in bold

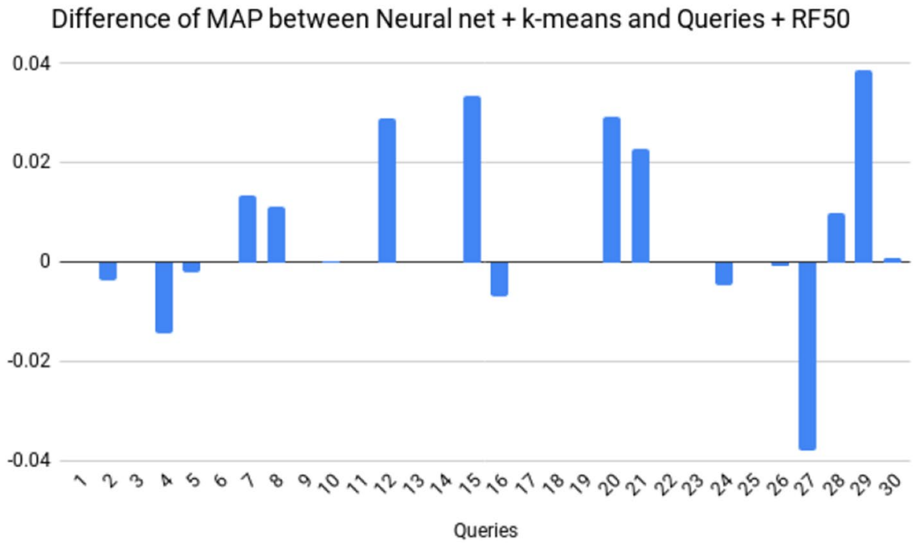

Fig. 1 Querywise difference graph of MAP between feedback document discovery with feature weighting and relevance feedback

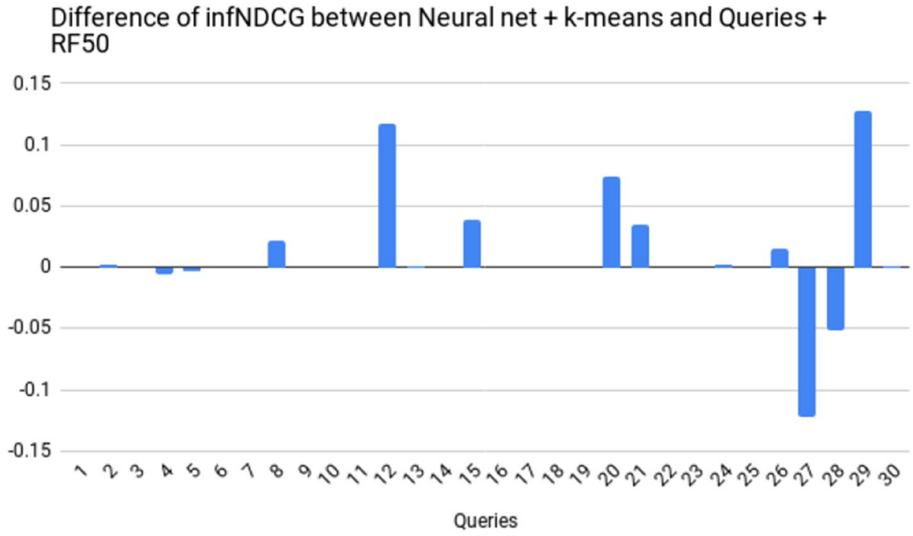

Fig. 2 Querywise difference graph of infNDCG between feedback document discovery with feature weighting and relevance feedback

Conclusion

This paper describes feature weighting for finding good feedback documents for query expansion in biomedical document retrieval. The two approaches are discussed for feature weighting based on type of the query and type of the term. The experiments performed on CDS 2014, CDS 2015 and CDS 2016 datasets show that the feature weighting approach gives improved results in terms of MAP and infNDCG as compared to baselines. In the future, the research can be carried out in the direction of automatically determining the weights of the medical terms.

Compliance with Ethical Standards

Conflict of interest Author has received research grant from TCS research scholar program.

Feedback Document Discovery: Finding Good Feedback Documents

First Algorithm

The first algorithm is based on classification. If we have human judgements available for some of the feedback documents, then it will serve as a training data for classification. The documents are represented as a collection of bag-of-words, the TF–IDF scores of the words represent features and human relevance scores provides the classes. Using this as a training data, we want to predict the relevance of other top retrieved feedback documents represented by TF–IDF scores of words.

```
Algo1 : classification
For each query Q
1. $D_N$ - set of N top retrieved documents $\{d_1, d_2, ..., d_N\}$
2. $D_k$ - set of k top retrieved documents for which human judgements are
   available $\{d_1, d_2, ..., d_k\}$
3. $D_l$ - set of l=N-k top retrieved documents for which human judgements are
   not available $\{d_{k+1}, d_{k+2}, ..., d_N\}$
4. $D_F$ - set of feedback documents
5. $D_F = \{d_i; relevance\ of\ d_i > 0, d_i \in D_k\}$
6. Train a classifier C on $D_k$ using relevance as a class label and generate model
   $M_c$
7. For each document $d_j$ in $D_l$, $k+1 \leq j \leq N$
8.         Predict the relevance $r_j$ of $d_j$ using trained model $M_c$
9.         If $r_j > 0$, then $D_F = D_F \cup \{d_j\}$
```

Second Algorithm

The second algorithm is an extension of first algorithm. The analysis of results of first algorithm shows that the feedback document set still contains some non-relevant docs and it is responsible for insignificant improvement. The detailed analysis is given in "Experiments and Results". This approach further removes non-relevant documents from relevant document class identified by classification approach. The idea is to perform clustering on the relevant identified documents with number of clusters as two: one from actually relevant documents and second from non-relevant documents. This approach relies on the statement that the relevant documents tend to cluster within the space. k-means clustering is used with $k = 2$. Since, the convergence of k-means clustering depends on the initial choice of cluster centroids, the initial cluster centroids are chosen as the average of relevant documents' vectors and the average of non-relevant documents' vectors from training data.

```
Algo2 : classification + clustering
For each query Q
1. $D_N$ - set of N top retrieved documents $\{d_1, d_2, ..., d_N\}$
2. $D_k$ - set of k top retrieved documents for which human judgements are
   available $\{d_1, d_2, ..., d_k\}$
3. $D_l$ - set of l=N-k top retrieved documents for which human judgements are
   not available $\{d_{k+1}, d_{k+2}, ..., d_N\}$
4. $D_F$ - set of feedback documents
5. $D_F = \{d_i; relevance\ of\ d_i > 0, d_i \in D_k\}$
6. Train a classifier C on $D_k$ using relevance as a class label and generate model
   $M_c$
7. $D_R = \phi, D_{NR} = \phi$
8. For each document $d_j$ in $D_l$, $k+1 \leq j \leq N$
9.         Predict the relevance $r_j$ of $d_j$ using trained model $M_c$
10.        If $r_j > 0$ then
                   $D_R = D_R \cup \{d_j\}$
11.        else
                   $D_{NR} = D_{NR} \cup \{d_j\}$
   \\ $D_R$ contains predicted relevant documents from $D_l$
12. Perform K-means clustering on $D_R$ with k=2 (relevant docs and non-relevant
    docs)
13. $D_F = D_F \cup \{documents\ from\ relevant\ docs\ cluster\}$
```

References

1. Allan J. Incremental relevance feedback for information filtering. In: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. pp. 270–278. ACM; 1996.
2. Aronson AR, Rindflesch TC. Query expansion using the UMLS metathesaurus. In: Proceedings of the AMIA annual fall symposium. p. 485. American Medical Informatics Association; 1997.
3. Azad HK, Deepak A. Query expansion techniques for information retrieval: a survey. Inf Process Manag. 2019;56(5):1698–735.
4. Boag W, Wacome K, Naumann T, Rumshisky A. Cliner: a lightweight tool for clinical named entity recognition. In: AMIA joint summits on clinical research informatics (poster); 2015.
5. Cao G, Nie JY, Gao J, Robertson S. Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. pp. 243–250. ACM; 2008.
6. Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. ACM Comput Surv (CSUR). 2012;44(1):1.

7. Díaz-Galiano MC, Martín-Valdivia MT, Ureña-López L. Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput Biol Med. 2009;39(4):396–403.
8. Dong L, Srimani PK, Wang JZ. Ontology graph based query expansion for biomedical information retrieval. In: Bioinformatics and biomedicine (BIBM), 2011 IEEE international conference on. pp. 488–493. IEEE; 2011.
9. Oh HS, Jung Y. Cluster-based query expansion using external collections in medical information retrieval. J Biomed Inform. 2015;58:70–9.
10. Pan M, Zhang Y, He T, Jiang X. An enhanced hal-based pseudo relevance feedback model in clinical decision support retrieval. In: International conference on intelligent computing. pp. 93–99. Springer; 2018.
11. Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. Natural language processing using very large corpora. New York: Springer; 1999. p. 157–76.
12. Roberts K, Simpson M, Demner-Fushman D, Voorhees E, Hersh W. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. Inf Retr J. 2016;19(1–2):113–48.
13. Salton G, Buckley C. Improving retrieval performance by relevance feedback. J Am Soc Inf Sci. 1990;41(4):288–97.
14. Sankhavara J. Biomedical document retrieval for clinical decision support system. In: Proceedings of ACL 2018, student research workshop. pp. 84–90; 2018.
15. Sankhavara J, Majumder P. Biomedical information retrieval. In: Working notes of FIRE 2017—Forum for information retrieval evaluation. pp. 154–157; 2017.
16. Sankhavara J, Thakrar F, Sarkar S, Majumder P. Fusing manual and machine feedback in biomedical domain. Tech. rep., Dhirubhai Ambani Inst of Information and Communication Technology. 2014.
17. Stokes N, Li Y, Cavedon L, Zobel J. Exploring criteria for successful query expansion in the genomic domain. Inf Retr. 2009;12(1):17–50.
18. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552–6.
19. Xu J, Croft WB. Quary expansion using local and global document analysis. In: ACM sigir forum. vol. 51, pp. 168–175. ACM; 2017.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.