Relevance Model Augmented with Inter-Collection Co-occurrence Statistics for Ad-hoc Retrieval in Precision Medicine

Sonal Mishra[1], Procheta Sen[2], Debasis Ganguly[3] and Prasenjit Majumder[1]

[1]*Dhirubhai Ambani Institute of Information and Communication and Technology, Gandhinagar, India*
[3]*University College London, London, United Kingdom*
[3]*University of Glasgow, Glasgow, United Kingdom*

Abstract

The ubiquitous growth of the volume of medical literature has necessitated the use of search systems to retrieve the relevant literature from a large collection of medical documents. Such a retrieval based approach has been shown to be useful in the precision medicine task, where instead of keyword-based queries of ad-hoc IR, the queries comprise of a disease name and the personal context associated with the disease, e.g., the gene sequence of the person or their demographics. The two fundamental sources of information useful for the precision medicine task are the medical literature (e.g. PubMed articles) and randomized control trials (RCTs). In this paper, we propose a symbiotic relevance feedback model that is able to utilize the information not only individually from each collection, but is also able to operate across collections improving the retrieval effectiveness of each collection by leveraging information from the top-retrieved documents of the other. Our experiments on the TREC 2019 and 2020 precision medicine track show that our proposed symbiotic relevance feedback model outperforms post-hoc combinations of standard relevance feedback models operating independently on respective collections.

Keywords

Relevance feedback, Smoothing across collections, Precision Medicine

1. Introduction

The volume of biomedical literature has increased steadily during the recent years. Availability of such large volumes of published research articles in PubMed[1] or findings of clinical trials[2] has not only provided opportunities for effective and efficient methods of information access (e.g., ad-hoc search on these collections), but also conduct a systematic review of different (often conflicting) findings and conclusions of these studies [1].

CIRCLE'22: Joint Conference of the Information Retrieval Communities in Europe, July 04–07, 2022, Samatan, Gers, France
sonal.mishra1297@gmail.com (S. Mishra); p.sen@ucl.ac.uk (P. Sen); debasis.ganguly@gla.ac.uk (D. Ganguly); p_majumder@daiict.ac.in (P. Majumder)
https://procheta.github.io/sprocheta/ (P. Sen); https://gdebasis.github.io/ (D. Ganguly); http://intranet.daiict.ac.in/~p_majumder/ (P. Majumder)
0000-0002-3814-5462 (P. Sen); 0000-0003-0050-7138 (D. Ganguly)

[1]https://pubmed.ncbi.nlm.nih.gov/
[2]https://clinicaltrials.gov/

Accessing relevant information from medical literature or clinical trials is beneficial for the task of precision medicine, where given the disease and a particular genetic variant of a patient, the task is to find information that is relevant (e.g., medicines, treatments - what works and what not) in the context of the specified gene variant [2, 3]. A major difference of 'precision medicine information retrieval' (PMIR) with respect to ad-hoc search over news or web documents is that the query terms corresponding to the disease and the gene represent two different aspects of information need, the information retrieved for each of which then needs to be collated in an effective manner.

Existing approaches in PMIR [4] has reported that the use of external collections or knowledge bases contribute to enrich the information needs in an effective manner with the use of synonyms and other related concepts. These enriched information needs eventually help improve the retrieval quality, because it is usually the case that the same concept is represented by the choice of different words across different documents. In fact, the best performing system of the TREC Precision Medicine track of 2019 made use of a domain-specific thesaurus[3] to expand gene variant names with additional information, such as aliases of the gene name. For instance, the term 'BRAF' may be expanded with additional terms such as 'B-RAF1', 'NS7', 'RAFB1' etc., which potentially contributes in bridging the vocabulary gap across retrieved documents.

A limitation of using external resources is that such resources may not be available for all application domains. Instead, in this paper, we propose a pseudo-relevance feedback (PRF) based approach which does not need to rely on the existence of external resources for effective retrieval. While PRF is a classic IR technique that has proven, over the years, to be effective in improving the overall retrieval effectiveness [5, 6, 7, 8], a straight-forward application of an existing PRF algorithm is less likely to be effective for PMIR. This is because in contrast to standard IR test collections comprised of documents from a single genre (e.g., web articles or news), the test collection for PMIR essentially comprises of documents from two different sources, namely the medical literature and the clinical trials, that are characteristically different from one another. While the former constitutes peer reviewed scientific articles, structured in a consistently formal manner and backed by empirical findings, the latter, in turn, reports initial findings of ongoing randomized control trials (RCTs) or even publishes eligibility criteria of upcoming RCTs.

In this paper, we propose a novel PRF solution that addresses this characteristically unique situation of retrieving from two different genres of document collections in PMIR. The proposed approach seeks to model the semantic relationship of terms both within and across the two collections, thus complementing each other, to improve the overall retrieval quality.

2. Related Work

Pseudo-relevance feedback (PRF). PRF essentially relies on term level manipulations, e.g., while Ogilvie et. al. [9] learn the appropriate number of feedback terms for query expansion, Cao et. al. [10] selectively use good feedback terms for query expansion. Traditional PRF methods, such as Okapi [11], the relevance model (RM) [5] and its variants [6, 12, 13], primarily rely on the set of top-retrieved documents for selecting candidate expansion terms. For term-level

[3] https://www.ncbi.nlm.nih.gov/gene

manipulations, researchers have also leveraged on semantic matching with embedded vectors to learn retrieval-specific semantic relationships from top documents retrieved with a large number of queries from a query log [14], or to combine the effects of global term semantics within the framework of RM [6].

In a recent work, Lu et al. [15] explored different fusion techniques to combine multiple relevance models estimated on different query variants. A dual cross-media relevance model for image annotation was proposed by [16] which estimates the joint probability of image and annotation words. The study in [17] proposed a positional relevance model (PRM) which preferentially weighs the term co-occurrence relations with the positional information of words.

A cross-lingual extension of RLM was proposed in [18], where given a query in the source language, the corresponding topic model was estimated in the target language with the help of dictionary-based word alignment. It was shown by [19] that the estimation of relevance models can be improved by using large external corpora. The results reported in [19] showed that using external corpus was 10% more effective in terms of mean average precision compared to standard relevance language model [5].

The difference of our proposed cross relevance language model with existing modifications of relevance language model is that rather than using extra information from external corpus [19] or parallel corpora [17] or the position information of the words in documents [18] we exploit the different types of document structure within a corpus to improve the effectiveness of the relevance models.

Precision Medicine Retrieval. Wang and Akella [20] used a concept based relevance language model, where the documents and queries are transformed from the term space into concept space to capture the semantic relations between different biomedical terms. Agosti et al. [2] investigated the effect of different query formulation techniques for the TREC precision medicine task. The study in [21] proposed the use of knowledge graphs for explainable biomedical document retrieval using neural models. An online system to experiment with different gene based query expansion techniques in the TREC precision medicine track setup was developed by [22]. Domain knowledge extracted from external corpus has been shown to improve retrieval effectiveness of biomedical documents [23]. While [24] investigated the effect of different tokenization techniques in biomedical document retrieval, [25] proposed a learning to rank framework for biomedical document retrieval. An explanation framework for biomedical document retrieval was proposed in [26] with an objective to capture the different aspects of relevance.

3. Cross Relevance Model (CRM)

Although we apply the proposed cross relevance model for the PMIR task, which involves two independent collections, we present our model in a generic manner, i.e., we assume that the task is to retrieve documents from m different collections and then collate them into a single ranked list.

Relevance Model (RLM). We start with a brief technical introduction to the relevance model (RLM) [5]. RLM essentially estimates a probability distribution of term weights based on the co-occurrence likelihoods. These co-occurrences are computed between the query terms and the ones present in the top-retrieved documents. Formally speaking, the weight of a term w in the RLM distribution computed for a given query Q is

$$P(w|Q) = \sum_{D \in M_k(Q)} P(w|D) \prod_{q \in Q} P(q|D), \tag{1}$$

where M is the set of top-k ranked documents retrieved for Q with an initial retrieval model, e.g. BM25 [11].

Inter-collection RLM. We now generalize the notion of RLM to span across different collections, as required in PMIR. In general, denoting the set of $m \in \mathbb{N}$ documents collections as $C = \cup_{i=1}^{M} \{C_i\}$, for a given query q, we first retrieve M different ranked lists, one for each C_i by a retrieval model such as BM25. Let the ranked list of top-k documents retrieved from the collection C_i be denoted by $M_k(Q, C_i)$.

Our objective now is to model the term dependence relationships between the terms of a collection C_i with respect to the ones that are outside the collection, i.e., $C - C_i$. Towards that objective, we generalize the RLM distribution of 1 as

$$P(w|Q, C_i) = \lambda \underbrace{\sum_{D \in M_k(Q, C_i)} P(w|D) \prod_{q \in Q} P(q|D)}_{\text{RLM for } C_i} + (1 - \lambda) \underbrace{\sum_{D \in M_k(Q, C - C_i)} P(w|D) \prod_{q \in Q} P(q|D)}_{\text{RLM for } C - C_i}. \tag{2}$$

In Equation 2, we see that the term distribution weights are now a linear combination of the term co-occurrence likelihoods within the current collection C_i, and also those in the set of all collections external to C_i. The parameter $\lambda \in [0, 1]$ controls the relative importance of the intra-collection likelihoods with respect to the inter-collection ones.

The top-k ranked documents retrieved from the complementary set $C - C_i$ is obtained by computing the top-k from the COMBSUM [27] over each retrieved list external to the collection C_i, i.e.,

$$M_k(Q, C - C_i) = \text{TOP}_k(\text{COMBSUM}(\bigcup_{j=1, j \neq i}^{m} M(Q, C_j))). \tag{3}$$

After estimating the RLMs (term weight distributions) within and across each collection C_i as per Equation 2, we execute the standard steps of RM3 [28] based retrieval, i.e., employ a linear combination of the original query terms along with the weights for additional terms estimated by RLM for a second pass retrieval. This gives us m different ranked lists for each C_i. Finally, we employ another COMBSUM to retrieve the top-1000 results.

	Dataset Statistics	
Property	Trec 2019	Trec 2020
#Documents	26,909,414	32,122,364
#PubMed abstracts	26,669,007	31,818,247
#Clinical trial doc	241,006	306,238
Avg Doc. Length	200	200
#Queries	50	40

Table 1
Dataset statistics for TREC 2019 and 2020 Precision Medicine Track.

4. Experimental Setup

4.1. Datasets

We use the TREC 2019 [3] and TREC 2020 [2] precision medicine track datasets for our experiments. The dataset is comprised of a document collection collated from two different sources, namely the PubMed articles and the documents from the ClinicalTrials[4] collection. While the former contains abstracts of published studies, the latter contains early findings of ongoing or recently completed randomized control trials. The value of m, i.e., the number of document collections in CRM, thus corresponds to 2 for our experiments. Table outlines the characteristics of the dataset 1.

Dataset Pre-Processing. The documents within the two respective sub-collections are structured into several different fields, e.g., 'MeshHeading', 'Abstract', 'Keywords' etc. While an option is to index the information from each field separately, after some initial experiments we found that a flat index structure, i.e., information from all the fields being merged onto a single one, yielded better retrieval results. Consequently, for all our reported experiments, the document representations followed a flat bag-of-words approach.

For the purpose of indexing, we used Lucene[5], a Java-based framework. We employed the standard steps of stopword removal (SMART stop-list) and stemming (Porter stemmer).

Each query in the TREC precision medicine track [2, 3] is structured into three different fields, namely 'Gene Name', 'Disease Name' and 'Demographic Category'. As with the documents, we also conducted initial experiments seeking to investigate the effect of structured query representations on retrieval effectiveness. Similar to our observations for the document representation, it turned out that an unstructured (flat) bag-of-words representation of the queries yielded better results in comparison to a structured one. Consequently, all our reported experiments use the unstructured query representation.

[4] https://clinicaltrials.gov/
[5] https://lucene.apache.org/

4.2. Baselines

We explored a number of baselines to compare against our proposed approach of PRF. Since our focus is on investigating how effectively we can exploit all the information present in a given corpus, we did not consider baseline methods that make use of external corpora or resources. Each baseline is described below.

1. **BM25**: We used the standard BM25 retrieval model [11] to retrieve from the index constructed from the overall collection (i.e., comprising both clinical trials and PubMed articles). The objective of using BM25 as one of our baselines is to observe how a retrieval model performs in the absence of any pseudo-relevance feedback on the PMIR task.

2. **BM25+RM**: In our second baseline, we apply the standard RM-based PRF [5] on an initial list retrieved with BM25. This baseline is, in fact, an ablation to our proposed model and corresponds to the situation when term semantics is modeled only within individual collections.

3. **COMBSUM**: In this approach, rather than executing search on a single combined index, we rather apply BM25 on two separate indexes - one for the PubMed abstracts, and the other for the clinical trial documents). We then combine the two ranked lists obtained from these two indexes into a single list using the COMBSUM method [27]. This baseline represents a relatively simple way of addressing the information of the two sub-collections in a separate manner, a term in each index having its own *local* collection statistics and term weights.

The TREC 2019 [3] and 2020 [2] Precision medicine tracks mainly used nDCG as the official metric. In addition to nDCG, we use MAP, P@5, P@10 to report the results of our experiments. Since the objective of all our experiments is to demonstrate the retrieval effectiveness in the presence of different pseudo relevance feedback, in the context of our research scope, MAP is the most important evaluation metric.

4.3. Proposed Approaches

We now describe the settings for our proposed PRF approach, namely the CRM and its variants.

1. **Self-only RM (SRM)**: This approach, which we call 'Self-only RM' (or SRM) uses the same setup as the COMBSUM method on separate indexes (Section 4.2). The only difference is that, in contrast to retrieving results only with the initial retrieval and merging the result lists, in this method we first execute RM-based PRF on each retrieved list. The post-feedback list of documents are then merged with the COMBSUM method. The objective of using this baseline is to study the effects of independent PRFs on separate indexes followed by a post-hoc merging of the results. In this sense, the SRM method is an ablation to our method because instead of jointly modeling within and across feedback, this method simply applies a post-hoc merging of results obtained with intra-feedback models.

2. **Cross-only RM**: This is another ablation for our proposed PRF model of Equation 2, where we set $\lambda = 0$ to prune the intra-RLM components. This model exclusively relies on the cross-collection term semantics dependencies.

Dataset	Method	Model	Metrics			
			MAP	NDCG	P@10	P@5
TREC 2019	Baseline	BM25	0.1901	0.4653	0.4980	0.5560
		BM25 + RM	0.1966	0.4741	0.5160	0.5600
		COMBSUM	0.2010	0.4983	0.5100	0.5760
	Proposed	SRM ($\lambda = 1$)	0.2104	0.5088	0.5200	0.5680
		Cross-only RM ($\mathrm{CRM}_{\lambda=0}$)	0.2129	**0.5156**	0.5180	0.5560
		CRM ($\lambda = 0.7$)	**0.2136**	0.5131	**0.5240**	**0.5840**
TREC 2020	Baseline	BM25	0.1564	0.4476	0.3800	0.3950
		BM25 + RM	0.1622	0.4556	0.3675	0.4100
		COMBSUM	0.1602	0.4597	0.3451	0.3900
	Proposed	SRM ($\lambda = 1$)	0.1701	0.4754	0.3725	0.3751
		Cross-only RM ($\mathrm{CRM}_{\lambda=0}$)	0.1665	0.4701	0.3625	0.4050
		CRM ($\lambda = 0.7$)	**0.1746**	**0.4852**	**0.3902**	**0.4250**

Table 2
A comparative analysis of the performance of CRM against the baselines and ablations on TREC 2019 and 2020 Precision Medicine datasets. The best values along each column are bold-faced. For each method, the BM25 parameters values for k and b are 1 and 0.6, respectively. The optimal values of the number of feedback documents and the number of terms for expansion were found to be 5 for all the PRF approaches.

3. **CRM**: This is our proposed approach of Equation 2, with $m = 2$ for the TREC PM collection. It models term semantics both for the PubMed articles and the clinical trials separately, as well as modeling the cross term semantics across the two collections.

4.4. Parameter Setting

For all our experiments, we used BM25 as our initial retrieval model on top of which the PRF models are eventually applied. The hyper-parameters of each IR model investigated were tuned with grid search and the best results (along with the parameter configurations) are reported in Table 2. In particular, the two parameters of BM25, namely k and b, were tuned with a grid search. Both k and b were varied within the interval of 0.1 to 1 in steps of 0.1.

As for the feedback models, there are two parameters common to all the methods that we investigated, namely the number of feedback documents (F), and the number of terms (K) with the top RM weights used for query expansion. We tuned F in a range from 5 to 20 for each PRF method and reported the best results in Table 2. Similarly the optimal value of K was found out to be 5 after tuning it in the range of 5 to 30. Similarly, the value of λ, the mixing parameter for CRM, was also varied in steps of 0.1 within the interval $[0, 1]$. The optimum value for λ for CRM is reported in Table 2.

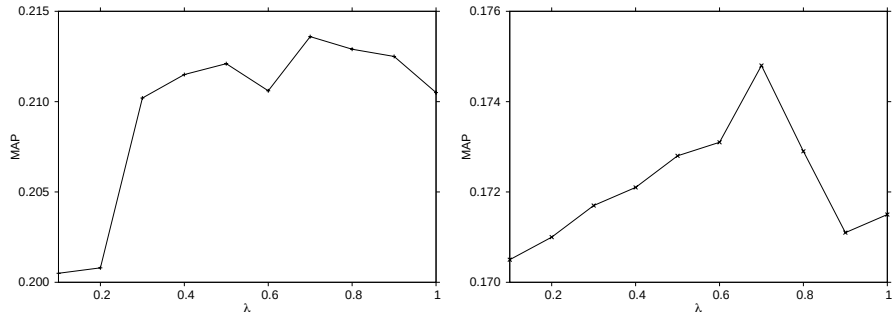

Figure 1: Effect of varying the CRM parameter λ on MAP values for TREC Precision Medicine query sets - 2019 (left) 2020 (right).

5. Results

The results obtained for both TREC 2019 and 2020 precision medicine track datasets are presented in Table 2. The interesting observations that can be made from Table 2 are as follows.

Firstly, BM25+RM performs better than BM25 in terms of all the evaluation metrics. Similar observation follows for COMBSUM and SRM. This observation confirms that the use of PRF improves retrieval effectiveness for the PMIR task.

Secondly, the COMBSUM model outperforms BM25. This observation confirms our hypothesis that the use of sub-collection helps to improve retrieval effectiveness.

Thirdly, our proposed approach, CRM, significantly outperforms all the baseline and the ablation approaches for both TREC 2019 and 2020 datasets. This observation supports our hypothesis that jointly modeling the self and cross relevance models separately on each sub-collection improves the overall retrieval effectiveness.

In TREC 2019 precision medicine task, the best nDCG value reported was 0.5783 [4]. The method in [4] used a supervised approach using external information. Table 2 shows that the best NDCG value for our proposed approach is 0.5156. This shows that without using any external information the performance of our proposed approach is comparable to approaches using external knowledge bases.

Figure 1 shows the effect of varying the parameter λ for CRM. It is observed from Figure 1 that for both TREC 2019 and TREC 2020 the best MAP value is obtained for $\lambda = 0.7$. This observation indicates that a higher relative importance should be assigned to the intra-collection term co-occurrence likelihoods in comparison to the inter-collection ones.

6. Conclusions and Future Work

In this paper we proposed a cross relevance model (CRM) for precision medicine document retrieval. The proposed approach used a mixture of both intra-collection and inter-collection relevance models to better model the two different types of term dependencies. In future, we would like to investigate the effect of cross relevance model on collections that do not have distinct information sources. Instead, we could apply ideas like clustering to partition a collection into a number of different pseudo sub-collections.

References

[1] D. Papaioannou, A. Sutton, C. Carroll, A. Booth, R. Wong, Literature searching for social science systematic reviews: consideration of a range of search techniques, Health Information & Libraries Journal 27 (2010) 114–122.

[2] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, Overview of the trec 2020 precision medicine track, The ... text REtrieval conference : TREC. Text REtrieval Conference 26 (2020).

[3] K. Roberts, D. Demner-Fushman, E. Voorhees, W. Hersh, S. Bedrick, A. Lazar, S. Pant, F. Meric-Bernstam, Overview of the trec 2019 precision medicine track, 2019.

[4] E. Faessler, M. Oleynik, U. Hahn, Julie lab & med uni graz @ trec 2019 precision medicine track, 2020.

[5] V. Lavrenko, W. B. Croft, Relevance based language models, in: Proc. of SIGIR '01, ACM, New York, NY, USA, 2001, pp. 120–127.

[6] D. Roy, D. Ganguly, M. Mitra, G. J. Jones, Word vector compositionality based relevance feedback using kernel density estimation, in: Proc. of CIKM '16, 2016, pp. 1281–1290.

[7] H. Zamani, J. Dadashkarimi, A. Shakery, W. B. Croft, Pseudo-relevance feedback based on matrix factorization, in: Proc. of CIKM '16, ACM, 2016, pp. 1483–1492.

[8] A. Montazeralghaem, H. Zamani, J. Allan, A reinforcement learning framework for relevance feedback, in: SIGIR, ACM, 2020, pp. 59–68.

[9] P. Ogilvie, E. Voorhees, J. Callan, On the number of terms used in automatic query expansion, Information Retrieval 12 (2009) 666–679.

[10] G. Cao, J.-Y. Nie, J. Gao, S. Robertson, Selecting good expansion terms for pseudo-relevance feedback, in: SIGIR '08, 2008, p. 243–250.

[11] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, A. Payne, Okapi at trec-4, 1996.

[12] A. Chakraborty, D. Ganguly, A. Caputo, S. Lawless, A factored relevance model for contextual point-of-interest recommendation, in: Proc. of ICTIR '19, ACM, New York, NY, USA, 2019, pp. 157–164.

[13] S. Datta, D. Ganguly, D. Roy, F. Bonin, C. Jochim, M. Mitra, Retrieving potential causes from a query event, in: SIGIR, ACM, 2020, pp. 1689–1692.

[14] H. Zamani, W. B. Croft, Relevance-based word embedding, in: Proc. of SIGIR '17, ACM, New York, NY, USA, 2017, p. 505–514.

[15] X. Lu, O. Kurland, J. S. Culpepper, N. Craswell, O. Rom, Relevance modeling with multiple query variations, in: Proc. of ICTIR '19, 2019, p. 27–34.

[16] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, S. Ma, Dual cross-media relevance model for image annotation, in: Proc. of the 15th ACM International Conference on Multimedia, Association for Computing Machinery, 2007, p. 605–614.

[17] Y. Lv, C. Zhai, Positional relevance model for pseudo-relevance feedback, in: Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, p. 579–586.

[18] V. Lavrenko, M. Choquette, W. B. Croft, Cross-lingual relevance models, in: Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, p. 175–182.

[19] F. Diaz, D. Metzler, Improving the estimation of relevance models using large external

corpora, in: Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, p. 154–161.

[20] C. Wang, R. Akella, Concept-based relevance models for medical and semantic information retrieval, in: Proc. of the 24th ACM International on Conference on Information and Knowledge Management, Association for Computing Machinery, 2015, p. 173–182.

[21] Z. Yang, Biomedical Information Retrieval Incorporating Knowledge Graph for Explainable Precision Medicine, 2020, p. 2486.

[22] V. Nguyen, S. Karimi, B. Jin, An experimentation platform for precision medicine, in: Proc. of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, 2019, p. 1357–1360.

[23] S. Wang, M. Hauskrecht, Improving biomedical document retrieval using domain knowledge, in: Proc. of the SIGIR, Association for Computing Machinery, 2008, p. 785–786.

[24] D. Trieschnigg, W. Kraaij, F. de Jong, The influence of basic tokenization on biomedical document retrieval, in: Proc. of SIGIR 2007, 2007, p. 803–804.

[25] B. Xu, H. Lin, Y. Lin, Y. Ma, L. Yang, J. Wang, Z. Yang, Improve biomedical information retrieval using modified learning to rank methods, IEEE/ACM Trans. Comput. Biol. Bioinformatics 15 (2018) 1797–1809.

[26] J. Qu, J. Arguello, Y. Wang, Towards Explainable Retrieval Models for Precision Medicine Literature Search, 2020, p. 1593–1596.

[27] J. A. Shaw, E. A. Fox, Combination of multiple searches, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 105–108. URL: http://trec.nist.gov/pubs/trec3/papers/vt.ps.gz.

[28] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at TREC 2004: Novelty and HARD, in: TREC 2004, 2004.