MeSH-based dataset for measuring the relevance of text retrieval

Won Kim, Lana Yeganova, Donald C Comeau,
W John Wilbur, Zhiyong Lu
National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA
{wonkim, yeganova, comeau, wilbur, luzh}@mail.nih.gov

Abstract

Creating simulated search environments has been of a significant interest in information retrieval, in both general and biomedical search domains. Existing collections include modest number of queries and are constructed by manually evaluating retrieval results. In this work we propose leveraging MeSH term assignments for creating synthetic test beds. We select a suitable subset of MeSH terms as queries, and utilize MeSH term assignments as labels for retrieval evaluation. Using well studied retrieval functions, we show that their performance on the proposed data is consistent with similar findings in previous work. We further use the proposed retrieval evaluation framework to better understand how to combine heterogeneous sources of textual information.

1 Introduction

PubMed is a search engine processing on average 3 million queries a day and is recognized as a primary tool for scholars in the biomedical field (M. Falagas, Pitsouni, Malietzis, & Pappas, 2008; Lu, 2011; Wildgaard & Lund, 2016).

PubMed provides access to a collection of approximately 28 million biomedical abstracts as of 2018, of which about 4.5 million have full text document available in PubMed Central. With the growing availability of full-text articles, an essential question to consider is how to leverage full text information to improve PubMed retrieval? While a number of studies have pointed out the benefits of full text for various text mining tasks (Cohen, Johnson, Verspoor, Roeder, & Hunter, 2010; Westergaard, Stærfeldt, Tønsberg, Jensen , & Brunak, 2018), combining these two resources for information retrieval is not a trivial endeavor. Naïvely merging full text articles with abstract data, naturally increases the recall, but at a cost for precision, generally degrading the overall quality of combined search (Lin, 2009).

Research is required to understand how to best combine abstracts and full texts, examine the relative importance of different sections in full text, investigate the performance of different scoring functions, etc. A major obstacle in such efforts is the lack of large-scale gold standards for retrieval evaluation. Hence, creating such large-scale retrieval evaluation framework is the goal of this work.

Gold standards are typically assembled by using human judgments, which are time consuming, expensive and not scalable. Pioneering examples are a TREC collection (Hersh, Cohen, Ruslen, & Roberts, 2007) and a BioASQ collection (Tsatsaronis et al., 2015). Simulating test collections for evaluating retrieval quality offers a viable alternative and has been explored in the literature (Azzopardi & de Rijke, 2006; Azzopardi, de Rijke, & Balog, 2007; Kim, Yeganova, Comeau, Wilbur, & Lu, 2018). In this work we create an evaluation framework based on MeSH term assignments, and use that framework to test the performance of several classic ranking functions.

We examine the utility of MeSH terms as query surrogates and MeSH term assignments as pseudo-relevance rankings. We describe how we select a subset of MeSH terms as candidate MeSH queries and discuss the retrieval results using five different retrieval functions available in SOLR. MeSH queries are representative of real user queries. This approach allows us to create a large-scale relevance ranking framework that is based on human judgements and is publicly available. MeSH queries are available for download at:

Proceedings of the BioNLP 2018 workshop, pages 161–165
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

https://ftp.ncbi.nlm.nih.gov/pub/wbur/mesh_queries/.

2 MeSH Term Based Queries for Retrieval Evaluation

Each paper indexed by MEDLINE® is manually assigned on average thirteen MeSH terms (Huang, Neveol, & Lu, 2011) by an indexer, who has access to both the abstract and full text of articles. It is plausible to assume that MeSH terms assigned to a document are highly reflective of its topic, and the document is highly relevant to that MeSH term.

In this work we propose using a subset of MeSH terms as queries and rely on the assumption that documents with the MeSH terms assigned are relevant to the query. As queries, we aim to select MeSH terms that satisfy certain frequency requirements, and those that are correlated with real user queries. We will refer to the final set of MeSH terms that we use as queries as MeSH queries. Using MeSH terms for evaluation of various NLP tasks has been described in the literature (Bhattacharya, Ha−Thuc, & Srinivasan, 2011; Yeganova, Kim, Kim, & Wilbur, 2014). However, to our knowledge, using MeSH terms as query surrogates and MeSH assignments as relevance rankings has not been yet described.

2.1 MeSH term preprocessing

We preprocess the MeSH terms by applying several processing steps, which include lowercasing, removing all non-alphanumeric characters, and dropping stop words from MeSH term strings. We further drop tokens in the remaining MeSH term string that are pure digits.

2.2 Frequency Threshold

We apply frequency threshold to remove MeSH terms that are not likely to be useful as queries. Some MeSH terms such as *Humans*, are very general, and are not useful for evaluation of retrieval results. *Humans* is assigned to an overwhelming fraction of PubMed documents, even to those that are not directly discussing the topic. For example, an article studying *dietary experiments on rats involving the hormone "insulin"* is assigned *humans* because it studied animals to understand diabetes for humans. Another complication are ambiguous MeSH terms. With the frequency threshold, our goal is to limit the analysis to those MeSH terms that tend to carry the same meaning across the corpus.

For a single token MeSH term, we consider two frequencies: the number of PubMed documents the MeSH term is assigned to, and the frequency of the token used as a text word in PubMed abstracts. For a single token MeSH term, we required that the smaller of the two frequencies is at least half as big as the larger. For multi-token MeSH terms, the frequency with which each individual token in the MeSH term appears in the text is at most ten times as high as the frequency of the MeSH term. These requirements lead to 5,117 single-token and 1,735 multi-token MeSH terms for use as queries.

2.3 Presence in User Queries

The second essential consideration is to select MeSH terms that are likely to be used as queries. We collected PubMed queries issued in the 2017 calendar year. We normalized these user queries in the same manner as MeSH terms. We found that among the 5,117 single token MeSH terms, about half of them appeared as queries. Among the 1,735 multi-token MeSH terms 96% have been issued as a query. Based on this analysis, we decided to proceed with the multi-token MeSH queries for our experiments. We will refer to that set of MeSH terms as MeSH queries.

3 SOLR Retrieval Functions

SOLR is an open source search platform built on Apache Lucene which has been widely used in the search industry for more than a decade. It offers a number of useful features including fast speed, distributed indexing, replication, load-balanced querying, and automated failover and recovery. Lucene-based SOLR search engine is a popular industry standard for indexing, search and retrieval. SOLR provides several ranking options, and our interest is in evaluating them using MeSH queries and pseudo-relevance judgements.

We investigated most of the weighting formulas available in the native SOLR/Lucene search engine, and report the top five best performing ones: tf.idf, BM25, DFR, IBS and Dirichlet.

tf.idf is the SOLR default ranking algorithm and one of the most basic weighting schemes used in information retrieval (Robertson, 2004).

	MAP	BE
tf .idf	0.380	0.506
BM25	**0.413**	**0.532**
DFR	**0.417**	**0.536**
IBS	0.404	0.524
Dirichlet	0.305	0.454

Table 1. Retrieval results for multi-word queries, based on the top 2K retrieved documents. Presented are averages over 1,735 multi-word MeSH queries.

BM25 is the ranking algorithm described in (Robertson SE, 1995) and (Sparck Jones, Walker, & Robertson, 1998).

DFR is the implementation based upon the *divergence from randomness (DFR)* framework introduced in (Amati & Van Rijsbergen, 2002) .

IBS is based upon a framework for the family of information-based models, as described in *(Clinchant & Gaussier, 2010).*

Dirichlet is an language model for Bayesian smoothing using Dirichlet priors from (Zhai & Lafferty, 2004).

4 Results

MeSH terms are assigned based on article abstracts and full texts, hence it is natural to include in the retrieval experiments not only PubMed articles, but also corresponding PubMed Central full text articles. To that end, we created a retrieval environment which included all PubMed articles (~27 million abstracts) and their available PMC full text counterparts (~4 million full texts) in a unified system. The search environment was created in such a way that we can distinguish PubMed and PMC records, and identify which PMC record corresponds to a PubMed abstract. The retrieval system, however, treated all PubMed and PMC documents independently. For PubMed records, we indexed the title and the abstract fields, for the PMC full text records we indexed title, abstract and full text fields. We evaluated each retrieval method available in SOLR by querying the unified database using MeSH queries. Retrieved documents (both PubMed and PMC) where scored using SOLR weighting functions and returned in the order of diminishing score.

For each MeSH query, we retrieved the top 2,000 documents. Among those, we considered only documents to which MeSH terms have already been assigned (recent documents may not have been assigned MeSH terms yet) and call them the *retrieved set*. Documents in the *retrieved set* that are assigned MeSH query as a MeSH term are treated as *positive*, while the rest are considered *negative*. Given these assignments, we can compute Mean Average Precision (MAP) and Precision-Recall Break Even (BE) (M. Falagas, Pitsouni, E., Malietzis, G., & Pappas, G., 2008) to measure the success of each retrieval function.

Table 1 presents the summary of the retrieval results from SOLR using the five different weighting formulas, averaged over the 1,735 multi-token MeSH queries. Table 1 shows that BM25 outperforms tf.idf in terms of both MAP and BE. This result is consistent with results reported in (Lin, 2009). We also observe that BM25 and DFR outperform the other three ranking methods, with DFR showing slightly better results than BM25.

A common consideration with document ranking formulas is how robust they are to document length. This next experiment examines whether different ranking formulas favor shorter PubMed abstracts to longer PMC full text documents, or the opposite. Among the top 2,000, we considered *positive* retrieved documents for which both PubMed and PMC records exist. For such articles, it is possible for both PubMed and PMC records to be included in top 2K or just one of them to be present. For each query, we counted the total number of *positive* documents as PMC articles that are ranked higher than PubMed articles (denoted as PMC > PM), as well as the number of positive documents for which PubMed articles are ranked higher (PM > PMC).

The counts are presented in Table 2. We observe that tf.idf pulls more PubMed abstracts into the highest scoring 2,000, thus favoring relatively short (PubMed) documents. Dirichlet, on the other hand favors PubMed Central full text articles. These experiments suggest that tf.idf and Dirichlet are more extreme. By contrast, BM25, DFR and IBS favor PubMed abstracts, but not as strongly.

Our next goal is to consider the value of full text articles for retrieval. We analyze the retrieval performance by computing MAP and BE measures in retrieving 1) PubMed articles only 2) PMC articles only and 3) both PubMed and PMC articles using BM25 and DFR retrieval functions. For the combined retrieval, we assign each article the maximum of its PubMed and PMC score

	Total # of positives	PMC>PM	PM>PMC
tf.idf	136K	13.5K (10%)	122.5K (90%)
BM25	190K	52K (27%)	138K (73%)
DFR	190K	59K (31%)	131K (69%)
IBS	199K	72K (37%)	126K (63%)
Dirichlet	468K	455K (97%)	13K (3%)

Table 2. Comparison of PubMed and PMC scores for multiword queries based on top 2K retrieved documents. The counts are included only for the articles for which both PubMed and PMC versions exist, and one or both are in the top 2K.

and evaluate based on that maximum. We observe from Table 3, that both BM25 and DFR performed better in retrieving PubMed articles than PMC articles. Using the maximum of the PubMed score and PMC score does not yield improved performance over the abstract-only search for both BM25 and DFR.

		BM25	DFR
PMC	MAP	0.265	0.273
	BE	0.353	0.360
PubMed	MAP	**0.305**	**0.309**
	BE	**0.390**	**0.394**
Combined	MAP	0.167	0.279
	BE	0.270	0.376

Table 3. The value of full text PMC articles in the retrieval performance. In combined retrieval, we assign each article the maximum of its PubMed and PMC score and evaluate based on that maximum.

5 Conclusion and Discussion

In this work we propose a large-scale collection for relevance testing. The collection represents a subset of MeSH terms that we use as queries and MeSH term assignments as pseudo relevance rankings. The value of this resource is significant not only in its simplicity and intuitiveness, but also in the quality of relevance judgements achieved though leveraging decades of manual curation. Moreover, by using MeSH terms we are guaranteed to include as queries significant and important PubMed topics. Many of these terms are frequently used as queries. To summarize, MeSH queries provide a reliable and high-quality collection of queries.

To further validate the feasibility of this collection, we used well studied retrieval functions on the set. In the future, we plan to use the proposed test collection to understand how to leverage full text documents for better search.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems, 20*(4), 357-389. doi: Doi 10.1145/582415.582416

Azzopardi, L., & de Rijke, M. (2006). *Automatic construction of known-item finding test beds*. Paper presented at the SIGIR '06.

Azzopardi, L., de Rijke, M., & Balog, K. (2007). *Building Simulated Queries for Known-Item Topics*. Paper presented at the SIGIR'07, Amsterdam, The Netherlands.

Bampoulidis, A., Lupu, M., Palotti, J., Metallidis, S., Brassey, J., & Hanbury, A. (2016). Interactive exploration of healthcare queries. *14th International Workshop on Content-Based Multimedia Indexing (CBMI)*.

Bhattacharya, S., Ha−Thuc, V., & Srinivasan, P. (2011). MeSH: a window into full text for document summarization. *Bioinformatics, 27*(13).

Clinchant, S., & Gaussier, E. (2010). Information-based models for ad hoc IR (2010). *SIGIR'10, conference on Research and development in information retrieval.*

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., & Hunter, L. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics, 11*(492).

Falagas, M., Pitsouni, E., Malietzis, G., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal, 22*(2), 338-342.

Falagas, M., Pitsouni, E., Malietzis, G., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal, 22*(2), 338-342.

Hersh, W., Cohen, A., Ruslen, L., & Roberts, P. (2007). *Genomics Track Overview.* Paper presented at the Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007).

Huang, M., Neveol, A., & Lu, Z. (2011). Recommending MeSH terms for annotating biomedical articles. *J Am Med Inform Assoc, 18*(5), 660-667. doi: 10.1136/amiajnl-2010-000055

Kim, S., Yeganova, L., Comeau, D. C., Wilbur, W. J., & Lu, Z. (2018). PubMed Phrases, an open set of coherent phrases for searching biomedical literature. *Scientific Data, in press*.

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics, 10*, 46. doi: 10.1186/1471-2105-10-46

Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. Database: the journal of biological databases and curation. *Database (Oxford), 2011*.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation., 60*(5), 503–520.

Robertson SE, W. S., Hancock-Beaulieu M, Gatford M, Payne A. (1995). Okapi at TREC-4. *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, 73-96.

Sparck Jones, K., Walker, S., & Robertson, S. E. (1998). A probabilistic model of information retrieval: development and status (pp. 1-75): University of Cambridge.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., . . . Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics, 16*, 138. doi: 10.1186/s12859-015-0564-6

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen , L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *Plos Computational Biology, 14*(2).

Wildgaard, L. E., & Lund, H. (2016). Advancing PubMed? A comparison of 3rd-party PubMed/MEDLINE tools. *Library Hi Tech, 34*(4), 669-684. doi: https://doi.org/10.1108/LHT-06-2016-0066

Yeganova, L., Kim, W., Kim, S., & Wilbur, W. J. (2014). Retro: concept-based clustering of biomedical topical sets. *Bioinformatics, 30*(22).

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179-214.