Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw331
Advance Access Publication Date: 10 June 2016
Applications note

OXFORD

Data and text mining

Meshable: searching PubMed abstracts by utilizing MeSH and MeSH-derived topical terms

Sun Kim*, Lana Yeganova and W. John Wilbur

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on November 3, 2015; revised on May 17, 2016; accepted on May 22, 2016

Abstract

Summary: Medical Subject Headings (MeSH®) is a controlled vocabulary for indexing and searching biomedical literature. MeSH terms and subheadings are organized in a hierarchical structure and are used to indicate the topics of an article. Biologists can use either MeSH terms as queries or the MeSH interface provided in PubMed® for searching PubMed abstracts. However, these are rarely used, and there is no convenient way to link standardized MeSH terms to user queries. Here, we introduce a web interface which allows users to enter queries to find MeSH terms closely related to the queries. Our method relies on co-occurrence of text words and MeSH terms to find keywords that are related to each MeSH term. A query is then matched with the keywords for MeSH terms, and candidate MeSH terms are ranked based on their relatedness to the query. The experimental results show that our method achieves the best performance among several term extraction approaches in terms of topic coherence. Moreover, the interface can be effectively used to find full names of abbreviations and to disambiguate user queries.

Availability and Implementation: https://www.ncbi.nlm.nih.gov/IRET/MESHABLE/

Contact: sun.kim@nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

PubMed is the most comprehensive bibliographic database of biomedicine and life sciences. As PubMed continues to grow, it becomes more difficult to discover what one is looking for. But what if we could search and browse PubMed through subject areas? Searching documents by subjects not only can help focus the search, but can also help explore the landscape for a subject of interest.

Many studies have pointed out the usefulness of finding subjects in search results (Sarkar *et al.*, 2009; Struble and Dharmanolla, 2004; Theodosiou *et al.*, 2011). The common approach in these studies is to group the search results on the fly; many methods make use of ontological information provided by the Medical Subject Headings (MeSH) resource for grouping. Our goal is to provide a convenient way to use the MeSH resource at query time. There are a few options for doing this. One is by directly querying with a MeSH term, e.g. *otitis[MeSH]*. This, however, does not appear to be the most convenient way for accessing the information because MeSH is a controlled vocabulary with a specific syntax. More frequently the MeSH resource is used indirectly by expanding a text query with the related MeSH term. For example, the query *ear infection* is automatically expanded to include the MeSH term *otitis[MeSH]*. However, such expansions are limited to terms listed as synonyms or variants for a given MeSH term. An alternative way to search using MeSH terms is by starting from the MeSH interface (http://www.ncbi.nlm.nih.gov/mesh) and using the PubMed query builder. Then again this approach is not intuitively available at the PubMed search page and may only be known to more experienced users.

In this study, the goal is to search PubMed abstracts through the lens of the MeSH vocabulary. The key question we address here is how to link a text query with relevant MeSH concepts and we approach it by setting a supervised task as follows. First, for every MeSH term we define the MeSH-Doc set as the group of documents

assigned the MeSH term. We then identify significant terms appearing in the MeSH-Doc set based on the occurrence frequency of these terms in the set versus the rest of PubMed. We refer to these terms as topic terms associated with the particular MeSH term. As a result, topic terms along with corresponding weights are computed for every MeSH term. At query time, MeSH concepts are ranked based on how they weight the terms in the query.

2 Methods

As of July 2015, a total of 386 960 MeSH term/subheading combinations had been assigned to two or more PubMed abstracts. Our goal is (i) to identify significant topic terms appearing in each MeSH-Doc set and (ii) to link a query to the relevant MeSH terms or subheadings through the identified terms. It is known that controlled vocabularies such as UMLS and MeSH have low usage in biomedical literature (Kim *et al.*, 2015a). Since MeSH terms are manually curated, we expect prominent text terms of a MeSH-Doc set to be either the same or closely related to the MeSH term defining the MeSH-Doc set. The major benefit of applying this strategy is that it enables us to associate the controlled vocabulary with the phrases that actually appear in literature. Users can leverage this association to search relevant MeSH terms/subheadings and to define a more focused document set that was curated by humans.

Topic modeling is a popular approach to identify latent topics and their topic terms. However, it is not needed for our purposes because topics (e.g. MeSH) are already known in our application. Supervised topic models (Zhu *et al.*, 2009) utilize known labels, but they are mostly used for topic classification, not topic-term identification problems. Moreover, LDA (Latent Dirichlet Allocation), a popular topic modeling approach, is not an especially effective way to provide coherent topic terms (Kim *et al.*, 2015b). Thus, we explore the term identification process based on the theme generation framework introduced in Wilbur (2002) and Kim and Wilbur (2012).

In Wilbur (2002), a theme is defined as a set of documents and significant terms appearing in the document set. Starting from seed documents, a theme is obtained by iteratively choosing a set of documents and then a set of terms based on an EM (Expectation-Maximization) framework. Since our setup has labeled data for each MeSH term/subheading, this iterative process is unnecessary. Hence, only the term extraction procedure of the theme framework is utilized for prioritizing terms from PubMed abstracts (see Supplementary Material for more details). The significance of a term is obtained from the α parameter (defined in Supplementary Material) in the approach of Wilbur (2002) and Kim and Wilbur (2012), which is the difference between the contribution coming from the term depending on whether it is considered as belonging to the topic or not. This approach not only calculates the importance of terms efficiently, but also shows better topic coherence scores than some other feature extraction methods (see Section 3).

For a given text query, *Meshable* scores each MeSH entry by finding the rank of each query term among the MeSH topic terms (The current system finds the rank among top 20 topic terms, but this may be changed depending on user feedback.). If a query consists of multiple words (up to three), the ranks of individual words are averaged for scoring (If more than three words are entered as a query, only first three words are used for processing. If no match is found for an individual word, a high value is assigned as a rank, and the MeSH terms without this individual word will not appear as a result.). The MeSH terms found are then listed as a search result. The interface contains tooltips to show top-ranked topic terms for each MeSH entry, links to MeSH descriptions and the option to perform a PubMed search based on identified MeSH ('MESH' and 'KEY' buttons) or its combination with the query ('KEY+' button). Figure 1 shows the screenshot of the result for the query, 'diabetes'. For the first entry, 'hypoglycemic agents', three buttons are provided for PubMed search. 'MESH' (blue) and 'KEY' (green) use 'hypoglycemic agents' for searching PubMed. The difference is that the 'MESH' only finds the documents to which the MeSH was assigned, but 'KEY' retrieves all the documents that include 'hypoglycemic agents' in any fields, e.g. titles, abstracts, journals, MeSH, etc. 'KEY+' (pink) adds the original query, 'diabetes' to the query produced by 'KEY'.

3 Results and discussion

Our topic term extraction module has been applied to PubMed abstracts (July 2015) that had both titles and abstracts. For each MeSH term/subheading, the documents with the MeSH entry are considered as positives and other documents as negatives. Unigrams and bigrams are used as features for the term extraction process. In order to evaluate how well our extraction process works, we applied two coherence measures to a sample of our results.

Coherence measures (Aletras and Stevenson, 2013; Mimno *et al.*, 2011) are commonly used to evaluate how focused selected topic terms are for a topic, and they are known to be correlated with human judgments (Aletras and Stevenson, 2013). Thus, we used UMASS (Mimno *et al.*, 2011) and NPMI (normalized point-wise mutual information) (Aletras and Stevenson, 2013) for evaluating topic terms. Both UMASS and NPMI measure the average of pairwise similarities for topic terms, with the difference that UMASS also takes the order of topic terms into account. UMASS is defined as

$$\text{UMASS} = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \log \frac{p(t_i, t_j) + \epsilon}{p(t_j)},$$

where $p(t_i, t_j)$ is the fraction of documents containing both terms t_i and t_j and $p(t_j)$ is the fraction of documents containing the term t_j. n indicates the number of top topic terms. $\epsilon = \frac{1}{N}$ is the smoothing factor, where N is the size of the dataset. NPMI is also computed by

$$\text{NPMI} = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{\log \frac{p(t_i,t_j)+\epsilon}{p(t_i)p(t_j)}}{-\log(p(t_i, t_j) + \epsilon)}.$$

Table 1 shows the performance comparison between our method and other feature extraction approaches based on UMASS and NPMI measures. In this experiment, 100 MeSH terms were

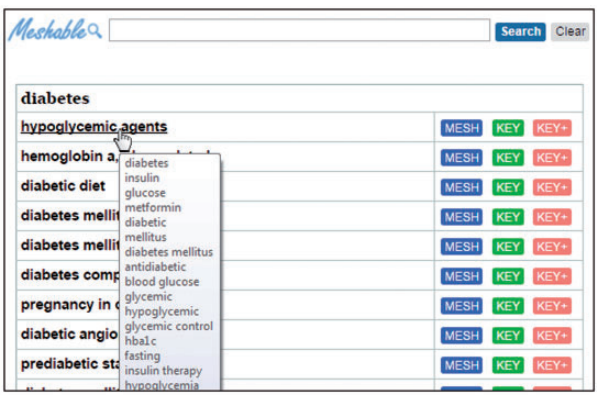

Fig. 1. Screenshot of the result for the query, 'diabetes'

Table 1. Performance comparison between our theme method and other feature extraction approaches for Top 5 and Top 10 topic terms

Methods	Top 5		Top 10	
	UMASS	NPMI	UMASS	NPMI
Theme	−12.3707	8.2984	−76.2566	34.9573
Bayes weights	−60.6455	4.9002	−280.0770	21.0571
Chi-square	−27.6835	7.0750	−189.4230	27.4682
Hypergeometric test	−12.3707	8.2984	−76.2657	34.9573

The coherence measures, UMASS and NPMI, are used to evaluate topic terms obtained from 100 random MeSH entries, and the scores are averaged.

randomly selected among those assigned to between 1000 and 100 000 PubMed documents, and the scores shown are averages over the 100 MeSH topics. As shown in the table, the theme method and the hypergeometric test (Kim and Wilbur, 2001) show the best performance. In contrast, Bayes weights (Kim and Wilbur, 2001) and chi-square (Liu and Motoda, 2007) provide significantly lower coherence scores. Although the hypergeometric test is closely related and ties with our approach here, they are not identical. Since it takes more time to evaluate the hypergeometric test, we use our approach.

A useful application of our method is to find full forms for abbreviations, e.g. *AD* to *Alzheimer disease* and *FMF* to *familial Mediterranean fever*. To evaluate this, 100 terms were randomly selected from disease or biomedical abbreviations in Wikipedia (https://en.wikipedia.org) and All Acronyms (http://www.allacronyms.com). As a result, our interface returned at least one full form for 91% of the abbreviations. Among these, 74 full names were exactly matched with the ones given in the websites (see Supplementary Material for more details). Current PubMed automatically expands abbreviations to retrieve more relevant documents, however it often does not work (e.g. SARS).

Another use case scenario is to disambiguate multiple concepts from a query. For instance, given the query *cat*, our interface finds the concepts, *cats (animal)*, *chloramphenicol o-acetyltransferase (protein)*, *cat's claw (plant)*, etc (The results for the example queries, *cat* and *diabetes*, are shown in Supplementary Material.). The MeSH interface works similarly, but it only matches queries with MeSH descriptions. Our approach associates MeSH with information from PubMed abstracts by distributional semantic analysis, hence it allows queries to retrieve important MeSH terms which are closely related in meaning.

Funding

Intramural Research Program of the NIH, National Library of Medicine.

Conflict of Interest: none declared.

References

Aletras,N. and Stevenson,M. (2013). Evaluating topic coherence using distributional semantics. In *Proc. International Conference on Computational Semantics (IWCS)*, pp. 13–22.

Kim,S. and Wilbur,W.J. (2012) Thematic clustering of text documents using an EM-based approach. *J. Biomed. Semant.*, **3**, S6.

Kim,S. *et al.* (2015a) Identifying named entities from PubMed for enriching semantic categories. *BMC Bioinformtics*, **16**, 57.

Kim,S. *et al.* (2015b). Summarizing topical contents from PubMed documents using a thematic analysis. In: *Proc. Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pp. 805–810.

Kim,W. and Wilbur,W.J. (2001) Corpus-based statistical screening for content-bearing terms. *J. Am. Soc. Inform. Sci. Technol.*, **52**, 247–259.

Liu,H. and Motoda,H. (2007). *Computational Methods of Feature Selection*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Mimno,D. *et al.* (2011). Optimizing semantic coherence in topic models. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 262–272.

Sarkar,I.N. *et al.* (2009) LigerCat: using "MeSH clouds" from journal, article, or gene citations to facilitate the identification of relevant biomedical literature. In: *AMIA Annu. Symp. Proc.*, **2009**, pp. 563–567.

Struble,C.A. and Dharmanolla,C. (2004). Clustering MeSH representations of biomedical literature. In *Proc. HLT-NAACL 2004 Workshop on Linking Biological Literature Ontologies and Databases*, pp. 41–48.

Theodosiou,T. *et al.* (2011) MeSHy: Mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms. *J. Biomed. Inform.*, **44**, 919–926.

Wilbur,W.J. (2002) A thematic analysis of the AIDS literature. In *Proc. Pacific Symposium on Biocomputing*, pp. 386–397.

Zhu,J. *et al.* (2009) MedLDA: maximum margin supervised topic models for regression and classification. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1257–1264.