# MeSH Term Suggestion for Systematic Review Literature Search

Shuai Wang  
University of Queensland  
Brisbane, Australia  
shuai.wang5@uq.net.au  

Hang Li  
University of Queensland  
Brisbane, Australia  
hang.li4@uq.net.au  

Harrisen Scells  
University of Queensland  
Brisbane, Australia  
h.scells@uq.net.au  

Daniel Locke  
University of Queensland  
Brisbane, Australia  
locke.dj@hotmail.com  

Guido Zuccon  
University of Queensland  
Brisbane, Australia  
g.zuccon@uq.edu.au  

## ABSTRACT

High-quality medical systematic reviews require comprehensive literature searches to ensure the recommendations and outcomes are sufficiently reliable. Indeed, searching for relevant medical literature is a key phase in constructing systematic reviews and often involves domain (medical researchers) and search (information specialists) experts in developing the search queries. Queries in this context are highly complex, based on Boolean logic, include free-text terms and index terms from standardised terminologies (e.g., MeSH), and are difficult and time-consuming to build. The use of MeSH terms, in particular, has been shown to improve the quality of the search results. However, identifying the correct MeSH terms to include in a query is difficult: information experts are often unfamiliar with the MeSH database and unsure about the appropriateness of MeSH terms for a query. Naturally, the full value of the MeSH terminology is often not fully exploited.

This paper investigates methods to suggest MeSH terms based on an initial Boolean query that includes only free-text terms. These methods promise to automatically identify highly effective MeSH terms for inclusion in a systematic review query. Our study contributes an empirical evaluation of several MeSH term suggestion methods. We perform an extensive analysis of the retrieval, ranking, and refinement of MeSH term suggestions for each method and how these suggestions impact the effectiveness of Boolean queries.

ACM Reference Format:
Shuai Wang, Hang Li, Harrisen Scells, Daniel Locke, and Guido Zuccon. 2021. MeSH Term Suggestion for Systematic Review Literature Search. In ADCS '21. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1234/123456

## 1 INTRODUCTION & RELATED WORK

A medical systematic review is a comprehensive review of literature for a highly focused research question. Systematic reviews are seen as the highest form of evidence and are used extensively in healthcare decision making and clinical medical practice. In order to synthesise literature into a systematic review, a search must be undertaken. A major component of this search is a Boolean query.

The Boolean query is often developed by a trained expert (i.e., an information specialist), who works closely with the research team to develop the search, and usually has knowledge about the domain.

The most commonly used database for searching medical literature is PubMed. Due to the increasing size and scope of the PubMed database, the Medical Subject Headings (MeSH) ontology was developed to conceptually index studies [17, 25]. MeSH is a controlled vocabulary thesaurus arranged in a hierarchical tree structure (specificity increases with depth in a parent–child relationship, e.g., Anatomy → Body Regions → Head → Eye... etc.). Indexing and categorising studies with MeSH terms enables queries to be developed which incorporate both free-text keywords and MeSH terms — enabling more effective searches. The use of MeSH terms in queries has been shown to be more effective than free-text keywords alone [1, 7, 17, 23], e.g. they increase precision [14] and are far less ambiguous than free-text [24]. However, it is still difficult even for expert information specialists to be familiar with the entire MeSH controlled vocabulary [13, 14] — at the time of writing, MeSH contains 29,640 unique headings.

One way that PubMed has attempted to overcome this difficulty is to develop a method called Automatic Term Mapping (ATM). ATM is an automatic query expansion method which attempts to seamlessly map free-text keywords in a query to one of the three categories (index tables): MeSH, journal name or author name [16]. Although ATM is applied by default for all queries issued to PubMed, it has several semantic limitations: it is inaccurate when used to expand free-text acronyms into MeSH terms [21]; will produce different MeSH expansions even though synonymous free-text terms are used [2], and has difficulty disambiguating between MeSH terms and journal names [22]. Despite these limitations, the use of ATM for MeSH term suggestion has been shown to increase the precision of free-text searches in the genomic domain [15]. However, its use has, to the best of the authors knowledge, not been empirically evaluated in the context of improving the effectiveness of systematic review literature search queries.

Our paper introduces the task of MeSH term suggestion for Boolean queries used in systematic review literature search. We model this task within the context of an information specialist looking for MeSH terms to add to a query without MeSH terms present. In addition to new MeSH suggestion methods, we also propose a framework to evaluate the effectiveness of the suggestion of MeSH terms on an established collection of systematic review literature search queries.
The contributions of this paper are:

(1) The introduction of the new task of suggesting MeSH terms for systematic review literature search (Boolean queries), modelled within the context of an information specialist looking for MeSH terms to add to a query without MeSH terms present.

(2) An empirical evaluation of the effectiveness of MeSH suggestion methods for this task (i.e., ranking MeSH terms for a query).

(3) An empirical evaluation of the effectiveness of Boolean queries using the suggestions made by different suggestion methods (i.e., retrieving abstracts for a query given different suggested MeSH terms).

2 MeSH Term Suggestion

Next, we outline how we perform MeSH term suggestions for Boolean queries. As the Boolean queries used for systematic review literature search are highly complex, containing nested Boolean clauses, MeSH terms are suggested not globally, but instead locally, for query fragments. A query fragment is a clause of a Boolean query containing semantically related text clauses (i.e., free-text or MeSH terms). Each text clause in a query fragment is grouped into a Boolean clause using the OR operator. To give an intuition for how query fragments are derived and utilised for MeSH term suggestions, see Figure 1. The OR operators in Figure 1 are implicit. We exploit these fragments to perform a fine-grain evaluation for MeSH term suggestion (i.e., in terms of retrieval performance, ranking performance, and refinement of the ranking performance). However, we also perform defragmentation to obtain a Boolean query with suggested MeSH terms for comparison to the original Boolean queries.

We propose to suggest MeSH terms in a pipeline of three steps: retrieval, ranking, and refinement. The following three sections provide a description of how we approach each of these steps.

2.1 MeSH Term Retrieval

The first step in our MeSH term suggestion pipeline is the retrieval of MeSH terms. The retrieval of MeSH terms is facilitated by three different methods:

- **ATM** The entire free-text only query fragment is submitted to the PubMed entrez API [18] for ATM. When free-text clauses without specific qualifiers are present in a query, the three index tables (MeSH, journal name, author name) are searched sequentially to determine if a mapping exists. If there is no mapping found initially, the free-text clause is divided into individual terms and the process is repeated. Mapped terms are filtered to only include those that are MeSH terms.

- **MetaMap** Each free-text clause in a query fragment is submitted to MetaMap [3]. The results from MetaMap are filtered to only include those entities derived from the MeSH source. All of the mapped MeSH terms are recorded for each of the free-text terms in a query fragment. Additionally, the MetaMap score is recorded for each MeSH term.

- **UMLS** We index the UMLS [5] (version 2018 with options set to default values). Qualifiers are the terminology PubMed uses for field restrictions. Keywords in a query containing semantically related text clauses (i.e., free-text or MeSH terms) are present in a query, the three index tables (MeSH, journal name, author name) are searched sequentially to determine if a mapping exists. If there is no mapping found initially, the free-text clause is divided into individual terms and the process is repeated. Mapped terms are filtered to only include those that are MeSH terms.

For the MetaMap and UMLS approaches, the same MeSH term may be retrieved multiple times for a given free-text clause. To overcome this issue, we re-score the MeSH terms using rank fusion (CombSUM) [8]. The intuition for this re-scoring is that highly common MeSH terms that also obtain a high score from these retrieval methods should be scored highly overall (thus ranked higher than common MeSH terms and highly scoring MeSH terms).

2.2 MeSH Term Ranking

Once MeSH terms have been retrieved, they are ranked according to the approach for entity ranking described by Jimmy et al. [9] by adapting features proposed by Balog [4]. In total, we use eleven features, each described in Table 1. For the description of MeSH terms ($d_e$), we scrape the corresponding Wikipedia page. We generate features for each MeSH term retrieval method (i.e., ATM, MetaMap, UMLS). Positive instances correspond to MeSH terms in the original query fragment, negative instances correspond to MeSH terms not

1Qualifiers are the terminology PubMed uses for field restrictions. Keywords in a query may be explicitly restricted to certain fields, e.g., title, abstract, MeSH terms etc.

2Version 2018 with options set to default values.
Whenever we encounter ties, we treat all of the tied MeSH terms as
Therefore, tied MeSH terms at the top of rankings are more likely
tied MeSH terms account for much larger accumulations of gain.

t of tied MeSH terms at the boundary of the cut-off defined by
resulting in at least one MeSH term suggested for every query
re-scoring MeSH terms becomes apparent when used with the
The 
boundary of the cut-off (i.e., a refinement of the ranking). The 
method: the highest-ranking MeSH term will receive a score of 
the percentage of total 
for the MeSH term.
We use topics from the CLEF TAR task from 2017, 2018, and 2019
We evaluate the effectiveness of MeSH term suggestions retrospectively using the MeSH terms identified from pre-existing queries as a gold standard. In doing so, we make the assumption that the MeSH terms in these pre-existing queries are the ideal choices. As such, this gold standard may be biased to favour the PubMed ATM method, as it could have been used to suggest MeSH terms originally. The MeSH term suggestion methods proposed above are likely to identify MeSH terms that were not originally in pre-existing query fragments. To combat this assumption, we also evaluate the retrieval effectiveness achieved by the queries with the proposed suggestions. We therefore evaluate both (i) the effectiveness of query suggestion given the assumption that MeSH terms in pre-existing queries are a gold standard; and (ii) the effectiveness of the query at retrieving studies.

Note that (ii) also has limitations: that query fragments must be combined back into the original query structure in order to properly evaluate the query; and new MeSH terms may retrieve studies that are unjudged (it is unknown if the retrieved unjudged studies are relevant or not). To account for these unjudged studies, we use the approach proposed by Scells et al. [20], which calculates, in addition to the lower bound typically assumed (i.e., all unjudged studies are irrelevant), an upper bound (i.e., assume all unjudged studies are relevant) and a balance between the two (i.e., assume some unjudged studies to be relevant given a maximum likelihood estimation over the judged studies).

The effectiveness of the MeSH term suggestion is evaluated using reciprocal rank, nDCG@5,10, recall@5,10, precision, and recall. Precision and recall measure the effectiveness of the retrieval of MeSH terms by the three retrieval methods. nDCG and reciprocal rank measure the effectiveness of the LTR entity ranking model for each of the three retrieval models.

To evaluate the effectiveness of the suggested MeSH terms for the task of systematic review literature search, once query fragments are defragmented, the retrieval effectiveness is evaluated using typical systematic review literature search measures: precision, recall, and Fβ=[0.5,1,3]. To obtain retrieval results, the PubMed entrez API is used to directly issue defragmented Boolean queries. For reproducibility purposes, as PubMed is constantly updated with new studies, we apply a date restriction to all queries.

For both evaluation settings (i.e., ranking MeSH term suggestion and Boolean query retrieval), we evaluate the quality of ranking in two settings: (i) all, where all retrieved MeSH terms are considered; and (ii) cut, where a score-based cut-off is determined to filter the suggested MeSH terms.

| Feature | Description |
|---------|-------------|
| | Total free-text terms in a fragment |
| | Length of description of MeSH term e |
| | Sum IEF of free-text terms |
| | Sum TF of free-text terms in d_e |
| | Sum TF of free-text terms in d_e |
| | LM score of free-text terms for d_e |
| | BM25 score of free-text terms for d_e |
| | SDM score of free-text terms for d_e |
| | Whether the free-text terms contain e |
| | Whether e contains any free-text terms |
| | Whether e is equal to the free-text terms |

Table 1: Features used in MeSH term ranking.
are discarded because of retrieval issues (2017: CD010276, CD010173, CD012019; 2018: CD011926; 2019: CD010038), likely resulting from the fact that some queries are automatically translated from queries in one format (Ovid Medline) into another format (PubMed). In total we used 242 topics across all three datasets (114 unique, as each year has partial overlap). For each topic, we divide the Boolean query for that topic into several query fragments. We create these fragments using the transmute tool [19]. Each fragment contains at least one MeSH term. This results in a total of 302 unique query fragments for the three years (2.65 fragments per query on average). For each of the query fragments, we corrected any errors (e.g., spelling mistakes, syntactic errors), extracted MeSH terms, keywords, query fragment with MeSH terms, and query fragments without MeSH terms. For training the LTR model for MeSH term ranking, the pre-split training and test portions from the CLEF datasets are used. The 2019 topics are split also on systematic review type (intervention and diagnostic test accuracy — indicated as I and D respectively in the results), while those for 2017 and 2018 are all diagnostic test accuracy. We use the quickrank library [6] for LTR, instantiated with LambdaMART trained to maximise nDCG. We leave other settings as per default.

4 RESULTS

All of the results in this section are presented on the testing portions of each CLEFT TAR year (i.e., 2017, 2018, 2019/I, 2019/D).

4.1 Suggestion Effectiveness

4.1.1 Retrieval of MeSH Terms

Firstly, we investigate the MeSH term suggestion methods’ effectiveness in retrieving terms given a query fragment. Table 2 reports precision (P) and recall (R) for the retrieval of MeSH terms. When comparing the three base retrieval methods, UMLS generally retrieves more relevant terms than ATM and MetaMap, as suggested by the higher recall value for UMLS than the other two methods. However, the UMLS method achieves lower precision than the other two methods, indicating that it retrieves too many MeSH terms. The fusion method achieves the highest recall across all datasets. However, it never outperforms the other methods in terms of precision (naturally because it combines all the MeSH suggestions). We find that: (i) UMLS is the most effective MeSH retrieval method for recall, (ii) ATM is the most effective retrieval method for precision, and (iii) that fusion of multiple MeSH retrieval methods generally leads to the highest recall and lowest precision.

4.1.2 Ranking of MeSH Terms

Next, we investigate the effectiveness of the LTR model at ranking the retrieved MeSH terms for each retrieval method. For this task, we observe the reciprocal rank (RR), R@5 (Recall@5), and nDCG@k, and reciprocal rank (RR).

Table 2: Effectiveness of the MeSH term suggestion methods with respect to precision (P), recall@5 (R@5), nDCG@k, and reciprocal rank (RR). M indicates MetaMap, U indicates UMLS, F indicates fusion, Two-tailed statistical significance (p < 0.05) with Bonferroni correction between ATM, and the other methods, for each year is indicated by *.
Meaning of the MeSH term suggestion when used in a Boolean query to search literature for systematic reviews. M indicates MetaMap, U indicates UMLS, F indicates fusion. For evaluation measures, Opt indicates optimistic treatment of residuals, MLE indicates maximum likelihood estimation treatment of residuals. Two-tailed statistical significance (p < 0.05) with Bonferroni correction between the ORIGINAL query for each year and queries with new MeSH suggestions is indicated by *.  

4.1.3 Refinement of MeSH Terms. Finally, we investigate the effect of refining the ranked MeSH terms by cutting off the ranking at a certain point and discarding the remainder. We estimate this cut-off point through a parameter. Our tuning results on the training portions of the datasets are presented in Figure 2. We believe that the spikes in these plots generally correspond to the inclusion and exclusion of ties. These spikes are most prominent in the MetaMap, and ATM methods as these methods do not assign highly discriminative scores to MeSH terms. Furthermore, note that the UMLS and fusion methods have considerably smoother shapes, as these methods have highly discriminative scores.  

We investigate the effect that this refinement has on the MeSH term suggestion performance in Table 2 (i.e., with -CUT). We find that refinement generally improves precision while lowering recall. The loss in recall attributed to the refinement negatively affects ranking effectiveness. In most cases, refinement is worse than ranking effectiveness. In most cases, refinement generally improves precision while lowering recall.  

### 4.2 Search Effectiveness

We next investigate the impact in performance that MeSH term retrieval, ranking, and ranking refinement has on the retrieval effectiveness of Boolean queries.  

First, we examine the retrieval effectiveness when we consider unjudged studies to be irrelevant. This assumption is a typical retrieval evaluation scenario and provides a lower bound on effectiveness. For the 2017 and 2018 datasets, we find that few methods increase precision over the original queries (both are refined rankings); however, for the two 2019 datasets, there is no method where we see an increase in precision over the original queries. However, in terms of recall, the unrefined fusion ranking improved recall except for a single case (2018). This result is likely because this method adds the most MeSH terms. For all of the results in precision and recall, the results are statistically significant.  

Comparing these results to our optimistic and MLE residual treatments of unjudged studies, we find that the unrefined fusion ranking achieves the highest results in all evaluation measures across all datasets, likely a result of the fact that it retrieves the most MeSH terms; (ii) Although we find that unranked fusion still
achieves the highest recall for the MLE treatment, it generally performs worse than other methods.

To our surprise, the refined fusion method did not achieve the highest result among any evaluation measure or dataset. Indeed, refinement of rankings generally lowered recall and had a negligible effect on precision for all methods. This result is interesting for us, as it suggests that in practice, it is beneficial to add as many MeSH terms to a query as possible rather than to select a few key MeSH terms.

5 CONCLUSIONS

In this paper, we presented the new task of suggesting MeSH terms within the context of systematic review literature search (suggestion for Boolean queries). We provided a comprehensive evaluation of the effectiveness of MeSH suggestion methods (in terms of retrieval, ranking, and refinement). We compared these methods to the existing method that PubMed uses to suggest MeSH terms (ATM). We found that both the MetaMap and UMLS suggestion methods can improve the retrieval effectiveness of Boolean queries. Unsurprisingly, when we combined the three methods using rank fusion, we found the highest gains in retrieval effectiveness.

Our methods overcome the semantic limitations of ATM: the MetaMap and UMLS methods both suggested more relevant MeSH terms than ATM, and the addition of these terms positively impacted retrieval performance. Often this came with a minor loss in recall. Note that there are generally between 10–100 relevant terms than ATM, and the addition of these terms positively impacted retrieval performance. Often this came with a minor loss in recall. Note that there are generally between 10–100 relevant studies per topic: the actual impact of loss in recall is attributed to only a handful of studies and is likely not to impact the results of a systematic review.

Identifying MeSH terms to add to a Boolean query for systematic review literature search is known to be a difficult task for humans to accomplish. The outcomes of this paper have implications for both the information retrieval and systematic review communities. Firstly, our methods can be used in automatic query formulation situations (see, e.g., tasks in CLEF TAR). Secondly, they can be integrated into existing tools to assist information specialists in formulating more effective queries.

REFERENCES

[1] Samir Abdou and Jacques Savoy. 2008. Searching in Medline: Query expansion and manual indexing evaluation. Information Processing & Management 44, 2 (2008), 781–789.

[2] KP Adlassnig et al. 2009. Optimization of the PubMed automatic term mapping. In Medical Informatics in a United and Healthy Europe: Proceedings of MIE 2009, the XXII International Congress of the European Federation for Medical Informatics, Vol. 150. IOS Press, 238.

[3] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium. American Medical Informatics Association, 17.

[4] Kristian Balog. 2018. Entity-oriented search. Springer.

[5] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32, suppl_1 (2004), D267–D270.

[6] Gabriele Capani, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto. 2016. Quality versus efficiency in document scoring with learning-to-rank models. Information Processing & Management 52, 6 (2016), 1161–1177.

[7] Angela A Chang, Karen M Heskett, and Terence M Davidson. 2006. Searching the literature using medical subject headings versus text word with PubMed. The Laryngoscope 116, 2 (2006), 336–340.

[8] Edward A Fox and Joseph A Shaw. 1994. Combination of multiple searches. NIST special publication SP 243 (1994).

[9] Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Card Retrieval for Consumer Health Search: An Empirical Investigation of Methods. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM ’19). Association for Computing Machinery, New York, NY, USA, 2405–2408. https://doi.org/10.1145/3357384.3358128

[10] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. 2017. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In CLEF’17.

[11] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In CEUR Workshop Proceedings, Vol. 2380.

[12] Evangelos Kanoulas, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview: In CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS.

[13] Ying-Hsang Liu. 2009. The impact of MeSH (Medical Subject Headings) terms on information seeking effectiveness. Ph.D. Dissertation. Rutgers University-Graduate School-New Brunswick.

[14] Ying-Hsang Liu and Nina Wacholder. 2017. Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searches. Information Processing & Management 53, 4 (2017), 851–870.

[15] Zhiyong Lu, Won Kim, and W John Wilbur. 2009. Evaluation of query expansion using MeSH in PubMed. Information retrieval 12, 1 (2009), 69–80.

[16] AM Nahin. 2003. Change to PubMed’s automatic term mapping affects phrase searching. NLM Tech Bull 331 (2003).

[17] Randy R Richter and Tricia M Austin. 2012. Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. Physical therapy 92, 1 (2012), 124–132.

[18] Eric Sayers. 2010. A General Introduction to the E-utils. Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information (2010).

[19] H. Scells, D. Locke, and G. Zuccon. 2018. An Information Retrieval Experiment Framework for Domain Specific Applications. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.

[20] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean Query Refinement for Systematic Review Literature Search. In The Web Conference (WebConf ’19). 1646–1656.

[21] Stefan Schulz, Martin Honeck, and Udo Hahn. 2001. Indexing medical WWW documents by morphemes. Studies in health technology and informatics 1 (2001), 266–270.

[22] Aida Marissa Smith. 2004. An examination of PubMed’s ability to disambiguate subject queries and journal title queries. Journal of the Medical Library Association 92, 1 (2004), 97.

[23] Carol Tenopir. 1985. Full text database retrieval performance. Online Review (1985).

[24] Nina Wacholder, Yael Ravin, and Misook Choi. 1997. Disambiguation of proper names in text. In Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 202–208.

[25] Yuri L Zierman and Howard L Bleich. 1997. Conceptual mapping of user’s queries to medical subject headings. In Proceedings of the AMIA Annual Fall Symposium. American Medical Informatics Association, 519.