International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424
Volume 3 Issue 8 , August 2026
DOI: https://doi.org/10.71366/ijwos

PEDAGOGICAL INTERVENTIONS TO ENHANCE MESH TERM SELECTION SKILLS FOR HEALTH RESEARCHERS

S. Narmatha[1], Dr. V. Maniraj[2]
[1] *Research Scholar, Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli), Poondi, Thanjavur, Tamilnadu, India.*
[2] *Research Advisor, PG and Research Dept. of Computer Science, AVVM Sri Pushpam College (Affiliated to Bharathidasan University, Tiruchirappalli)., Poondi, Thanjavur, Tamilnadu, India..*

Article Info

Article History:

Published: 6 August 2026

Publication Issue:
Volume 3, Issue 8
August-2026

Page Number:
66-74

Corresponding Author:
S. Narmatha

Abstract:

Systematic reviews in the medical field require comprehensive and carefully structured literature searches to ensure the validity and reliability of their findings. These searches are typically developed through collaboration between health researchers and information specialists who combine subject expertise with advanced knowledge of information retrieval methods. Search strategies frequently involve complex Boolean logic that integrates free-text keywords with controlled vocabularies, particularly Medical Subject Headings (MeSH). While the use of MeSH terms can significantly enhance both the accuracy and completeness of search results, selecting the most appropriate terms remains a challenging task due to the complexity of the MeSH system and varying levels of user familiarity. Insufficient expertise may result in inefficient searching or underutilization of MeSH features. This study explores automated approaches for recommending relevant MeSH terms based on initial keyword-only search queries. The primary aim is to assist users in identifying high-value MeSH terms that can strengthen systematic review search strategies. The performance of multiple recommendation methods is assessed through their impact on retrieval effectiveness, ranking quality, and the refinement of iterative search queries.

Keywords: Medical Subject Headings

1. Introduction

Systematic reviews in medicine provide structured and comprehensive syntheses of existing evidence to answer well-defined clinical or scientific questions. As a fundamental component of evidence-based healthcare, these reviews are essential for guiding clinical practice, supporting policy development, and advancing medical knowledge. A critical element of the systematic review process is the construction of thorough and precise literature search strategies, which are most often implemented using carefully designed Boolean queries. These strategies are typically developed through collaboration between domain experts and information specialists, integrating subject knowledge with expertise in information retrieval.

PubMed is the primary resource for searching biomedical literature and organizes its rapidly growing content using the Medical Subject Headings (MeSH) controlled vocabulary. MeSH is structured hierarchically, encompassing broad conceptual domains such as “Anatomy” as well as highly specific descriptors like “Eye.” When combined with free-text keywords, MeSH terms enhance search effectiveness by improving recall and precision, minimizing ambiguity, and enabling more consistent retrieval of relevant publications.

Previous research has demonstrated that search strategies incorporating MeSH terms retrieve more relevant and comprehensive results than strategies based solely on free-text keywords. Despite these advantages, the identification of appropriate MeSH terms remains a complex and time-intensive task, even for experienced information professionals. This challenge is primarily attributed to the scale and intricacy of the MeSH vocabulary, which currently includes over 29,000 unique descriptors.

To assist users in formulating effective queries, PubMed employs Automatic Term Mapping (ATM), a mechanism that links free-text search terms to relevant MeSH descriptors, journal titles, or author names. While ATM has shown utility in specific domains, such as genomics research, it also presents several limitations, including inconsistent synonym recognition, difficulties in processing acronyms, and occasional misinterpretation of terms as journal titles rather than MeSH descriptors. Moreover, despite its widespread adoption, the performance of ATM within systematic review search workflows has not been comprehensively assessed.

This study explores automated methods for recommending MeSH terms to support the construction of Boolean search strategies for systematic reviews. Specifically, it focuses on scenarios in which information professionals begin with keyword-based queries and require automated assistance to identify relevant MeSH terms that can improve the accuracy, completeness, and reliability of literature searches.

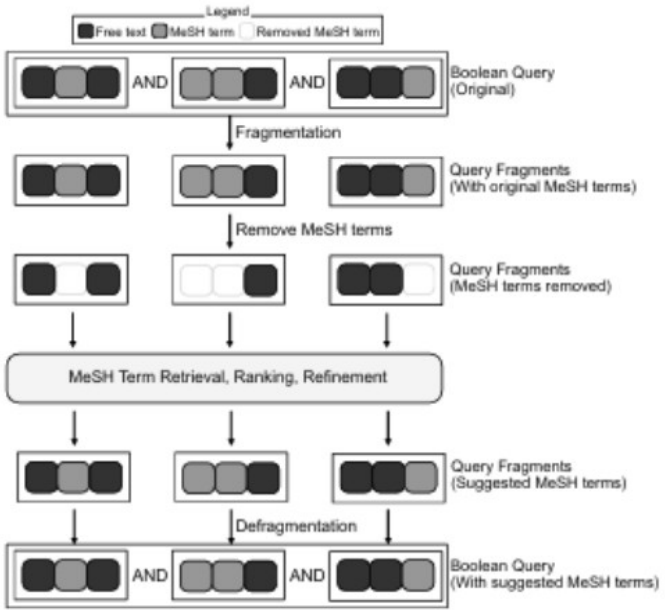


Figure 1: Workflow for Recommending MeSH Terms

Rewritten Version

The MeSH term recommendation process is organized as a structured pipeline consisting of three stages: retrieval, ranking, and iterative refinement. To compare and assess the performance of different MeSH recommendation methods, four key evaluation criteria are applied:
(1) the accuracy with which relevant MeSH terms are identified;
(2) the relevance and practical usefulness of the ranking assigned to recommended terms;
(3) the degree to which ranking quality improves through successive refinement steps; and
(4) the effect of integrating the recommended MeSH terms into incomplete Boolean queries on overall literature retrieval performance.
It should be noted that the number of MeSH terms produced for a given query fragment may vary, sometimes exceeding and at other times falling below the number of original free-text terms.

2. Related Work

This study presents an evaluation framework designed to assess how effectively MeSH term recommendations can enhance search strategies used in systematic reviews. Using a collection of Boolean queries extracted from published systematic reviews, we propose novel methods for generating MeSH term suggestions. Our approach builds upon

existing research that explores computational support for the construction and optimization of Boolean queries in evidence synthesis.

The primary contributions of this study include:

1. **Task Formulation:** We introduce a new task focused on MeSH term recommendation for systematic review searches, framed from the perspective of an information specialist who begins with a Boolean query composed exclusively of free-text terms.
2. **Applied Evaluation:** We conduct a systematic evaluation of MeSH recommendation techniques, emphasizing the accuracy and usefulness of term rankings relative to the original query intent.
3. **Comparative Analysis:** We empirically compare multiple MeSH recommendation approaches by assessing their impact on improving the retrieval of relevant abstracts when incorporated into Boolean search strategies.

3. MeSH Term Recommendation

Our framework generates MeSH term suggestions tailored specifically to Boolean search queries. Given the complex and often deeply nested nature of systematic review queries, MeSH recommendations are applied at the level of query fragments rather than to the query as a whole. A query fragment is defined as a semantically cohesive component of the query, typically composed of clauses connected by the OR operator and containing either free-text or MeSH terms. Each clause within a fragment represents an individual Boolean condition.

Figure 1 illustrates the construction of query fragments and their role in the MeSH recommendation workflow, with OR relationships between clauses implied. Operating at the fragment level enables focused evaluation of MeSH suggestions in terms of retrieval effectiveness, ranking accuracy, and improvements achieved through ranking refinement.

Following the recommendation stage, a defragmentation step reintegrates the selected MeSH terms into the complete Boolean query. The revised query is then evaluated against the original to measure changes in retrieval performance. The MeSH recommendation process consists of three main stages—retrieval, ranking, and refinement—which are described in detail below.

MeSH Term Retrieval

The retrieval stage identifies candidate MeSH terms from query fragments composed solely of free-text clauses. Three complementary retrieval techniques are employed:

1. **Automatic Term Mapping (ATM):** Each query fragment is submitted to the PubMed Entrez API using unqualified free-text terms. PubMed attempts to map these terms to MeSH headings, journal titles, or author names. When no direct mapping is identified, the fragment is decomposed into individual terms. Only MeSH headings are retained in the final candidate set.
2. **MetaMap:** Each sentence within a query fragment is processed using MetaMap, a tool developed by the National Library of Medicine. MetaMap identifies biomedical concepts and maps them to entries in the UMLS Metathesaurus. Results are filtered to retain only MeSH-related concepts, each accompanied by a confidence score representing the likelihood of a correct mapping.
3. **UMLS Elasticsearch Indexing:** Selected UMLS tables (MRCONSO, MRDEF, MRREL, and MRSTY, version 2019AB) are indexed using Elasticsearch (version 7.6). Non-MeSH concepts are excluded, and each query clause is searched against the index. Candidate MeSH terms are scored using the BM25 ranking function.

As MetaMap and UMLS-based methods may produce overlapping MeSH candidates for a given clause, a rank fusion strategy (CombSUM) is applied to merge scores. This approach favors MeSH terms that consistently appear across multiple methods with strong individual scores, while reducing the influence of common or weakly ranked terms.

MeSH Term Ranking

Following retrieval, candidate MeSH terms are ordered using an entity ranking approach adapted from Balog et al. A total of eleven ranking features are employed, as detailed in Table 1. To enrich term representations (denoted as d_c), supplementary contextual information is extracted from the corresponding Wikipedia pages.

Each retrieval technique—ATM, MetaMap, and UMLS indexing—generates method-specific features for the MeSH terms it produces. Terms already present in the original query fragment are labeled as positive instances, while absent terms are labeled as negative instances, creating binary training data for a learning-to-rank (LTR) model developed separately for each retrieval method.

A rank fusion mechanism is then applied to combine normalized scores across all three methods into a unified ranking. This strategy ensures that MeSH terms identified frequently and consistently by multiple retrieval techniques are prioritized, increasing the likelihood of recommending highly relevant terms.

MeSH Term Refinement

The final stage of the framework refines the ranked list by filtering out less relevant MeSH terms and retaining only the most informative candidates for each query fragment. A rank-based cut-off is applied using the parameter κ, which specifies the top percentage of ranked terms to retain. Values of κ are evaluated from 5% to 95% in increments of 5%. To ensure that each fragment contributes at least one MeSH term, the highest-ranked term is always retained and assigned a score of zero.

When multiple MeSH terms share identical scores at the κ threshold, all tied terms are treated as a single group, and their contributions are aggregated as combined gain. This conservative strategy increases the likelihood that highly ranked ties are preserved while allowing lower-ranked ties to be excluded. As a result, the refinement process promotes the inclusion of the most relevant MeSH terms and avoids arbitrary exclusion of potentially valuable candidates.

Evaluation

To evaluate the effectiveness of our MeSH term recommendation approach, we perform a **retrospective assessment** using previously developed systematic review queries as a benchmark. These existing queries, which already contain manually curated MeSH terms, serve as the **gold standard**. However, it is important to note that this baseline may naturally favor terms generated through PubMed's Automatic Term Mapping (ATM), as ATM may have influenced the original query construction.

Our approach aims to identify additional potentially relevant MeSH terms that were not included in the original query fragments. To assess the value of these recommendations, we go beyond simple comparison with the original terms and evaluate how well the enhanced queries improve retrieval of relevant research. This dual evaluation allows us to question the assumption that the original MeSH terms are always the most effective.

Several limitations must be acknowledged in this evaluation. First, the query fragments must be **reassembled into complete Boolean queries** to allow meaningful performance comparisons. Second, our retrieval experiments may uncover relevant studies that were not part of the original relevance assessments, meaning they have not been formally judged. To address these challenges, we implement a **three-pronged evaluation strategy** that measures performance from multiple perspectives.

1.Lower bound: Assumes all unjudged studies are irrelevant.

2.Upper bound: Assumes all unjudged studies are relevant.

3.MLE-based estimate: Uses maximum likelihood estimation based on the judged studies to estimate how many unjudged studies might be relevant. This estimation is used to randomly sample unjudged documents likely to be relevant, based on the proportion of relevant documents in the original candidate pool.

We evaluate how well each of the three MeSH term retrieval methods (ATM, MetaMap, UMLS) performs using precision and recall. To measure the quality of MeSH term ranking produced by our learning-to-rank (LTR) models, we use normalized Discounted Cumulative Gain (nDCG) and reciprocal rank.

In evaluating the impact of suggested MeSH terms on systematic review searches, we reconstruct full Boolean queries from the refined query fragments and test their retrieval performance. For this, we apply standard information retrieval metrics, including precision, recall, and Fβ scores for β = {0.5, 1, 3}. Since the results for β = 0.5 and β = 3 follow the same patterns as β = 1, only the F1 scores are reported for clarity.

All retrieval experiments are executed via the PubMed Entrez API, with a fixed date range applied to ensure consistency and reproducibility across queries—important given the continuously evolving nature of the PubMed database. Finally, in both ranking and retrieval evaluations, we measure performance in two scenarios:

(i) using all retrieved MeSH terms, and

(ii) applying a score threshold to filter out lower-ranking MeSH suggestions.

Method	P	R	RR	R@5	R@10	nDCG@5	nDCG@10
2017/A	0.3027	0.3718	0.4614	0.3504	0.3576	0.3601	0.3494
2017/A-C	0.3600	0.2421*	0.4047	0.2362*	0.2403*	0.2713*	0.2652*
2017/M	0.3496	0.3818	0.5730	0.3659	0.3793	0.4218	0.4102
2017/M-C	0.4333	0.2868	0.4739	0.2800	0.2868	0.3105	0.3024
2017/U	0.2571	0.4659	0.5910	0.4214	0.4475	0.4518	0.4469
2017/U-C	**0.4819**	0.2846	0.5295	0.2831	0.2846	0.3446	0.3329
2017/F	0.2446	**0.5281**	**0.6207**	**0.4519**	**0.4958**	**0.4915**	**0.4971**
2017/F-C	0.4517	0.3458	0.4897	0.3391	0.3450	0.3581	0.3475
2018/A	0.3287	0.3772	0.4967	0.3261	0.3703	0.3719	0.3838
2018/A-C	0.3742	0.2129*	0.3933*	0.1983*	0.2024*	0.2417*	0.2392*
2018/M	0.3088	0.3360	0.4630	0.3007	0.3353	0.3470	0.3380
2018/M-C	0.3704	0.2257*	0.3940	0.2237	0.2257	0.2689	0.2523
2018/U	0.2885	0.4641	**0.6007**	0.4188	0.4565	0.4615	0.4569
2018/U-C	**0.4711**	0.2643	0.5278	0.2575	0.2643	0.3405	0.3292
2018/F	0.2661	**0.5024**	0.5793	**0.4316**	**0.4798**	**0.4633**	**0.4629**
2018/F-C	0.4212	0.3456	0.4754	0.3233	0.3387	0.3646	0.3550
2019/D/A	0.3399	0.3558	0.5933	0.3503	0.3558	0.3910	0.3671
2019/D/A-C	0.5071	0.2628	0.5583	0.2628	0.2628	0.3222	0.2990
2019/D/M	0.3864	0.3053	0.6167	0.3003	0.3053	0.3810	0.3530
2019/D/M-C	**0.5458**	0.2303	0.5417	0.2253	0.2303	0.3209	0.2963
2019/D/U	0.2651	0.4528	0.6058	0.4428	0.4478	0.4680	0.4348
2019/D/U-C	0.4850	0.2619	0.5167	0.2619	0.2619	0.3159	0.2866
2019/D/F	0.2266	**0.4778**	**0.6725**	**0.4622**	**0.4678**	**0.5000***	**0.4680***
2019/D/F-C	0.5068*	0.3419	0.4933	0.3419	0.3419	0.3618	0.3302
2019/I/A	0.3110	0.3631	0.4469	0.3379	0.3560	0.3409	0.3445
2019/I/A-C	0.3703	0.2195*	0.3868	0.2195*	0.2195*	0.2511*	0.2484*
2019/I/M	0.2651	0.3395	0.4253	0.3071	0.3368	0.3131	0.3205
2019/I/M-C	0.3368	0.2178	0.3682	0.2088	0.2178	0.2389	0.2370
2019/I/U	0.2779	0.4175	0.4340	0.3765	0.4114	0.3516	0.3663
2019/I/U-C	0.3415	0.2209	0.3769	0.2146	0.2209	0.2447	0.2464
2019/I/F	0.2566	**0.4462**	**0.5283**	**0.4120**	**0.4373**	**0.4190**	**0.4233**
2019/I/F-C	**0.4074**	0.3229	0.4336	0.3112	0.3166	0.3271	0.3255

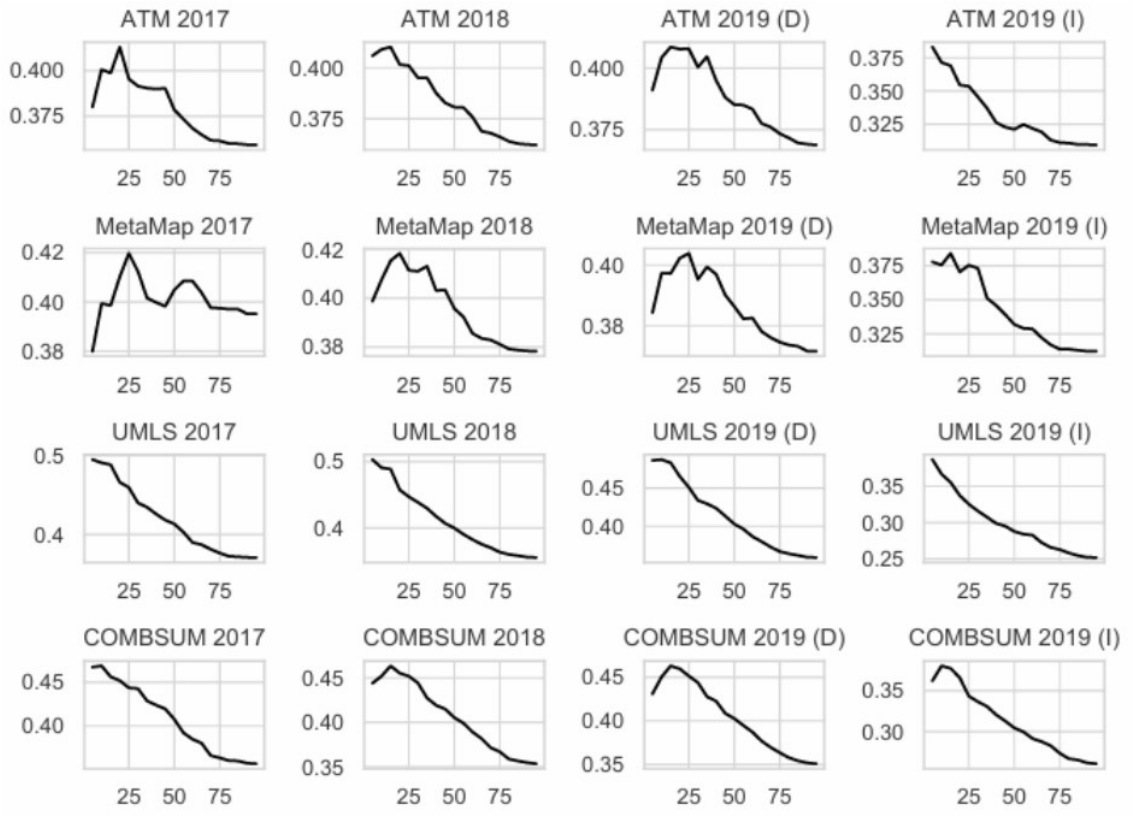
Table 2:Accuracy of MeSH Term Suggestion Methods Measured by Precision

Ranking of MeSH Terms

Next, we evaluate how effectively the **learning-to-rank (LTR) model** orders the MeSH terms retrieved by each method. The evaluation uses standard ranking metrics, including **reciprocal rank (RR)**, **Recall at rank k (R@k)**, and **normalized Discounted Cumulative Gain at rank k (nDCG@k)**, as summarized in Table 2.

Our findings highlight two key observations:

1. The **UMLS-based retrieval method** generally achieves higher recall compared to ATM and MetaMap, resulting in superior overall ranking performance across most metrics.
2. The **fusion strategy**, which combines rankings from all three methods, typically produces the most effective MeSH term ordering. The only exception is the reciprocal rank metric for the 2018 dataset, where the fusion approach performs slightly below UMLS alone.



Refinement of MeSH Terms

Finally, we assess the impact of applying a **ranking-based cut-off** to the retrieved MeSH terms, removing those that fall below a specified threshold. This threshold is defined by the parameter κ. Figure 2 illustrates the effect of tuning κ on the training datasets, with the x-axis showing different κ values and the y-axis representing the corresponding F1 scores.

The curves display noticeable spikes, likely caused by the inclusion or exclusion of groups of tied MeSH terms. These fluctuations are particularly evident for the MetaMap and ATM methods, where term scores are often less distinct. In contrast, the UMLS and fusion approaches produce smoother curves, reflecting clearer differentiation in term scores.

Table 2 (labeled with "-C") summarizes the impact of refinement on MeSH term recommendation performance. The results indicate that applying refinement generally **improves precision** but also leads to a **drop in recall**, which can reduce overall ranking quality. In many cases, using a refinement cut-off results in poorer performance compared to retaining all MeSH terms, and it is often less effective than using ATM without refinement.

The Impact of Fusion

In terms of **recall**, the fusion method without refinement generally achieved higher results, with the exception of one instance in the 2018 dataset. This improvement is likely due to the fusion process aggregating all MeSH terms suggested by ATM, MetaMap, and UMLS, increasing the coverage of potentially relevant terms. However, this straightforward fusion approach **did not improve precision** for Boolean queries, indicating that simply combining terms is not sufficient to enhance overall search accuracy.

These findings suggest that **semi-automatic MeSH term recommendations** can be valuable for information specialists, who can use their domain expertise to select the most relevant terms and optimize query performance. Interestingly, the **refined fusion strategy** did not outperform other methods on any metric or dataset. Applying refinement generally reduced recall while having minimal effect on precision across all methods. This highlights that, for effective systematic review searches, the **careful selection of appropriate MeSH terms** is more critical than increasing the total number of terms included.

4. Case Study

To further understand why the performance of MeSH term recommendation methods varies, we conducted a detailed **case study** using query **CD010542** from the 2017 CLEF TAR dataset as a representative example. Multiple queries generated by our MeSH suggestion techniques were reviewed, and this query was chosen for detailed analysis.

Following the workflow illustrated in Figure 1, we first removed all MeSH terms from the original query and divided it into individual fragments. For each fragment, MeSH term suggestions were generated, after which the query was reconstructed by integrating these recommendations into the Boolean structure. The reconstructed queries were then executed in PubMed, and their retrieval performance was compared with the **original query** and a **version without any MeSH terms**.

The query fragments and their corresponding MeSH term suggestions are presented in Table 4. Table 5 demonstrates that results for ATM and MetaMap were identical, despite originating from different query fragments. This reinforces the observation from Section 4.2.3: **the choice of MeSH terms is pivotal** to the effectiveness of Boolean queries in systematic review searches.

Table 3 summarizes the evaluation of MeSH term recommendations within Boolean queries for the systematic review search of CD010542. Here, **O** denotes the original query, **R** represents the version with MeSH terms removed, **A** corresponds to Automated Term Mapping (ATM), **M** to MetaMap, and **U** to the Unified Medical Language System (UMLS).

Method	P	P (MLE)	P (Opt)	F1	F1 (MLE)	F1 (Opt)	R	R (MLE)	R (Opt)
O	0.0207	0.0598	0.6622	0.0405	0.1126	0.7968	0.9000	1.0000	1.0000
R	0.0274	0.0608	0.6140	0.0531	0.1143	0.7594	0.9000	0.9524	0.9951
A	0.0167	0.0583	0.7574	0.0327	0.1100	0.8611	0.9000	0.9692	0.9976
M	0.0167	0.0583	0.7574	0.0327	0.1100	0.8611	0.9000	0.9692	0.9976
U	0.0256	0.0598	0.6382	0.0499	0.1126	0.7778	0.9000	0.9545	0.9956

When comparing **UMLS** with other MeSH term recommendation methods, we observed that UMLS generally achieved **higher precision** than the other approaches, except under optimistic evaluation metrics. This finding suggests that certain MeSH terms—such as the one included in fragment 1 of the original query—may actually **reduce query effectiveness**, demonstrating that the manually selected MeSH terms are not always optimal. Interestingly, the

version of the query **without any MeSH terms** outperformed all other methods in terms of precision while maintaining similar recall. This indicates that MeSH terms do not universally improve search results and that a **more flexible approach** for selecting the most relevant MeSH terms could enhance the overall effectiveness of recommendation methods.

Rewritten Version (Same Research Meaning)

To address this challenge, we propose two complementary strategies:

1. **Fully Automated Strategy:** A classification model is developed to assess the usefulness of individual MeSH terms, together with a stopping criterion that identifies the optimal point at which adding additional terms no longer improves query effectiveness.
2. **Semi-Automated Strategy:** The classification model is applied iteratively to assign confidence scores to candidate MeSH terms, offering decision support to information specialists as they select the most relevant terms during the query formulation process.

5. Conclusion

This study presents a novel framework for recommending MeSH terms to strengthen Boolean search strategies used in systematic review literature searches. We performed a comprehensive evaluation of multiple MeSH recommendation techniques, examining their effectiveness across retrieval, ranking, and refinement stages, and benchmarking their performance against PubMed's Automatic Term Mapping (ATM) system.

The results demonstrate that both MetaMap- and UMLS-based approaches enhance Boolean query performance by generating more relevant MeSH term suggestions than ATM alone. The greatest improvements were observed when all three methods were combined using rank fusion, which capitalized on the complementary strengths of each approach and mitigated several semantic limitations inherent in ATM. Although these improvements occasionally resulted in a modest decrease in recall, the reduction was minimal and unlikely to affect the overall outcomes of systematic reviews.

Identifying appropriate MeSH terms for Boolean queries remains a complex task, even for experienced information professionals, particularly in the context of systematic reviews. The findings of this research provide meaningful contributions to both the information retrieval and systematic review communities. In addition, the proposed methods show promise for broader applications in automated query construction tasks, such as those explored in CLEF TAR, and could be integrated into existing search tools to support information specialists in developing more accurate and effective literature search strategies.

References

1. Hliaoutakis, Angelos. "Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline." *Master's thesis* (2005).

2. Hassan, Sahar, Franck Hétroy, and Olivier Palombi. "Ontology-guided mesh segmentation." *FOCUS K3D Conference on Semantic 3D Media and Content*. 2010.

3. Díaz-Galiano, Manuel Carlos, et al. "Integrating mesh ontology to improve medical information retrieval." *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8*. Springer Berlin Heidelberg, 2008.

4. Soualmia, Lina Fatima, Christine Golbreich, and Stéfan Jacques Darmoni. "Representing the MeSH in OWL: Towards a semi-automatic migration." *KR-MED*. Vol. 102. 2004.

5. Osborne, John D., et al. "Interpreting microarray results with gene ontology and MeSH." *Microarray Data Analysis: Methods and Applications* (2007): 223-241.

6. Elberrichi, Zakaria, Belaggoun Amel, and Taibi Malika. "Medical Documents

Classification Based on the Domain Ontology MeSH." *arXiv preprint arXiv:1207.0446* (2012).

7. Elberrichi, Zakaria, Malika Taibi, and Amel Belaggoun. "Multilingual medical documents classification based on mesh domain ontology." arXiv preprint arXiv:1206.4883 (2012).