Metadata Driven Semantically Aware Medical Query Expansion

Rituraj Ojha[1](✉) and Gerard Deepak[2]

[1] Department of Metallurgical and Materials Engineering, National Institute of Technology, Tiruchirappalli, Tiruchirappalli, Tamil Nadu, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tiruchirappalli, Tamil Nadu, India

Abstract. The query used to retrieve information related to the medical domain may not contain the technical terms which are used in the medical industry. The user query should include more relevant terms and therefore, query expansion technique is required in the medical domain for their Information Retrieval Systems. In this paper, a metadata driven semantically aware medical query expansion methodology is proposed. The proposed approach takes a query as an input which is preprocessed and then Latent Semantic Indexing is used to generate new topics for each query word. A set of ontologies of PubMed keywords are semantically aligned using Lesk similarity and Normalized Pointwise Mutual Information. A Knowledge Tree is formed which is used to classify the metadata generated from Google Books using Recurrent Neural Networks. Finally, the terms from the Knowledge Tree are enriched using Wikidata, CASNET, and Hepatitis Knowledge Base, and are semantically integrated with 25% of the classified metadata using Normalized Pointwise Mutual Information under the Social Spider algorithm. The proposed MDSA-MQE methodology achieves the Precision of 90.12%, Recall of 93.87%, Accuracy of 92.08%, F-Measure of 91.95%, and Normalized Discounted Cumulative Gain value of 0.94 making it a better approach than the baseline approaches.

Keywords: Entity enrichment · Information retrieval · Medical query expansion · Metadata

1 Introduction

The Information Retrieval Systems are software programs used for finding the relevant information or documents that the user needs. A web search engine is an excellent example of the Information Retrieval System. Information on the internet in the medical field is increasing at a fast rate. A proper Information Retrieval System is required which fetches relevant medical documents as per the user's query.

However, it is very tough to access this large amount of increasing data. Furthermore, the user query lacks the essential technical terms that are contained

B. Villazón-Terrazas et al. (Eds.): KGSWC 2021, CCIS 1459, pp. 223–233, 2021.
https://doi.org/10.1007/978-3-030-91305-2_17

in the relevant document set. As a result, the user might not receive the required medical information that he needs. Hence, a proper Information Retrieval System is required, which uses a query expansion methodology to enhance the user query with more technical and essential terms for fetching the relevant document set that is most suitable.

Motivation: Medical terms are highly technical, and their technical interpretation is very challenging for Information Systems, and therefore there is a necessity for a better understanding of medical terminology. When a system is queried with medical terminologies, a better query understanding or better query expansion recommendation system is required. Also, there are very few systems in this domain that return results with high accuracy. Therefore, building this system having higher accuracy than the existing approaches was a major motivating factor.

Contribution: A metadata driven semantically aware medical query expansion system is proposed. The input queries by users related to the medical domain drives the proposed framework. The query is preprocessed and Latent Semantic Indexing is used to generate new topics for each query word. A set of ontologies of PubMed keywords are semantically aligned using Lesk Similarity and Normalized Pointwise Mutual Information (NPMI). A Knowledge Tree is formed which is used to classify the metadata generated from Google Books using Recurrent Neural Networks (RNN). Finally, the terms from the Knowledge Tree are enriched using Wikidata, CASNET, and Hepatitis Knowledge Base and semantically integrated with 25% of the classified metadata using NPMI under the Social Spider algorithm. The values of metrics like, Precision, Accuracy, Recall, nDCG, and F-Measure of the proposed methodology are improved.

Organization: The flow of the remaining paper is as follows. A condensed summary of the related works is provided in Sect. 2. Section 3 depicts the architecture of the proposed system. Section 4 describes the architecture implementation. Section 5 presents the evaluation of results and performance. The conclusion of the paper is presented in Sect. 6.

2 Related Works

Arbabi et al. [1] have proposed a machine learning model which automatically recognizes medical concepts in a large unstructured text. They have proposed a neural dictionary model, called Neural Concept Recognizer (NCR), which can identify the synonyms which were unobserved previously, by using biomedical ontology. Kim et al. [2] have proposed a healthcare context information system based on ontology. The proposed model helps in providing the customized service environments by extracting and classifying the contextual information.

Yunzhi et al. [3] have presented a technique for the expansion of medical query which is based on Hepatitis ontology. Their methodology uses the semantic approach to retrieve better results by expanding the query with synonyms, hypernym and similar query related technical words. Gao et al. [4] have proposed a methodology for automatic expansion of query which will be used for retrieving relevant BIM resources from web. They have proposed a search engine using IFC IR ontology.

Kuzi et al. [5] have presented an approach for expanding query using word embeddings. Their approach expands the given query with the words which are semantically related using Word2Vec's CBOW embedding technique. Oh and Jung [6] have proposed a query expansion technique using several collections from external sources. Their approach uses pseudo-relevance feedback technique.

Keikha et al. [7] have proposed a supervised and unsupervised query expansion technique which uses pseudo-relevance feedback approach. In this approach, Wikipedia articles are extracted which are highly related to the user query. Dahir et al. [8] have proposed a methodology for the extraction of candidate terms for query expansion. Their methodology is based on DBpedia and WordNet.

Jain et al. [9] have proposed an approach to retrieve medical records using semantic query expansion technique for helping patient with their current symptoms. The proposed Electronic Medical Record (EMR) retrieval system uses domain ontologies, automatic sematic relationship learning, information retrieval, and domain knowledge from professionals. Raza et al. [10] have surveyed and presented the recent approaches and models for semantic query expansion. They have discussed strength and weakness of each model and organized them into a taxonomy. The papers [11] and [12] have proposed several approaches based on ontologies and machine learning models.

Mourao et al. [13] have proposed a medical IR system which supports image and text queries. The system can retrieve relevant PubMed articles and can expand the query using MeSH thesaurus. Hanauer et al. [14] reports about an IR system, namely EMERSE, which helps in retrieving information stored in clinical notes in EHRs.

From the literature analysis, it is inferable that either the models are based on traditional machine learning paradigm with static Ontologies or they use standard knowledge sources, like DBPedia and WordNet. Certain approaches have used semantic techniques alone. Word2Vec with clustering mechanisms are also incorporated along with a learning of relations. However, some of the models extract features from the dataset alone and do not focus on the standard knowledge stores. In some of the approaches, either a medical database which is highly domain specific or secondary modeled thesaurus has been used. It is to be noted that a combination of machine learning with semantics is quite rare in query expansion. However, encompassment of standard knowledge stores and specialized medical extracts infused with the highly stringent semantics with learning and inferencing is the need of the hour in medical query expansion which has been targeted in the proposed model.

3 Proposed System Architecture

The architecture for query expansion is depicted in Fig. 1. The proposed approach takes place in several steps. Initially, the query is taken as an input and is preprocessed using Tokenization, Lemmatization, stop word removal, and Named Entity Recognition (NER). Tokenization is the process of breaking the texts into small pieces called tokens. Byte Pair Encoding (BPE) is used for the tokenization process. Lemmatization involves grouping together several inflected kinds of the same word so that they are analyzed as one term. During stop word removal, the common ubiquitous words are removed as they add no value for the analysis and only increase the dimension of the feature set. The NLTK python library is used for stop word removal. NER is the process of finding and categorizing the data or entity into predefined categories. After the preprocessing stage, a Query word set is obtained. In the next step, Latent Semantic Indexing is used to generate new topics for each term in the query word set. Latent Semantic Indexing is the method for analyzing documents and producing a set of topics related to that document.

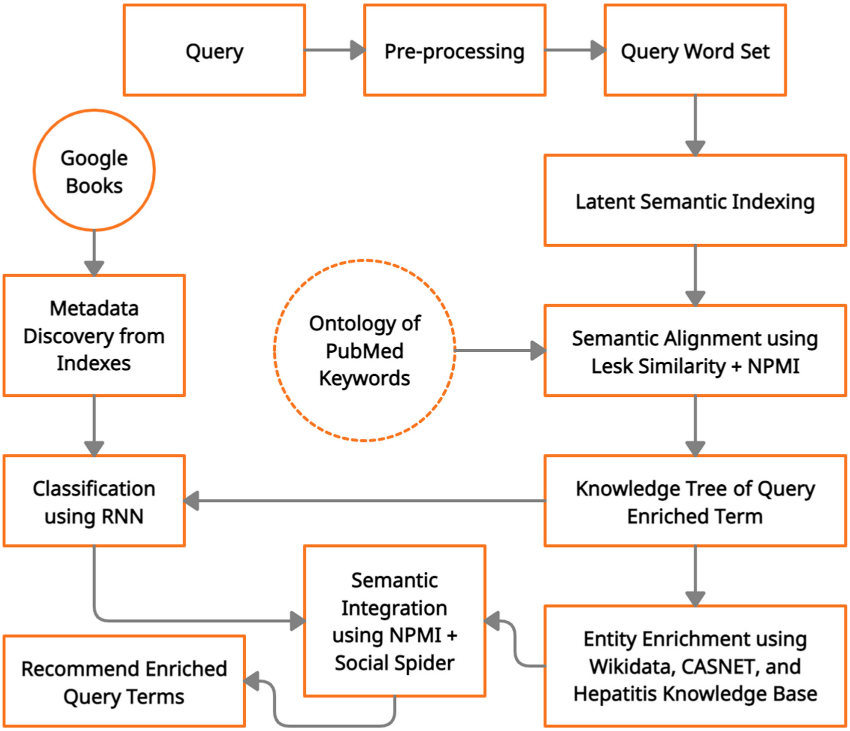
Query
Pre-processing
Query Word Set
Google Books
Latent Semantic Indexing
Metadata Discovery from Indexes
Ontology of PubMed Keywords
Semantic Alignment using Lesk Similarity + NPMI
Classification using RNN
Knowledge Tree of Query Enriched Term
Semantic Integration using NPMI + Social Spider
Recommend Enriched Query Terms
Entity Enrichment using Wikidata, CASNET, and Hepatitis Knowledge Base

Fig. 1. Proposed system architecture

Now, the ontology of PubMed keywords which has been modeled and generated is used for further decision making and knowledge inclusion into the framework. PubMed is used for accessing the MEDLINE database containing data on life sciences and biomedical. For every word in the query word set coming from

Latent Semantic Indexing, a set of ontologies of PubMed keywords are aligned. The Semantic alignment is achieved by computing the semantic similarity using Lesk Similarity [15] and NPMI measure. The Lesk Similarity measure assumes that words in the same neighborhood will share a common topic. It is mainly used for word sense disambiguation. The NPMI is the normalized form of PMI (Pointwise mutual information) measure. The value for NPMI ranges between the bounds −1 and 1 that assists the process of selective filtering and separates the items relevant to query from irrelevant items [16].

$$pmi(x; y) = J(x) + J(y) - J(x, y) \tag{1}$$

$$npmi(x; y) = \frac{pmi(x; y)}{J(x, y)} \tag{2}$$

Equation (1) represents the Pointwise Mutual Information. The J(x) represents the self-information, or $-\log_2 p(X = x)$. Equation (2) represents Normalized Pointwise Mutual Information. The J(x, y) represents joint self-information, which is equal to $-\log_2 p(X = x, Y = y)$.

Furthermore, the Knowledge Tree of the query enriched terms is formulated by computing the semantic similarity using Lesk and NPMI measures between the instances in the Ontology and the Latent Semantic Indexing enriched initial query words. Now, taking the tree as the input, we will classify using the RNN. From Google Books, the metadata is generated from indexes using OpenCalais and RDF Distiller. The top 25% of the classified results are retained for Semantic Integration. For terms coming from Knowledge Tree, entity enrichment is performed using Wikidata, CASNET, and Hepatitis Knowledge Base. The two medical knowledge bases, namely the CASNET and Hepatitis Knowledge Base are accessed using Python APIs. Wikidata will be queried using SPARQL endpoints. A CASNET is a semantic net, a knowledge structure consisting of nodes representing the concepts, characteristics, events, and branches specifying the relationships between nodes. It helps in treatment of Glaucoma.

Now, all the enriched entities along with the Knowledge Tree have been integrated with the 25% of the classified contents from the medical books from Google Books. The Semantic Integration is done using NPMI under Social Spider algorithm, which is a heuristic algorithm. It was created by imitating the behavior of spiders in nature. After all these steps, the enriched query terms are recommended by the system.

4 Implementation

The algorithm for the proposed technique is depicted in Table 1. The dataset used for this approach is the STS Benchmark dataset (https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark). It contains a collection of the English datasets and there are about 8628 sentence pairs contained in the benchmark. The input taken is the query along with google books and ontology of PubMed keywords. The algorithm returns the recommended enriched query words as output.

Algorithm 1. Proposed System Architecture

```
Input: Query, Google Books, Ontology of PubMed Keywords
Output: Enriched Query Terms
Begin
Step 1: The Query Q is subjected to query pre-processing.
    Q is tokenized, lemmatized, and NER and
    stop word removal is performed to obtain query word set Qw.
Step 2: while (Qw.next()!=NULL)
    for each Qw as label
        Qw ← Latent Semantic Indexing Generated Topics
    end for
    end while
Step 3: for each Qw
    Semantic alignment of set of ontologies of PubMed Keywords
    end for
Step 4: Knowledge Tree will be formed
Step 5: 5.1: Metadata generation using Google Books
    5.2: Classifying metadata taking knowledge tree as input using RNN
Step 6: Qw ← entity enrichment using Wikidata, CASNET, and Hepatitis
    knowledge base
Step 7: SemanticIntegration(Top 25% of classified contents from Google Books,
    Enriched entities from Knowledge Tree)
Step 8: Recommend enriched Query Terms
End
```

5 Results and Performance Evaluation

The performance of the proposed MDSA-MQE (Metadata Driven Semantically Aware Medical Query Expansion) approach is measured by considering Precision, Recall and Accuracy. Other measures including F-Measure and Normalized-Discounted Cumulative Gain (nDCG) are also used. The performance is evaluated for 6871 queries and the ground truth has been collected.

$$Precision = \frac{Retrieved \cap Relevant}{Retrieved} \tag{3}$$

$$Recall = \frac{Retrieved \cap Relevant}{Relevant} \tag{4}$$

$$Accuracy = \frac{ProportionCorrectsofeachQueryPassedtheGroundTruthTest}{TotalNumberofQueries} \tag{5}$$

$$F - Measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{6}$$

$$nDCG = \frac{DCG_\alpha}{IDCG_\alpha} \tag{7}$$

$$DCG = \sum_{i=1}^{\infty} \frac{rel_i}{\log_2(i+1)} \tag{8}$$

Equations (3), (4), and (5) represent Precision, Recall and Accuracy, respectively. Furthermore, the Eqs. (6), (7), and (8) represent F-Measure, nDCG and Discounted cumulative gain, respectively. The reason for considering these above metrics for evaluation is because they measure the relevance of the results. The Knowledge Base is extensively vast and, in this case, it is primarily used for entity population and increasing the density of the incoming auxiliary knowledge into the framework. The Recall computation is carried out with respect to the dataset and not with respect to the Knowledge Base. The number of relevant documents in the dataset is already known for each query which has been tested, as the ground truth collection was carried out by a simple voting mechanism by 714 candidates who were domain experts and top 25 relevant documents for each query was taken. However, 20 coherent relevant documents, based on the maximum voting, were shortlisted and considered as relevant for each subsequent query. The queries were distributed such that each candidate had to answer 20 queries and top 20 relevant documents from the dataset had to be recommended by them. Instead of using a query tool, a human-in-the-middle approach was followed for the collection of ground truth based on the documents in the dataset. For the query expansion, the keywords in the documents based on the frequency of occurrences and the uniqueness of terms were considered to be relevant.

Table 1. Performance comparison of the proposed MDSA-MQE with other approaches

Search technique	Average precision %	Average recall %	Average accuracy %	Average F-measure %	nDCG
MQES [19]	78.32	81.79	79.44	80.02	0.75
CRDESLM-QET [17]	74.36	77.12	75.69	75.71	0.74
QEMO [18]	83.78	87.92	85.56	85.80	0.91
Proposed MDSA-MQE	90.12	93.87	92.08	91.95	0.94

Table 1 represents the performance comparison of the proposed MDSA-MQE with other approaches. It is evident from the table that the proposed approach achieves the highest average accuracy with the Precision of 90.12%, Recall of 93.87%, Accuracy of 92.08%, F-Measure of 91.95%, and nDCG value of 0.94.

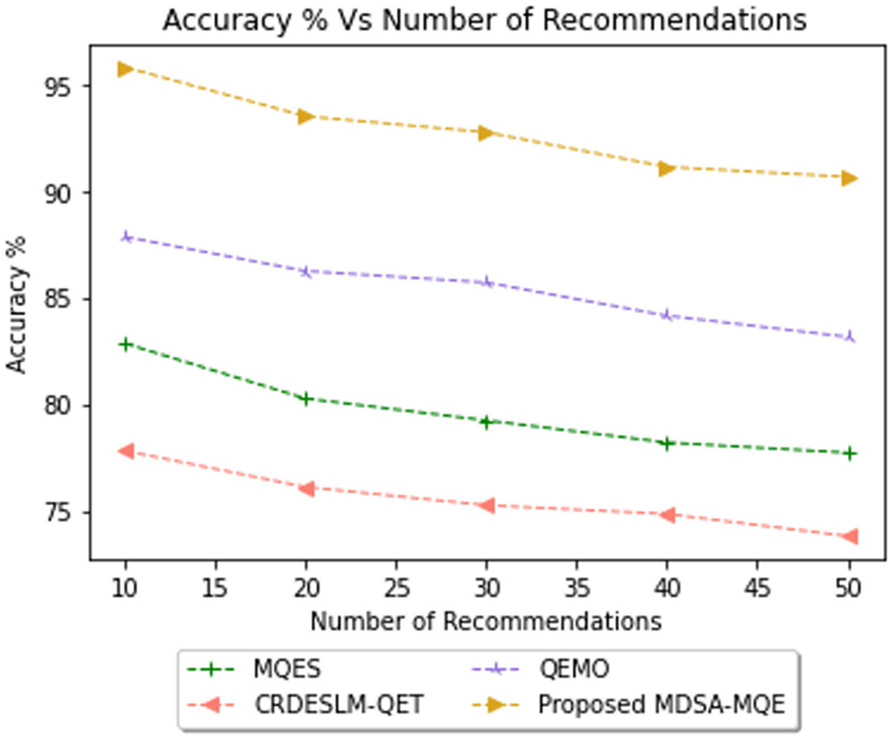

Fig. 2. Accuracy % vs number of recommendations

Figure 2 represents the Accuracy % vs Number of Recommendations of the proposed approach. The proposed MDSA-MQE approach is better than the other approaches because of several reasons. MQES [19] uses DBPedia and Wikidata as standard knowledge sources but for computing the relevance, it only uses Kullback–Leibler Divergence. As a result, MQES [19] achieves less accuracy than the proposed MDSA-MQE model. Generally, the greater the number of relevant knowledge bases, greater is the probability of relevant knowledge fed into the approach, thereby increasing the density of auxiliary knowledge and improving the overall accuracy of the approach. The CRDESLM-QET [17] uses a clustering method which is not as good as the proposed approach. In the QEMO [18] approach, they have used MeSH medical ontology to enrich the query with medical terms. Modelling ontology is highly complicated and depending on only modelled ontology at every instance of time is not the right strategy since every time ontology cannot be modelled. Also, a bag of terms has been used which is unsuitable when a lot of external knowledge sources are to be integrated into the model. Therefore, the density of enriched entities will be quite low in this approach. The proposed MDSA-MQE approach achieves a high density of entities because there are many knowledge sources. New topics are generated using Latent Semantic Indexing and metadata coming from Google Books is classified and semantically integrated with the query set. Also, ontology of PubMed keywords is used for better accuracy. Enrichment of the Knowledge Tree takes place using Wikidata, CASNET, and Hepatitis Knowledge Base. Hence, because of these reasons, the proposed MDSA-MQE approach is much better than the existing models.

Table 2 presents two examples from several of the expanded queries produced by the proposed model. Only the top 8 expanded queries results are displayed here.

Table 2. Examples of query expansion by proposed model

Model medical queries	Query expansion by proposed MDSA-MQE
Vitamin D status	1. Vitamin D status during gestation 2. VItamin D status during infancy 3. Vitamin D status of children 4. Vitamin D status of adults 5. Vitamin D status in geriatric patients 6. Vitamin D status in men 7. Vitamin D status in women after menopause 8. Vitamin D status in first trimester pregnancy
Oral Ciprofloxacin is the first line of antibiotic	1. Oral Ciprofloxacin is the first line of antibiotic for urinary tract infection 2. Oral Ciprofloxacin is the first line of antibiotic for gonorrhea 3. Oral Ciprofloxacin is the first line of antibiotic for chancroid 4. Oral Ciprofloxacin is the first line of antibiotic for skin 5. Oral Ciprofloxacin is the first line of antibiotic for bone 6. Oral Ciprofloxacin is the first line of antibiotic for joint infections 7. Oral Ciprofloxacin is the first line of antibiotic for typhoid fever 8. Oral Ciprofloxacin is the first line of antibiotic for plague

6 Conclusion

There is much necessity for satisfying the users by providing relevant information pertaining to medical areas. There is no medical query expansion technique which can enrich the user query with high accuracy. Therefore, the MDSA-MQE is proposed which takes queries as input. The query is preprocessed and Latent Semantic Indexing is used to generate new topics for each query word. A set of ontologies of PubMed keywords are semantically aligned using Lesk similarity and NPMI. A Knowledge Tree is formed which is used to classify the metadata generated from Google Books using RNN. Finally, the terms from the Knowledge Tree are enriched using Wikidata, CASNET, and Hepatitis Knowledge Base and semantically integrated with 25% of the classified metadata using NPMI under the Social Spider algorithm. The proposed methodology achieves the Precision of 90.12%, Recall of 93.87%, Accuracy of 92.08%, F-Measure of 91.95%, and nDCG value of 0.94. The overall accuracy of the proposed methodology is much better than the existing approaches.

References

1. Arbabi, A., Adams, D.R., Fidler, S., Brudno, M.: Identifying clinical terms in medical text using ontology-guided machine learning. JMIR Med. Inf. **7**(2), e12596 (2019)
2. Kim, J., Chung, K.-Y.: Ontology-based healthcare context information model to implement ubiquitous environment. Multimed. Tools Appl. **71**(2), 873–888 (2011). https://doi.org/10.1007/s11042-011-0919-6
3. Yunzhi, C., Huijuan, L., Shapiro, L., Travillian, R.S., Lanjuan, L.: An approach to semantic query expansion system based on Hepatitis ontology. J. Biol. Res.-Thessaloniki **23**(1), 11–22 (2016)
4. Gao, G., Liu, Y.S., Wang, M., Gu, M., Yong, J.H.: A query expansion method for retrieving online BIM resources based on industry foundation classes. Autom. Constr. **56**, 14–25 (2015)
5. Kuzi, S., Shtok, A., Kurland, O.: Query expansion using word embeddings. In: Proceedings of the 25th ACM international on Conference on Information and Knowledge Management, pp. 1929–1932, October 2016
6. Oh, H.S., Jung, Y.: Cluster-based query expansion using external collections in medical information retrieval. J. Biomed. Inform. **58**, 70–79 (2015)
7. Keikha, A., Ensan, F., Bagheri, E.: Query expansion using pseudo relevance feedback on Wikipedia. J. Intell. Inf. Syst. **50**(3), 455–478 (2017). https://doi.org/10.1007/s10844-017-0466-3
8. Dahir, S., Khalifi, H., El Qadi, A.: Query expansion using DBpedia and WordNet. In: Proceedings of the ArabWIC 6th Annual International Conference Research Track, pp. 1–6, March 2019
9. Jain, H., Thao, C., Zhao, H.: Enhancing electronic medical record retrieval through semantic query expansion. Inf. Syst. e-Bus. Manag. **10**(2), 165–181 (2012)
10. Raza, M.A., Mokhtar, R., Ahmad, N., Pasha, M., Pasha, U.: A taxonomy and survey of semantic approaches for query expansion. IEEE Access **7**, 17823–17833 (2019)
11. Panchal, R., Swaminarayan, P., Tiwari, S., Ortiz-Rodriguez, F.: AISHE-onto: a semantic model for public higher education universities. In DG. O2021: The 22nd Annual International Conference on Digital Government Research, pp. 545–547, June 2021
12. Gaurav, D., Rodriguez, F.O., Tiwari, S., Jabbar, M.A.: Review of machine learning approach for drug development process. In: Deep Learning in Biomedical and Health Informatics, pp. 53–77. CRC Press (2021)

13. Mourão, A., Martins, F., Magalhaes, J.: Multimodal medical information retrieval with unsupervised rank fusion. Comput. Med. Imaging Graph. **39**, 35–45 (2015)
14. Hanauer, D.A., Mei, Q., Law, J., Khanna, R., Zheng, K.: Supporting information retrieval from electronic health records: a report of university of Michigan's nine-year experience in developing and using the electronic medical record search engine (EMERSE). J. Biomed. Inform. **55**, 290–300 (2015)
15. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation, pp. 24–26 (SIGDOC 1986). Association for Computing Machinery, New York (1986). https://doi.org/10.1145/318723.318728
16. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL, pp. 31–40 (2009)
17. Keyvanpour, M., Serpush, F.: ESLMT: a new clustering method for biomedical document retrieval. Biomed. Eng./Biomedizinische Tech. **64**(6), 729–741 (2019)
18. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput. Biol. Med. **39**(4), 396–403 (2009)
19. Dahir, S., El Qadi, A., ElHassouni, J., Bennis, H.: Medical query expansion using semantic sources DBpedia and Wikidata. In: ISIC 2021: International Semantic Intelligence Conference, ISIC 2021, 2019 (2021)