ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(14), DOI: 10.17485/ijst/2016/v9i14/77644, April 2016* ISSN (Online) : 0974-5645

Medical Query Expansion using UMLS

K. Saravana Kumar* and K. Deepa

VIT University, Vellore - 632014, Tamil Nadu, India; ksaravanakumar@vit.ac.in, deepa.k@vit.ac.in

Abstract

Internet users have grown in recent years and they demand answers for many through online. Searching and retrieving documents is one of the most frequent thing most of the people do today. For retrieving medical related documents, we have to be extra careful and precise in the process of retrieval. Even though separate search engines are there to retrieve medical documents with care, the users are not well-known with MeSH terms (Medical Subject Heading). MeSH terms are terms used by medical professionals to retrieve documents. So the query has to be framed in such a way that only correct documents should be produced to the user. In this work, we proposed a method that deals with enriching the user query with the use of UMLS. The enriched user query after expansion is used to get accurate documents with less amount of time.

Keywords: Information Retrieval, MeSH Terms, Query Expansion

1. Introduction

Information retrieval is the method of finding appropriate information from collected information resources. The common duty of Information Retrieval (IR) is searching for relevant information in documents. Everyone has started to investigate information on-line which utilizes a smaller amount time and attempt. Medical related information retrieval has been gradually increased Medical information retrieval is the method of retrieving information based on the medical matter inquired by the user. According to the examination done by the association named Jupiter, 71% of people utilized the Internet to for searching health related data in the year 2007. This is a significant increase when compared with the data for the year 2005. Around 160 million people have started using Internet for health related information serach in US alone[1]. In an examination conducted by Pew Internet Project in the year of 2009, 83% of Internet consumers have looked for medical or health information. The percentage of internet users searching for health and related information rises to around 80% in United States and considerably increased worldwide[2,3]. The search is mostly associated with syndrome, information about doctors, hospitals and diet. Queries are particularly about particular diseases or medicinal problems to investigational medicine and treatments. Women frequently investigate for health related information for somebody connected to them while men investigate for medical information for friends. PubMed is a free catalog accessing mainly the MEDLINE database of references and abstracts on medical sciences and life science issues. Strong characteristic of PubMed is its capacity to mechanically connect the words to MeSH terms and subtitles. For example: "heart attack" connects to "myocardial infarction" where suitable MeSH terms are automatically prolonged which are more exact in medical term. This vital characteristic makes PubMed hunt automatically more susceptible and it keep us away from false information by balancing for the multiplicity of medical terms. NLM (National Library of Medicine) at the National Institutes of Health preserves the catalog as an element of the Entrez system of information retrieval. The consumers of PubMed are both medical and non-medical professionals. In case of medical users it is pretty uncomplicated for them to execute search since they have some knowledge about the medical terms. Formerly the database enclosed items opening from 1965, but this has been improved, and reports that are old also now accessible within the chief index. The database is liberally available on the Internet through the PubMed border and new credentials are appended from Tuesday to Saturday. This is how the database is maintained. This database needs MeSH terms to retrieve the information. It is suggested in many literature related to Medical Information Retrieval

* *Author for correspondence*

that the use of MeSH terms helps in understanding both query and the target document set[12]. When the user gives the query without medical term it takes time to retrieve documents. If medical terms are used in the query it takes less time to retrieve. The query enrichment is done to represent the query more meaningful. The original query is appended with the enriched query for better result. More than 5,500 biomedical papers are directed in MEDLINE. New papers are not incorporated robotically or instantaneously. Selection is based on the suggestion of a board called the Literature Selection Technical Review Committee, based on logical span and value of a paper.

2. Proposed System Architecture

Figure 1 shows the architecture of our proposed model. It consists of the major components, viz., query pre-processing, and scenario specific query expansion using UMLS.

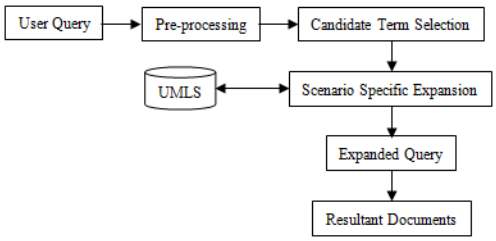

Figure 1. Medical IR using UMLS – architecture.

2.1 Preprocessing

Query preprocessing is the first and foremost step in information retrieval. The query given by the user has to be preprocessed before it starts to search for the information. Because the query given by the user may not be sufficient for search or it may have irrelevant information. Query is preprocessed to achieve various things like removal of redundant terms, quality improvement of a user query, increase in the standards of the result set and speeding up solution. Data that has to be searched may be incomplete. To make it complete some tasks has to be performed on the data such as cleaning, integration, transformation, reduction and discretization of data. Data cleaning is also done for fixing the missing values in a query. In most of the search queries, users failed to represent their requirement exactly as needed[4]. This may lead to wrong information identification or the search result may miss some right documents[5]. Malfunction of the equipment used for recording data, data inconsistency, and data misunderstanding are some of the causes of missing the data. Noisy data is also to be eliminated. Clustering is used to remove the noisy data. We need data cleaning while collecting data from various data sources and put the data in a different schema table with necessary data transformation. The requirement for the process of data cleaning increases when many data sources have to be combined and this is absolutely due to the redundant nature of heterogeneous data. To give access to precise data we need to consolidate dissimilar data and to remove duplicate information. The main thing in data cleaning is identifying the duplicate data and eliminating it. It is mostly done in AJAX framework in which data cleaning program is represented as directed graph that starts with the input that is given by the user and returns the clean data[6]. It is proved that AJAX is a widely used data cleaning tool[7]. Data cleaning process involves various stages like analysis, transformation rule development, transformation of data and backflow of cleaned data[8]. The process called data analysis is needed to identify an eliminate different errors and data inconsistencies. Analysis programs should be used to include the manual examination of different data samples, to increase the information regarding the data properties and to identify the quality related problems. Semantic integration is done to analyze the data and clean it[9].

Duplicate elimination is done after data transformation. Data transformation is done either by executing the workflow of ETL for loading and reviewing the data in the data warehouse or while answering multiple questions related to the data from multiple sources. After the elimination of errors the data that is cleaned has to be restored with the original query in order to provide accurate results. This is done in back flow of cleaned data phase. It avoids cleaning of data for future extractions. Many tools are available to perform data transformation and data cleaning. It is not easy to integrate the operations of many tools together[10]. Schema matching process[11] can be used to transform data from one schema into the other makes the data transformation process easier.

Data integration combines multiple databases and files. This is about merging of data from different sources and giving an integrated view of the combined data to the users. Data integration becomes important in a mixture of situations that contain both scientific and commercial domains. A theory of data integration is provided[24]. The source data is transformed into different format through the data transformation process. There are two steps in data transformation. They are data mapping and code

generation. Data mapping is used to map the data from source to destination and captures the transformation that happens during mapping. The real transformation program is created by code generation. Mapping is quite difficult because it requires one to one or many to many mapping transformation rules. Code generation produces an executable program in Java or XSLT. In another form of data transformation, the data in the database is transformed without removal from the database. This is called master data recast. In a well structured database, all data are connected to a restricted set of tables through a foreign key network directly or indirectly. The foreign keys in a foreign key network depends on single index from parent table. There are multiple languages that are used in data transformation. Normally data transformation requires a grammar to be followed. Some of the languages used are PERL, AWK, XSLT, TXL etc. Cassidy illustrated the usage of java language in his work[13]. Author in[14] used the TXL language in his work[1]. DMS software engineering toolkit is capable to solve the problem of data transformation[15]. Normally data warehouse consists of enormous amount of complex data. So data mining takes more time to run the complete set of data.

Data reduction is the process in which it obtains only a reduced data set which is very small in size but provides the same result. There are two steps in data reduction. They are feature selection and feature reduction. There are many data reduction strategies. Few of them are working by aggre gating data cubes, reducing data dimensions, generating hierarchies and discretizing data. The best algorithm for feature selection is correlation – based feature selection[16]. After feature selection, the feature extraction starts. Feature extraction begins from an preliminary set of calculated data and constructs resulting values planned to be informative, non redundant, helping the successive knowledge and simplification steps, in a few cases heading to improved human understandings. Normally, data cube aggregation is used when queries occur with aggregate functions[17]. Dimensionality reduction is the process of sinking the amount of random values under certain conditions. Dimensionality reduction is broadly divided into feature selection and feature extraction. Feature selection is used to find a subset of the original attributes. Feature selection can also be considered as selection of variables, attributes or subset of attributes. Redundant or repeated set of features supply the same set of features that we have. Feature selection provides three main profits. They are shorter training periods, improved generalization by sinking over fitting, enhanced model interpretability. A feature selection algorithm is the mixture of a search technique for suggesting new feature subsets, besides with an assessment measure that attains the dissimilar feature subsets. The easiest procedure is to examine every feature subset and choosing the feature that shown less error rate among the feature subsets. The assessment metrics that are chosen will influence this procedure greatly. Based on these assessment metrics we categorize the feature selection algorithms such as wrappers, filters and embedded methods[18].

2.2 Candidate Term Selection

Candidate term is the root word that has to be derived from the query. The root word is derived and selected by using stemming algorithms. Stemming is the process used in Information Retrieval and Linguistic Analysis to illustrate the development for sinking the words to their stem or root form. The stem words need not be the same word as the morphological origin of the word; it is typically enough that associated words map to the identical stem, even if this stem is not in itself a applicable root. Since 1960s, the algorithms for stemming have been considered and proposed in the field of computer science and natural language processing. Stemming programs are normally called as stemming algorithms or stemmers. Example for stemmer is if we give a word as "running" it is stemmed as run. Julie Beth Lovins has written the first available stemmer in the year 1968[19]. This paper was amazing for its early date and had huge power on later work in this area. In the year 1980, Martin Porter has developed a stemmer and published his work. This stemmer was called as Porter Stemmer and was very extensively used and became usual algorithm used for English stemming. The Tony Kent Strix award was received by Dr. Porter in 2000 for his work on information retrieval and stemming. Many performance of the Porter stemming algorithm were printed and generously spread; however, many of these stemmer implementations have limited subtle flaws. As the effect of these flaws, the stemmer did not able to reach its potential performance. To abolish this source of error, Martin Porter discharges an official free-software execution of the algorithm around the year of 2000. He expanded this task over the next few years by constructing Snowball, a structure for writing stemming algorithms, and executed an enhanced English stemmer jointly with stemmers for a number of other languages. There are many type of stemming algorithms that are

regarded for their performance and correctness and also they are able to overcome positive stemming blockages. These stemmers use the algorithms of suffix stripping, lookup, lemmatization, matching, affixing, and hybrid of these. Lemmatization process can also be used in medical information retrieval[20]. Lemmatization can be used in medical IR due to the fact that it stem a word to its actual root or dictionary form.

In medical information retrieval porter algorithm is used for stemming. This algorithm was introduced in 1980. It is commonly used algorithm in information retrieval. There are several steps in performing the algorithm. First, the plurals have to be removed from the given word. For example treatments are converted into treatment. Then –ings and –ed has to be eliminated. For example recognizing is converted into recognize. When there is any other vowel in the stem the terminal y has to be changed to i. For example furry is converted into furri. Double suffixes are mapped into single one. For example possibly is converted into possibli in second step. Then in this step it is converted into possible. In next step it deals with suffixes such as –ness, -full. For example completeness is converted into complete. In this way, the root word is stemmed.

2.3 UMLS

UMLS stands for Unified Medical Language System[23] that consists of various medical related vocabularies and medical literatures in the field of biomedical sciences. UMLS was developed and maintained by the National Library of Medicine, US to aid the recognition of keywords in the domain of biomedicine and health by the computer systems. As part of the UMLS, NLM constructs and allocates the UMLS Knowledge Sources. These are databases and connected software tools that are programs for system developers in constructing or attracting electronic information systems that generate, get back, and combine or collective biomedical and health statistics and information. The UMLS Knowledge Sources are not partial for exacting functions; they are multi-purpose. System developers will discover that they can be useful in systems that execute a range of functions connecting one or more types of information like, for example, patient records, systematic journalism, and public health data. The software tools of UMLS help out the system developers in modifying the knowledge sources of UMLS. The three tools of UMLS, namely, Metathesaurus, Semantic Network, and SPECIALIST Lexicon are terms collectively as the knowledge sources. Metathesaurus is a great, multi-purpose, and multi-lingual thesaurus that includes millions of biomedical and health correlated models, their identical names, and their associations. The Metathesaurus contains over 150 electronic reports of categorizations, code sets, thesauri, and lists of prohibited terms in the domain of biomedical. Semantic Network includes a set of large topic categories and their relationships, which supply a dependable classification of all thoughts symbolized in the Metathesaurus tool of UMLS and a set of functional and significant semantic associations that survive among semantic types. In other words, semantic network and the lexical tools are used to produce Metatheasurus. The SPECIALIST Lexicon presents the lexical information required for the specialist Natural Language Processing System. It contains frequently happening biomedical vocabulary and regular English words. The lexicon entry for each term or word records the logical, syntactical, morphological, and orthographic information about the terms. In our work, the query is expanded by assigning relevant medical terms for the selected candidate terms. The data set used was images and text[21]. The search can be improved by using UMLS query expansion[22].

It is unreasonable to request consumers or field experts to physically recognize scenario specific conditions for every question and all potential scenarios. As a result, a mechanical approach is extremely advantageous. However, the difference between scenario-specific expansion terms and non-scenario-specific ones possibly will seem noticeable to an individual specialist, but can be very complicated for an agenda. To extravagance this peculiarity, a domain-specific knowledge source is used. Knowledge sources are frequently not particularly planned for the principle of scenario-specific recovery. Therefore, scenarios normally materializing in medical queries may not be sufficiently carried by those knowledge sources. A knowledge-acquisition methodology is used to complement the accessible knowledge sources with supplementary knowledge that supports approximate scenarios.

3. Experimental Verification

The time taken by the search engine to retrieve the documents varies when using Medical term and the normal term. When Medical term is used the retrieval time is less whereas when natural language is used

the time is more. The retrieval time has been checked with many Medical terms and their equivalent natural language term. For that it is known that when Medical term is used the retrieval time is less whereas when we use natural language the time is more. This is because when we use Medical term, the term is more specific and it displays the accurate documents whereas when we use natural language, the documents is retrieved based on the terms used in the query. The results are accurate because we combine the natural term with the medical term. We tested these with many Medical terms and some of them are listed in Table 1. The Medical terms are referred from PubMed search engine. When we use natural term the results are not more effective, when we use medical term the time is less and when the medical term and the natural term is combined and given for search the retrieval time is more but the result is more accurate than others because the medical term is present in the query.

Example:

- User query: I have fever.
- Root word: fever.
- Medical term: Hyperthermia.
- Expanded query: Fever – Hyperthermia.

Here, in the above example, fever is the root word and the medical term for the word is 'hyperthermia'. To provide effective result the natural word and the medical word are combined.

Table 1. Comparison of time taken for retrieval

Normal Term		Medical term		Time taken by combining medical term and normal term (in sec)
Term	Time taken in sec	Term	Time taken in sec	
Fever	0.29	Hyper-thermia	0.20	0.38
Heart disease	0.28	Cardiovas-cular	0.18	0.38
Eye Infec-tion	0.32	Cornea	0.23	0.26
Pimple	0.39	Acne	0.27	0.35
Cancer	0.24	Epitheli-oma	0.22	0.37
Neck pain	0.22	Crick	0.16	0.33
Cold	0.44	Nasophar-yngitis	0.29	0.37

4. Conclusion

A method is proposed here to improve the queries in the field of Medical Information Retrieval. The enriched query is given to the database to retrieve the documents. When a normal word is given for search it is automatically converted as a medical term using UMLS. The medical term and the query given by the user are combined and searched which gives efficient results. It is difficult for the people who are not aware of the medical terms to understand the documents retrieved from medical search engine. It is found that the work proposed here showed some improvement in the query expansion process that would benefit both expert users in the medical domain and the novices to this field.

5. References

1. Boguski MS. Online health information retrieval by consumers and the challenge of personal genomics. In: Willard HF, Ginsburg GS, editors. Genomic and Personalized Medicine. Elsevier; 2009. p. 252–7.
2. Fox S. Health topics: 80% of internet users look for health information online. Washington, DC: Pew Internet and American Life Project; 2011 Feb 01. Available from: http://www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Health_Topics.pdf
3. Turan N, Kaya N, Aydın GO. Health problems and help seeking behavior at the internet. Procedia - Social and Behavioral Sciences. 2015; 195:1679–82.
4. Ruban, Sam B, Shetty A. A hybrid framework to refine queries using ontology. Indian Journal of Science and Technology. 2015; 8(24).
5. Jadidinejad AH, Sadr H. Improving weak queries using local cluster analysis as a preliminary framework. Indian Journal of Science and Technology. 2015; 8(15).
6. Galhardas H, Florescu D, Shasha D, Simon E. Declaratively cleaning your data using AJAX. In Journees Bases de Donnees; 2000 Oct. Available from: http://caravel.inria.fr/~galharda/BDA.ps
7. Galhardas H, Florescu D, Shasha D, Simon E. AJAX: An extensible data cleaning tool. Proceedings of the 2000 ACM SIGMOD International Conference on Management of data (SIGMOD '00); New York, NY, USA. 2000. p. 590–5.
8. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Computer Society Technical Committee on Data Engineering. 2000; 23(4):3–13.
9. Li WS, Clifton C. SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using

neural networks. Data and Knowledge Engineering. 2000; 33(1):49–84.
10. Do HH, Rahm E. On metadata interoperability in data warehouses. Germany: Univ of Leipzig. p. 1–21. Available from: http://lips.informatik.uni-leipzig.de/files/2000-13.pdf
11. Milo T, Zohar S. Using schema matching to simplify heterogeneous data translation. Proceedings of the 24th International Conference on Very Large Data Bases (VLDB '98); New York, NY, USA. 1998. p. 122-33.
12. Swetha S, Uma D, Suganya P, Nivedhitha V, Saravanakumar K. On the performance of medical information retrieval using MeSH terms – A Survey. Journal of Engineering Science and Technology Review. 2014; 7(4):137–42.
13. Cassidy T. Concurrency analysis of Java RMI using source transformation and verisoft. Kingston: Queen's University; 2003.
14. Cordy JR. The TXL source transformation language. Science of Computer Programming. 2006; 61(3):190–210.
15. Brabrand C, Schwartzbach MI. Growing languages with metamorphic syntax macros. Proceedings of the ACM SIGPLAN Workshop on Partial Evaluation and Semantics-based Program Manipulation (PEPM '02); Portland, Oregon, USA. 2000. p. 31–40.
16. Hall MA. Correlation-based feature selection for machine learning. Hamilton: The University of Waikato; 1999.
17. Pellow F, Pirahesh H. Data Cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery. 1997; 1(1):29–53.
18. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003; 3:1157–82.
19. Liu H, Hussain F, Tan CL, Dash M. Discretization: An enabling technique. Data Mining and Knowledge Discovery. 2002; 6(4):393–423.
20. Liu H, Christiansen T, Baumgartner WA, Verspoor K Jr. Bio Lemmatizer: A lemmatization tool for morphological processing of biomedical text. Journal of Biomedical Semantics. 2012; 3(3):1–29.
21. Diaz-Galiano MC, Martin-Valdivia MT, Urena-Lopez LA. Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in biology and medicine. 2009; 39(4):396–403.
22. Martinez D, Otegi A, Soroa A, Agirre E. Improving search over Electronic health records using UMLS based query expansion through random walks. Journal of Biomedical Informatics. 2014; 51:100–6.
23. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D267–70.
24. Lenzerini M. Data integration: A theoretical perspective. Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '02); Madison, WI, USA. 2002. p. 233–46.