EFFICIENT DIVERSITY AWARE RETRIEVAL SYSTEM FOR HANDLING MEDICAL QUERIES

V.Sudha[1], E.V.R.M. Kalaimani[2]

[1]*M.E Student, Department of Computer Science, Arasu Engineering College, Kumbakonam.*
[2]*Professor, Department of Computer Science, Arasu Engineering College, Kumbakonam.*

Abstract

Clinical research are meagre in healthcare today and most clinical data analytical tools have difficulty in extracting valid information from unstructured data.The hike of Electronic Medical Records (EMRs)causes an eruptive growth of the medical data. The current paper based medical search technologies are tedious to find meaningful patient information in the large medical database. Moreover, the good quality medical search in large collection of EMR is a challenging task especially due to the complex semantic relationships among medical concepts. Hence, to address the semantic issues, asearch result diversification and semantic based IR are proposed. Meta map concept identifier is introducedto map the biomedical query terms to the corresponding MeSH terms. With the help of clinical domain ontology, all the potential semantics from an input query are mined and consumed to model in to variety of query aspects. The extent of VSM is used to constitute the EMR log as vectors. Each group of words consists of multiple concepts and words. The documents are ranked based on its importance. In this project, aNeural Network based Classifier is used to train and test the medical documents and queries topredict the similarity measure between them.

Keywords: *Electronic Medical Records, Information Retrieval, Meta Map Concept Identifier, MeSH Ontology, Neutral Network Classifier*

----------***----------

1. INTRODUCTION

Electronic Medical Records (EMR)

EMR are the substratum of modern healthcare. It is widely deployed for the tracking of medical data over time and to improve the overall quality of care in practice. It also barter medical information among various healthcare related parties. This fact makes the medical search is becoming a critical technique for the rapid and effective access of patient information. In traditional topic-based approach, the bag-of-words scheme is used to represent the text as a multiset of its words. Medical information retrieval is challenging because of the inherent ambiguity within the posed queries .Such ambiguity is manifested in different ways: (1) A query expresses a clearly defined sense, but the genuine needs under this sense may cover a broad range. Taking a common scenario where an ordinary user performs medical search for example, he feels excruciating (he has a leg pain when walking and rash erupts on his body) but is dubious about his exact medical problems, so he inputs ‘‘leg pain’’ and ‘‘rash’’ as keywords into a search engine. In this case, as many diseases may cause these symptoms, the user may prefer to learn knowledge about all these diseases, so as to have a preliminary understanding about his situation and better prepare for the interview with doctors. (2) Query terms themselves are ambiguous, as most users have little medical knowledge. For instance, a pregnant woman feels pain in her stomach, so she submits a query composed of ‘‘pain in the stomach’’ and ‘‘pregnant’’. In this case, the term ‘‘pain’’ is ambiguous, which may mean ‘‘stabbing pain’’, ‘‘distending pain’’, ‘‘labor pain’’, etc. The user-cared reasons causing these different kinds of pain, however, may be totally different.

A. INFORMATION RETRIEVAL (IR)

Information Retrieval is the task of wrenching out the documents collection. It is the activity of obtaining information resources pertinent to end-user input. IR is used to retrieve the list of ranked log based upon the concept based input query. According to Abby (2000), the evaluation of an information retrieval systemis to assess how efficiently the system satisfies the expectations of the user needs. Orthodox evaluation technique is forBoolean retrieval or top-k retrieval. It also includeprecision and recall as an additional measure.

Data retrieval (IR) systems utilize uncomplicated data model whereas DB systems is very complex

1. Information is well-ordered as a group of logs.
2. Records are randomly ordered, it is schema-less. IR is used to retrieve or extract the relevant records based upon the user query.Suchas keywords or concepts.

Keyword Search

In datatext retrieval, all the words in each of the query log are determined to be the unique keywords. It allows query expansion formed using keywords and the analytical connectives such as: AND, OR, and NOT. Documents are ranked based upon the determination of the pertinent of a query such as:

Term frequency
Frequency of existence of the query keyword in log documents.

Inverse document frequency
It is a weigh of how copiousthe word replicatethat is, whether the key is familiar or sparse across all the log documents. If keywords in query arise close together in the document, the document has higher importance than if they occur far apart. Documents are returned in decreasing order of relevance score. Usually only top few documents are returned, not all.

Similarity Based Retrieval
Similarity based retrieval retrieve documents similar to a given document. User selects a few closely connected documents from those retrieved by keyword query, and system finds other documents which is similar.

Vector space model
Itis an n dimensional space, where n is the number of words in the document set. Vector for document d goes from origin to a point whose i^{th} coordinate is TF (d, t) / n (t). The cosine of the angle between the vectors of two documents is used as a measure of their similarity.

Precision vs Recall Tradeoff
In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced by a web search engine for a query) and a set of relevant documents (e.g. the list of all documents on the internet that are relevant for a certain topic), relevance. Measures of retrieval effectiveness:

Recall as a function of number of documents fetched,

$$\text{Recall} = \frac{t_p}{t_{p+}\, f_n}$$

Precision as a function of recall equivalently, as a function of number of documents fetched.

$$\text{Precision} = \frac{t_p}{t_{p+}\, f_p}$$

2. LITERATURE REVIEW

Luo et al [10] proposed an intelligent output interface for intelligent medical search engine. To facilitate ordinary people to search medical information, we have built an intelligent medical Web search engine called iMed. IMed uses medical knowledge and an interactive questionnaire to find multiple diseases serving as queries. This paper presents a new, intelligent search result output interface devoted to intelligent medical search. The new output interface automatically offers searchers what they want instead of waiting until they ask explicitly. We demonstrate the effectiveness of our techniques through an evaluation using USMLE (United States Medical Licensing Examination) medical exam cases.Electronic Health Records has been adopted in literature, Fridsma [7] a method to explain a model based on knowledge of the diversity of query subtopics to generate a diversified ranking for retrieved documents. We expand the original query into several related queries, assuming that query expansions expose subtopics of the original query.Betin and Medico Port [2] have dealt the medical search engine for all Computer Methods and Programs in Biomedicine. The semantic resources are exploited to represent queries in an expressive and meaningful context, through which to fill the semantic gap among the queries and documents and then to improve the medical search quality. Six existing domain-independent semantic similarity measures are adapted to the biomedical domain. However, all these proposed semantic based medical IR approaches consider the relevance as the only measure for medical record ranking, which makes their capabilities to handle the query ambiguity limited.The diversification strategy involves four steps: Query understanding, query transformation, candidate concept mapping and derived query generation. Meta map concept identifier is used to map the biomedical terms to MeSH (Medical Subject Heading) concepts.

3. PROPOSED METHODOLOGY

The proposed model encompasses six major processes such as EMR Pre-processing, Diversification strategy, Meta Map, MeSH ontology, Vector Space Ranking Model and Neural Network based Classifier. Pre-processing strategy is used to analyse the stop word, synonym, and white space present in the user query and the appropriate keyword is extracted from the input medical query. The diversification strategy involves four steps: Query understanding, query transformation, candidate concept mapping and derived query generation. Meta map concept identifier is used to map the biomedical terms to MeSH (Medical Subject Heading) concepts. The extent of VSM is used to constitute the EMR log as vectors**.** Each group of words consists of multiple concepts and words. The documents are ranked based on its importance. The importance of Neural Network based Classifieris used to train and test the medical documents and queries to predict the output based on the similarity between the document vectors and query vectors.

A. MEDICALDOCUMENT PRE-PROCESSING

Lucene is the search library used to provide a full-text indexing across both database objects and query log in various formats.

B. THE DIVERSIFICATION MEDICAL SEARCH APPROACH

Our proposed approach on disambiguating of medical records includes two steps, i.e., (1) query understanding to discover the implied aspects of the original query as multiple sub-queries; (2) diversity-aware medical retrieval to exploit multiple sub-queries to for disambiguating the medical search results. The following of this section gives a detailed description on each of the two steps.

Query Understanding

The given query keyword is understood and elucidated in a diversified manner. Our main aim is to understand the user based keyword and explicate their meanings and produce a set of uniquely retrieved medical records as the delivery outcome.In order to handle the ambiguity problem, the random catalogue of query keyword isexpounded in a varying fashion, i.e., the sense of the conceptsidentified in the user inputis perceived and discovered in to the potential aspects. The main motive is to understand the query input and transforms it into a set of derived queries to model different aspects of q. As medical ontology contains rich and accurate professional knowledge that is shared by domain experts.

Query Transformation

This sub-step carries out two functions, i.e., keyword phrase identification and expansion. With the support of available semantic resources, e.g., WordNet and Consumer Health Vocabulary (CHV), the former uses the maximum matching approach to scan the keywords in the query sequentially and find the longest matching sub sequences defined in the semantic resources as the keyword phrases. For example, given a query ‘‘difficulty breathing headache’’, the longest maximum matching approach can find ‘‘difficulty breathing’’ as a keyword phrase and ‘‘headache’’ as the other keyword phrase. For the latter, two types of expansions are conducted.

Meta Map

MetaMap is a device formed by NLM that plots free script to medicinal ideas in the UMLS, or equally, it determines metathesaurus ideas in script.Meta Map is a method to recognize entities from raw text by mapping them to MeSH terms with a scoring system. Take “mammary cancer” as an example, Meta Map will not only map entities to the MeSH term Malignant Neoplasm of Breast, but also provide information on the source vocabularies from which the term is originated. In this case, it is the MeSH term Breast Cancer, therefore, one can use this Meta Map feature to identify hierarchically related entities, which is exactly the main idea behind the first approach. However, it easier to further improve its performance. Meta Map based MeSH includes following steps,

- Entity processing.
- Apply Meta Map to processed entities.
- Generate candidate mappings from Meta Map results.
- Choose final mapping from candidates.

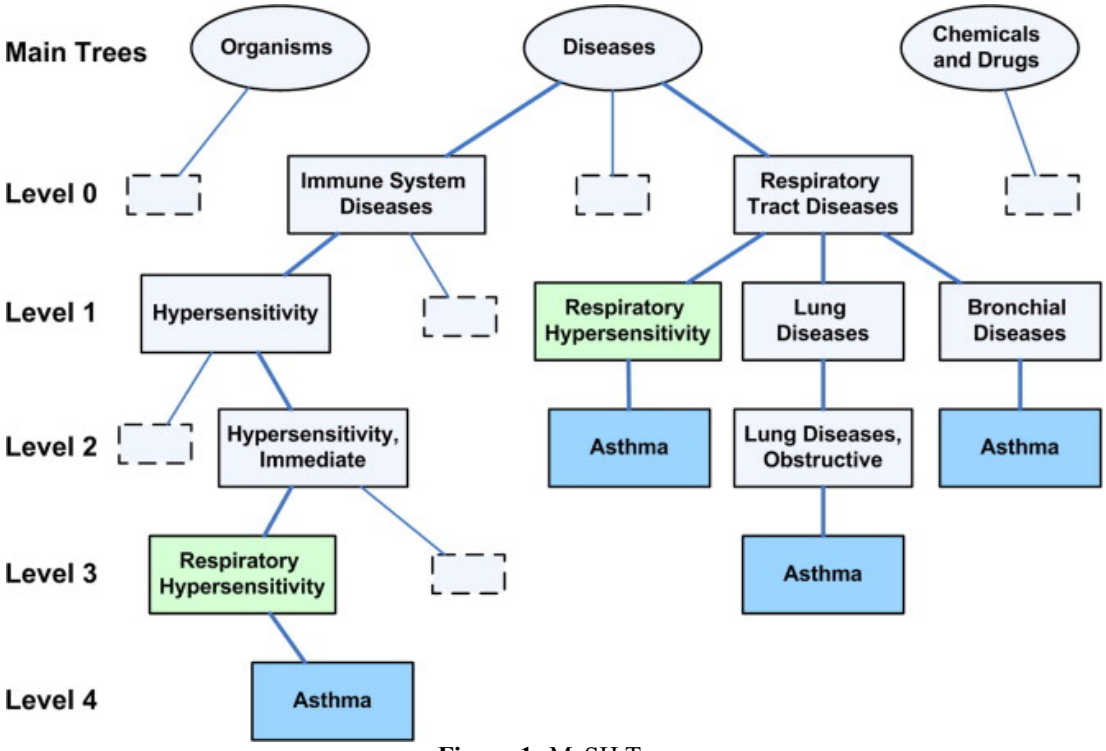


Figure 1: MeSH Tree

Replacement of Greek Letters:

Replace Greek letters by their full English names (e.g. α by alpha) .Greek letters and their full names are used interchangeably in articles, but only their full names are used in MeSH.

Extraction of non-English words:

Some chemicals and proteins are described by non-English words, such as “3-chloro-1, 2-propanediol” and “IGF-1”. Those words have exact matches in MeSH. It uses several features to identify non-English words, such as the presence of numbers, capital letters, and special characters (e.g. hyphen and comma).

Original Query	Query Transformation	Meta Mapping	Derived query generation
Headache at night	Two keyword phrases are identified: #1: headache; #2: at night; Keyword: phrase expansion: #1(headache) Same as:{ Cephalalgia; Cranial Pain; Hemicrania} #2(at night) same as:{night time;nocturnal}	Three concepts are identified for #1: C_{11}:Cephalalgia[ID:C10.597.617.470] C_{12}: Cranial Pain[ID:C15.598.645.500] C_{13}: Hemicrania [ID: C23.888.592.612.441] 12 concepts are identified for #2: C_{21}:Somnambulism[ID:C10.886.659.635.700] C_{22}:Somnambulism[ID:F03.870.664.635.700] C_{23}:Nocturnal Enuresis[ID:C12.777.934.284.500] C_{24}:Nocturnal Enuresis[ID:C13.351.968.934.252.500]	Three sub queries are generated: Q1:{(Headache disorders; cranial pain;nocturnal), Q2:(migraine ;hemicrania,headache at night) Q3:(signs and symptoms,headache; cephalagia,nocturnal; headache at night)}

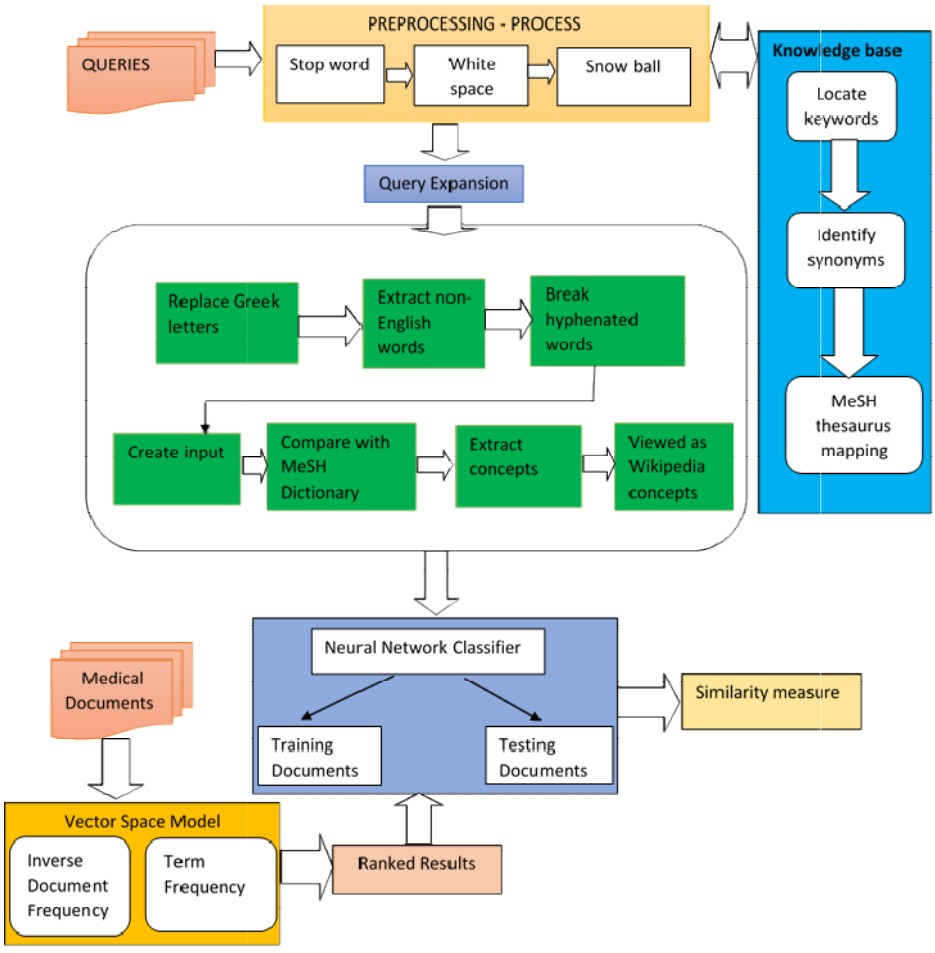


Figure 2: Diversity aware Retrieval System

generated by exploiting various concepts of the query. By this step the normal medical queries are transformed in to the MeSH query concepts. Table 3.1 depicts the diversification steps in pre- processing strategy.MeSH as the final mapping. When there are more than one such term, highest score term is selected.
Table 1: Diversification Flow

Break Hyphenated Words

Words joined by hyphens (e.g. soy-induced) are broken into atomic words. Meta Map often fails to break hyphenated words. Then it passes the pre-processed entities to Meta Map again and use its special feature to map entities to MeSH. As described earlier, Meta Map will give a list of MeSH terms as the potential mapping results of given entity. Each mapping is assigned with a score from 0 to 1000 indicating the probability of being the correct mapping. To generate candidate mappings, the MeSH terms related to disease, food, chemical, protein and gene, which are the focus of this study. Among those candidate mappings, the medical term originated from

Derived Query Generation

This sub-step constructs a list of derived queries to model the various aspects of q. The sub queries are.

C. VECTOR REPRESENTATION OF DOCUMENTS

Here, the documents are represented as a vector of identifiers. The basis of the vector space corresponds to unique terms in a set of documents. The importance of the keyword in the document is the weight of the documents vectors. In a naivest approach, the word is considered as a key-term. Yet, morphological variants like “Lymphocytosis” and “Lymphocyte” are so closely related that they are usually conflated into a single word stem, e.g., “Lymph,” by stemming. Word stems are usually treated as notational, rather than conceptual entities. Two word stems are considered unrelated if they are divergent. For example, the stem of “hyperthermia” and that of “febrile” are usuallyconsidered unrelated despite their apparent relationship. In stem-based VSM, word stems constitute the basis of the vector space. The base vectors are orthogonal to each other because different word stems are considered unrelated.

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

Each facet corresponds to a separate term. The value of vector is non-zero if a term occurs in the document. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting. The extent of VSM is used to constitute the EMR log as vectors. Pertinent rankings is calculated based on the presumption of query log documents. The deviation of the angles between input query and log documents is compared using the cosine similarity measure.

In exercise, the cosine of the angle between the query vectors and document log is calculated as:

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|} \quad (1)$$

Where $\mathbf{d_2} \cdot \mathbf{q}$ the intersection of the query log document and q is the query, $\|\mathbf{d_2}\|$ is the standard of vector d_2, and $\|\mathbf{q}\|$ is the standard of vector q.

The standard of a vector is calculated as such:

$$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2} \quad (2)$$

4. ALGORITHM

Neural Network Based Classifier Using Rear Distribution Algorithm

In Artificial Neural Networks, theRear distributionis a common method to teach theseveral layer of recurrent neural networks from the field of Computational Intelligence. Recurrent neural networks are stimulated by the data text retrieval processing of large number of neural cells.

Approach

The purpose of Rear-Distribution algorithm is to train the load in a several layer recurrent neural networks. The work of a recurrent network connotes in three stages. Where one layer stage is fully connected to the succeedingstage layer. A traditionalnetwork structure containsone input layer, one hidden intermediate layer, and one output layer. The method is applied directly from output layer then to the input layer i.e., backtracking process is undergoneit is primarily concerned with acclimatizing the load to the calculated fallacy in the presence of inputs.Activation=$(\sum_{k=1}^{n} wk * xki) + wbia * 1$ (3)

where n is the number of load weights and inputs, x_{ki} is the k^{th} ascribe on the i^{th} input pattern, and w_{bias} is the bias weight.

For the output nodes,this is the sum of the fallacy between the output nodes and the anticipated output.The method is applied directly from output layer then to the input layer i.e., backtracking process is undergoneit is primarily concerned with acclimatizing the load to the calculated fallacy in the presence of inputs.is given as

$es_j = (a_j - o_j) * cd_j$ (4)

Where es_j the fallacy sign for the j^{th}node,a_j is the anticipated output and$o_j$is the real output for the j^{th} node. The cd term is the imitative of the output of the j^{th} node.

$es_j = (\sum_{m=1}^{n} (wjm * esm)) * cdj$ (5)

Where $es_j$is the error signal for the j^{th}node, $w_{jm}$is the weight between the$j^{th}$and the m^{th} nodes, and $es_m$is the fallacy sign of the m^{th} node.

$ed_j = \sum_{m=2}^{n} esj * ym$ (6)

Where $ed_j$is the fallacyimitative for the j^{th}node, es_j is the fallacy sign for j^{th}node and ym is the intake from the m^{th} node in the precursory layer. Load weights are renovated in a way that decreases the fallacy imitative ed (fallacy assigned to the weight), metered by a learning coefficient.

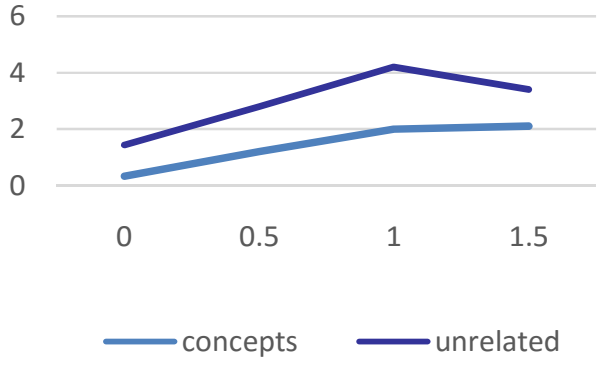
$W_c(c+1) = w_j(c) + (ed_m * learn_{rate})$ (7)

Where $w_j(c+2)$ is the renovated j^{th}weight, $ed_m$is the fallacy imitativefor m^{th}node and l_{rate} is an renovate coefficient factor.

Input: Query vectors, Document vectors, $iterations_{max}$, l_{rate}
Output: Classified output
Network ←ConstructNetworkLayers()
$Network_{weights}$ ← InitializeWeights(Network, Query vectors)
For (i=1To $iterations_{max}$)
  $pattern_i$←SelectInputPattern(InputPatterns)
  $output_i$←ForwardPropagate($pattern_i$, Network)
  BackwardPropagateError($pattern_i$,$output_i$, Network)
  UpdateWeights($pattern_i$,$output_i$, Network,l_{rate})
End
Return (Network)

5. RESULTS AND DISCUSSIONS

This result deliberates the similarity between the input queries and documents. One of the main assistances of this effort is to calculate the relevant documents and rank it based on its similarity between the document keyword and its respective query. The outcome gained is the evaluation of the precision and recall measures. The plotted graph indicates the better outcome when compared to the existing phrase based approach.



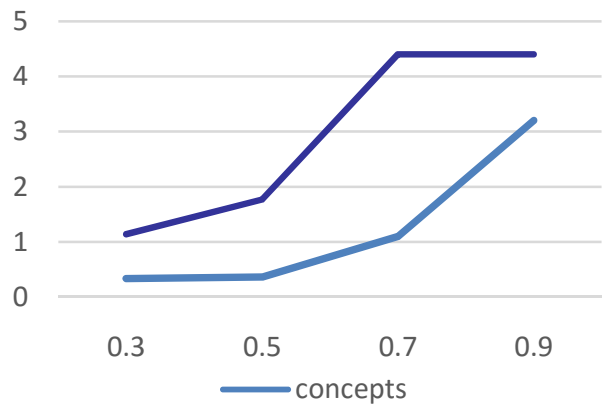


Figure 3: Performance comparison for Precision and Recall in concepts

The phrase based VSM lacks retrieval accuracy compared with the proposed work. In order to improve the accuracy of the concepts, similarity measure is calculated between medical queries and input vectors. The proposed approach balances both precision and recall accuracy for concept relevant and irrelevant queries or documents. The semantic based retrieval process is better than that of content based retrieval method. Hence the evaluation graph shows that the precision and recall values is calculated based on the positive predictive value and its sensitivity. Semantic similarity is stated in vector representation of the concepts and are acquired from knowledge origins. The aspect of the retrieval process mainly depends on the knowledge database of the sources.

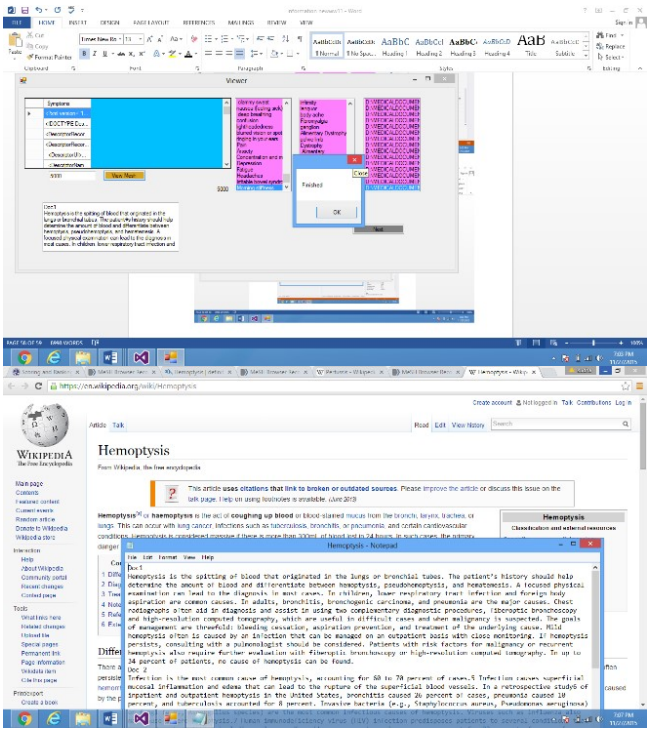


Figure 4: MeSH Query Loaded

CONCLUSION

This inferences of this modern approach is to retrieve the medical queries in a diversity fashion. The first challenge is

to extract the user query and consumed for query aspect modelling; then based on the modelled query aspects, the proposed approach not only considers the diversity aspect importance but also the similarity of the aspects. This empirical experiments on the real-world datasets illustrate the effectiveness of the proposed approaches. Our work displays that the finest outcome is reported on the application of the Diversity approach for the secondary use of medical queries, through which to enhance the healthcare quality and balance the utilization of medical resources. This paper gives the basic concept of the diversity-aware retrieval of medical records, which represents a beginning to combine the technologies of semantic IR and search result diversification for the secondary use of EMRs. In future, the current research is extended in the implementation of the cloud and data fusion of multiple sources for improvement of the secondary use of available medical data and then enhancing the healthcare quality.

REFERENCES

[1]. Agrawal.R, Gollapudi.S(2009), Diversifying search results, in: Proc. of WSDM,514.

[2]. Betin.A, MedicoPort(2007): a medical search engine for all, Computer Methods and Programs in Biomedicine 86 (April) 73–86.

[3]. Carbonell.J, Goldstein.J(1998) The use of MMR, diversity-based re ranking for reordering documents and producing summaries, in: Proc. of SIGIR'98,335–336.

[4]. Chen.H, Karger.D.R(2006), Less is more: probabilistic models for retrieving fewer relevant documents, in: Proc. of SIGIR'06, 429–436.

[5]. Demidova.E,et al(2010)., DivQ: diversification for keyword search over structured databases, in: Proc. of SIGIR'10, 331–338

[6]. Doyle, Lauren; Becker, Joseph (1975). Information Retrieval and Processing. Melville. pp. 410 pp. ISBN 0-471-22151-1.

[7]. Fridsma.D(2012), Electronic health records: the hhs perspective, Computer 45 (November(11))24–26, http://dx.doi.org/10.1109/MC.2012.371.

[8]. Goodrum, Abby.A (2000). "Image Information Retrieval: An Overview of Current Research". Informing Science 3 (2)

[9]. Hanbury.A(2012), Medical information retrieval: an instance of domain-specific search, in: SIGIR, 1191–1192.

[10]. Luo.G(2008), Intelligent output interface for intelligent medical search engine, in: Proc. of AAAI, 2008, 1201–1206.

[11]. Manevitz.L, MandYousef.M(2011), "One-class SVMs for document classification," J. Mach. Learn., vol. 2, pp. 139–154, Dec.

[12]. Radlinski.F, Dumais.S(2006), Improving personalized web search using result diversification, in: Proc. of SIGIR'06, 691–692.