INDEXING OF MEDLINE DOCUMENTS – A TEXT MINING APPROACH

1S.sagar Imambi, 2T.Sudha,
1Asst Professor, TJPS College, simambi@gmail.com
2Professor, SPMVV, Tirupathi

Abstract— Medline and PubMed repositories are rich in medical literature. Once the documents are retrieved from PUBMED, they need further analysis. The exponential growth of these online repositories and availability of large volume of documents has motivated the search for the hidden knowledge from these resources. Therefore, an automated system which could correctly extract information from PubMed is needed. We proposed a novel feature weighting schema to generate index terms. Our results show that the accuracy and precision are high when global relevant weight schema is used. Indexed terms generated from GRW are more relevant compared to terms extracted by other methods.

Keywords— Indexing-Medline documents, Online repositories, Text mining

I. INTRODUCTION
In recent years the availability of text has grown tremendously due to the Internet, digital libraries, news portals, organizational data, emails, blogs and other social networks. There is a growing need for tools helping people to find, filter and merge these sources[8]. PubMed and Medline are one of the sources of Biomedical literature. PubMed and Medline repositories have been growing at the rate of 500000 articles per year. Indexing and Classification of Biomedical literature from digital libraries is a time-consuming process that is prone to inconsistencies.[1]. The exponential growth of these online repositories and availability of large volume of documents has motivated the search for the hidden knowledge from these resources. Therefore, an automated system which could correctly extract information from PubMed, is needed.(N.Urimate, H.Matsuzawa 2004). Text mining can enhance the retrieval of useful information from PubMed and Medline.[2,28]. It can be viewed as an extension of data mining or knowledge discovery from databases.

Generally biomedical data repositories include different types of data like gene expression data, disease symptoms and diagnosis other diseases, protein identification data, genome sequence data gathered by Human Geneome Project, findings from the research publications etc. The goal of biomedical information system is to discover knowledge from online repositories and put it into practical use in the forms of diagnosis, prevention and treatment of diseases.

II. MEDLINE
MEDLINE is the largest component of PubMed, the freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine. Approximately 5,400 journals published in the United States and more than 80 other countries have been selected and are currently indexed for MEDLINE. PubMed Comprises more than 22 million biomedical literature from Medline, life science journals and online books. (http://www.ncbi.nlm.nih.gov/pubmed).

Since 1990, the MEDLINE database has grown faster than before with more documents available in electronic form. The cost of human indexing of the biomedical literature is high, so many attempts have been made in order to provide automatic indexing. A unique feature of MEDLINE is that the records are indexed with NLM's controlled vocabulary i.e the Medical Subject Headings (MeSH).

Medical Subject Headings (MeSH) mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several types of terms, such as Descriptors, Qualifiers, Publication Types, Geographics, and Entry terms. Descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to descriptors. For example, “Neoplasms” as a descriptor has the following entry terms (“Cancer,” “Cancers,” “Neoplasm,” “Tumors”, “Tumor”, “Benign Neoplasm,” “Neoplasm, Benign”). MeSH descriptors are organized in a MeSH Tree, which can be seen as the MeSH Concept Hierarchy. In the MeSH Tree, there are 15 categories (e.g., category A for anatomic terms), and each category is further divided into subcategories. For each subcategory, corresponding descriptors are hierarchically arranged from most general to most specific. (www.nlm.nih.gov/mesh)

MeSH descriptors have been used to index PubMed articles and used as features to extract information form PubMed articles.[4] used MeSH descriptors as the selected
features for classification and showed that there is a significant improvement of classification performance.

Only Mesh descriptors are not sufficient for indexing and classifying the PubMed documents. There are several studies to improve the performance of these tasks. Indexing using MeSH ontology reduces the dimension feature space and computational complexity. And it is proved by the following studies. Varelas et al. (2005) integrated domain ontology using term re-weighting for information retrieval application. Terms are assigned more weight if they are semantically similar with each other. YooHu, and Song (2006) applied MeSH domain ontology to clustering initialization and achieved best results. Terms are first clustered by calculating semantic similarity using MeSH ontology on PubMed document sets. Then the documents are mapped to the corresponding term cluster. Jing et al. (2006) adopted similar technique on document clustering. Xiaohua Hu (2006) perform comprehensive comparison study of document clustering on 44 medline corpora for seven document clustering approaches. STC provides better clustering solutions than hierarchal algorithms and domain ONTOLOGY MESH improves document clustering for MEDLINE articles. Xiaodan Zhang,(2008) evaluated the effects of nine semantic similarity measures with a term re-weighting method on document clustering of PubMed document sets using Mesh Ontology[5].

III. INDEXING

A. Challenges of Medline/PubMed Information Retrieval system

The availability of large medical online collections, such as MEDLINE poses new challenges to information and knowledge management like indexing.

- Ambiguity of lot of domain-specific terminology
- Only MESH terms are not provide conceptual information for searching and indexing
- PubMed will not apply any ranking strategy.
- Rapid proliferation of biomedical literature need unique information extraction tools

Feature weighing and selection for automatic indexing documents

Feature or term weighting is an important part in the process of information retrieval systems. Precise term weighting can greatly improve the process of finding index terms. The amount of influence of term in representing the document reflects on term weight. Traditional term weging schemas like TFIDF, LOG (TFIDF), Information gain etc. are not much useful in text documents.

There are several Novel weighting schemas that are proposed and tested by keen researchers. Some of them are WAKNN(Eui-Hong Han 2000), BTWS (Yunjung Jung 2000), STW (Franca Debole,2003), STFS (Q. Xu 2004),CONF (Pascal Soucy,2005), TFRF (M. Lan 2005). Random walk approach (Samer Hassan et al 2007) and ICF (Deqing Wang, Hui Zhang 2010) [9, 11,12,15]

Feature Selection can be defined as the task of selection of subset features that describe the documents. Feature selection is used to select the relevant terms to represent the documents. There are several studies that compared, summarized, analyzed and reviewed the feature selection methods like S.M.Ruger (2000), Imola (2002), Janma Novovicova, Antomin (2004), Shoushany, Yivan Saeyts et.al (2007), A.Dasgupta (2007) Ruixia (2009), [16,17,18].

Researchers have developed novel Feature selection methods to reduce the dimensionality and complexity. (Sanjay Chawla 2010, B.M.Vidyavathi, Dr.C.N.Ravikuma, Ngetal 2006, Li & Zong 2005 G.Forman 2003,Sebastian 2002). [20,21,22]

The computational complexities of the feature selection methods for biomeric applications were compared by Yvan Saeyts(2007), Marthis Sewll(2008), Z. Zhao, F. Morstatter(2010) and they recommended the techniques combined with evolution criteria for better performance.

Shasha Liao, Minghu Jiang(2005) developed a new method to extract feature selections based on concept attributes. They used the DEF terms of the Chinese word to extract concept attributes, and a Concept Tree (C-Tree) to give these attributes proper weights considering their positions in the C-Tree, as this information describe the expression powers of the attributes.
Qinghua Zou (2003) presented Index finder algorithm for generating all Valid UMLS concepts by permuting the set of words in the text and then filtering out the irrelevant concepts via synthetic and semantic filtering. Some researchers recently put their focus on the conceptual features extracted from text using ontologies and have shown that ontologies could improve the performance of text mining (A. Hotho 2003, S. Bloehdorn 2004).

PubMed and Medline documents provide very flexible resource for researchers. Padmini Srinivasan (2002) presented a system which is developed for the extraction of pairs of concepts from Medline dataset and to find the association between them. Padmini Srinivasan et al (2004) also developed Open discovery algorithm to uncovering implicit information from the Medline documents. They used this method to investigate the potential of turmeric or Curcumin Longa and identifies a ranked list of problems. Go-tag [27], NEWS-ML(Fai Wong 2002), MedMESH (P. Kanlear 2002), IndexFinder (Qinghua Zou, 2003) are the novel architectures developed to analyze Medline document collections. Thuy. T. T. Nguyen and Darryl. N. Davis. (2007) presented an improvement on k-mean algorithm k-mix and allows its applications to the mixer of attribute types found in the cardio vascular domain.

IV. PROPOSED ALGORITHM FOR GENERATING INDEX TERMS

The proposed model pre-process the documents to extract features. Preprocessing includes tokenizing, stemming, stop words removal etc. The unique terms from documents are weighted by using the novel global relevant weighting schema. The proposed schema is variation of global weight schema IDF. GRW(t) is calculated by using the below formula.

\[ \text{GRW}(t) = \text{TFIDF}_j \times \frac{P(T_{ij})}{P(C_i)} \]

Where TFIDF$_j$ is the tfidf value of jth term, P(T$_{ij}$) is probability of term ‘$j$’ belongs to class ‘$i$’ and P(C$_i$) is the probability of documents that belongs to class ‘$i$’ . Feature selection strategy is applied to select GRW of the term by using selection criteria GRW(t)= max{ GRW(t,C$_i$). For example,
If GRW(t1,c1)=0.25, GRW(t1,c2)=0.28 and
GRW(t1,c3) = 0.4 then term t1 is selected for class 3.

All the terms with high relevance are selected from each class and they are used as Index terms. Threshold ‘th’ is used to select the terms from each class. The relevance of the terms is measured with Baye’s classification. The Accuracy of the classification indicate relevance of the terms.

V. EXPERIMENTAL RESULT

For our experiment purpose, we used eight data sets which have been extracted from Pubmed and Medline. The document collection is a combination of documents related to complications of Diabetics like Diabetic Cardiopathy, Diabetic Retinopathy, Diabetic Nephro and Diabetic Neuropathy. Each complication represents a class. They are heterogeneous in terms of document size, number of classes, and document distribution. We observed non linear distribution of the documents among the classes. The aim of the experiments is to generate index terms model which helps in retrieving documents from Medline. Fig 2 shows the distribution of M1 data set with 3 classes. Table 1 shows the reduction terms to generate indexes. The relevance of the index terms is measured with classification accuracy and is showed in the table 1.

![Fig 2. The Non linear Distribution of M1 document set with 3 classes.](image)

<table>
<thead>
<tr>
<th>Validation Measure</th>
<th>Medline Data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>Accuracy</td>
<td>79.39%</td>
</tr>
<tr>
<td>Total Features</td>
<td>2765</td>
</tr>
<tr>
<td>No. of Selected Features</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. Reduction of index terms

VI. CONCLUSION

The indexing plays an important role in information retrieval process. Text mining enhance the performance of indexing. We proposed the global relevant weight schema based on the probability of term relevance to find relevant index terms. Our results show that the accuracy and precision are high when global relevant weight schema is used. We experimented on the text documents collected from diabetic literature of MEDLINE.
REFERENCES

1. Latha K, Kalimuthu et al., information extraction from biomedical literature using text mining framework, IJSE,GA,USA,ISSN:1934-9955,Vol.1,no.1, January 2007
7. Eui-Hong (Sam) Han George Karypis, Vipin Kumar, This work was supported by NSF grant ASC-9634719
11. Yun Jae Jung, Ha sun Park, Technical Report, Department of Computer Science and Engineering, University of Minnesota 4-192 EECs Building, 200 Union Street SE, Minneapolis, MN 55455-0159 USA
12. Pascal Soucy, Beyond TF-IDF Weighting for Text Categorization in the Vector Space Model, IJCAI05 Proceedings of the 19th international joint conference on Artificial Intelligence, 2005
16. SMRuger and SE Gauch- Feature Reduction for Classification and clustering, Technical Report, Computing Department Imperial College London, UK 2000
17. Jana Novovicova, Feature Selection using Improved Mutual Information for Text Classification Lecture Notes in Computer Science, 2004
23. Shasha Liao, Minghu Jiang: An Improved Method of Feature Selection Based on Concept Attributes in Text Classification. ICNC (1) 2005: 1140-1149