International Journal of Science, Environment
and Technology, Vol. 2, No 2, 2013, 267 - 274

EXTRACTION OF BIOMEDICAL INFORMATION FROM MEDLINE DOCUMENTS –A TEXT MINING APPRAOCH

[1]S. Sagar Imambi and [2]T. Sudha
[1]Asst Professor, TJPS College,
[2]Professor, SPMVV, Tirupathi
E-mail: simambi@gmail.com

Abstract: Medline and Pubmed repositories are rich in medical literature. 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. Text mining can enhance the retrieval of useful information from PubMed and Medline. It can be viewed as an extension of data mining or knowledge discovery from databases. 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. We proposed a novel text mining algorithm to enhance the performance of the information retrieval system. Our experiments conducted on various data sets showed that, it improves the accuracy and precision of the information retrieval.

Keywords: Online repositories -Indexing-Medline documents, 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 [21]. 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. Extraction of information from Biomedical literature and Classification of Biomedical literature from digital libraries is a time-consuming process that is prone to inconsistencies [10]. 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. Uramate, H. Matsuzuwa 2004). Text mining can enhance the retrieval of useful information from PubMed and Medline [13, 27]. It can be viewed as an extension of data mining or

*Received March 3, 2013 * Published April 2, 2013 * www.ijset.net*

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. BIOMEDICAL LITERATURE

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 that 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).

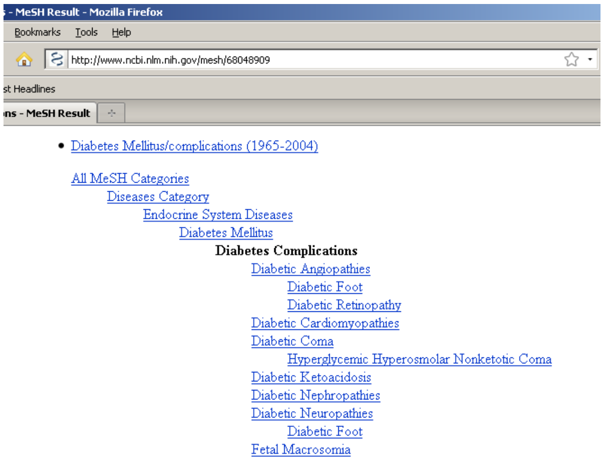
Medical Subject Headings (MeSH) mainly consists of the controlled vocabulary and a MeSH Tree. (www.nlm.nih.gov./mesh). MeSH descriptors have been used to index PubMed articles and used as features to extract information form PubMed articles [26] used MeSH descriptors as the selected features for classification and showed that there is a significant improvement of classification performance.

Only Mesh descriptors are not sufficient for extracting information and classifying the PubMed documents. There are several studies to improve the performance of these tasks. MeSH ontology as index terms, 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 [27].

A. *Challenges of Medline/PubMed Information Retreival 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



http://www.ncbi.nlm.nih.gov/mesh/68048909

- Diabetes Mellitus/complications (1965-2004)

All MeSH Categories
  - Diseases Category
    - Endocrine System Diseases
      - Diabetes Mellitus
        - **Diabetes Complications**
          - Diabetic Angiopathies
            - Diabetic Foot
            - Diabetic Retinopathy
          - Diabetic Cardiomyopathies
          - Diabetic Coma
            - Hyperglycemic Hyperosmolar Nonketotic Coma
          - Diabetic Ketoacidosis
          - Diabetic Nephropathies
          - Diabetic Neuropathies
            - Diabetic Foot
          - Fetal Macrosomia

Fig 1. Diabetes Complications from MESH

III. 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), and 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 (Yunjae 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) [5, 28, 14, 25].

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), Janna Novovicova, Antomin (2004), Shoushanli, Yvan Saeys et.al (2007), A. Dasgupta (2007) Ruixia (2009) [9,1,12].

Researchers have developed novel Feauter 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) [6, 22, 24].

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 identify a ranked list of problems. Go-tag [25], NEWS-ML (Fai Wong 2002), MedMESH (P. Kanlear 2002), Index Finder (Qingghua 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 GORITHM FOR GENERATING INDEX TERMS**

The proposed model pre-processes 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.

$$GRW(tC_i) = TFIDF_j \times \frac{P(T_{ij})}{P(C_i)}$$

Where TFIDFj is the tfidf value of jth term, P(Tij) is probability of term 'j' belongs to class 'i' and P(Ci) 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,Ci}.

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 indicates 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.

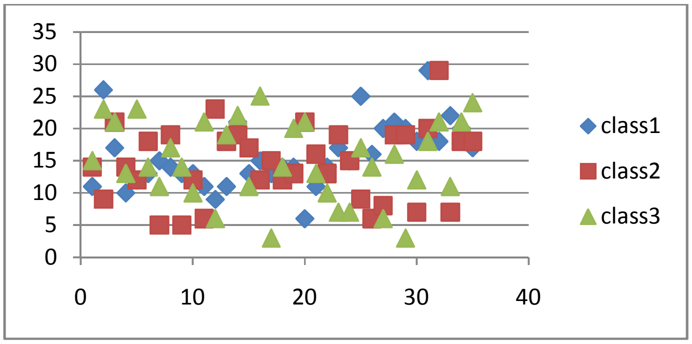


Fig 2. M1 document set with 3 classes

Validation Measure	Medline Data sets							
	M1	M2	M3	M4	M5	M6	M7	M8
Accuracy	79.39%	64.30%	98.05%	93.78%	99.95%	100.00%	99.94%	98.33%
Total Features	2765	2652	3117	4085	5026	10360	59680	62050
No. of Selected Features	4	30	61	24	80	16	15	31
%of reduction	99.85533	98.86878	98.04299	99.41248	98.40828	99.84556	99.99162	99.95004

TABLE 1: ACCURACY OF INFORMATION SSYTEM WITH VARIOUS ALGORITHMS.

VI CONCLUSION

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. Text mining enhances the performance of biomedical Information retrieval system. 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] A. Dasgupta, Feature selection methods for text classification, KDD'07, 2007.

[2] Deqing Wang, Hui Zhang, Wenjun Wu, Mengxiang Lin, The Computing Research

Repository, vol. 1012, 2010.

[3] Eui-Hong (Sam) Han George Karypis, Vipin Kumar, This work was supported by NSF grant ASC-9634719.

[4] Frakes, W.B., "Introduction to Information Storage and Retrieval Systems", in Frakes, William B. and Baeza-Yates, Ricardo (Eds.), Information Retrieval: Data Structures and Algorithms, Englewood Cliffs, NJ: Prentice-Hall, 1992, pp. 1-12.

[5] Franca Debole, Fabrizio Sebastiani, Proceedings of SAC03 18th ACM symposium on Applied Computing Melborne US, 2003, pp 784-788.

[6] George Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research 3, 2003, pp: 1289-1305.

[7] http://www.cse.unt.edu/~rada/papers/hassan.ieee07.pdf

[8] http://www.ncbi.nlm.nih.gov/pubmed

[9] Jana Novovi•cova, Feature Selection using Improved Mutual Information for Text Classification Lecture Notes in Computer Science, 2004.

[10] 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.

[11] M. Ghanem et al, GoTag: A Case Study in Using a Shared UK e-Science Infrastructure for the Automatic Annotation of Medline documents, I4th UK e-Science All-Hands Conference AHM 2005.

[12] M. Lan, S.Y. Sung, H.B. Low and C.L. Tan, A Comparative Study on Term Weighting Schemes for Text Categorization, Proc. International Joint Conf. Neural Networks, pp. 546-551, 2005.

[13] Padmini Srinivasan et al Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases, Workshop: Biolink 2004, pp 33-40.

[14] Pascal Soucy, Beyond TFIDF Weighting for Text Categorization in the Vector Space Model, IJCAI'05 Proceedings of the 19th international joint conference on Artificial Intelligence, 2005.

[15] Q. Xu, “A significance test-based feature selection method for the detection of prostate cancer from proteomic patterns” University of Waterloo, 2004.

[16] Qinghua Hu, Daren Yu, Yanfeng Duan, Wen Bao: A novel weighting formula and feature selection for text classification based on rough set theory, Natural Language Processing and Knowledge Engineering, 2003.

[17] R. Radha et al Fuzzy logic approach for diagnosis of Diabetics, Information

Technology Journal 6(1), 2007, pp 96-102.

[18] S. Bloehdorn, and A. Hotho, Text classification by boosting weak learners based on terms and concepts, Proc. of the 4th IEEE International Conference on Data Mining, pp 331-334, 2004.

[19] S. Sagar Imambi, T. Sudha: A Unified frame work for searching Digital libraries Using Document Clustering; International Journal of Computational Mathematical ideas, Vol 2-No1-2010, pp 28-32.

[20] S. Sagar Imambi, T. Sudha: Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining; International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010.

[21] S. Sagar Imambi, T. Sudha: Clinical Decision Support System for Heart Patients-International Journal of Computer Science, System Engineering and Information Technology, Vol 2-No2. (2009), pp 165-169.

[22] Sanjay Chawla Feature Selection, Association Rules Network and Theory Building, JMLR: Workshop and Conference Proceedings, The Fourth Workshop on Feature Selection in Data Mining, 2010 pp 14-21.

[23] Sebastiani, F.: Machine learning in automated text categorization, ACM Computing Surveys, 1(34) (2002) 1-47.

[24] Shasha Liao, Minghu Jiang: An Improved Method of Feature Selection Based on Concept Attributes in Text Classification. ICNC (1) 2005: 1140-1149.

[25] SMRuger and SE Gauch: Feature Reduction for Classification and clustering, Technical Report, Computing Department Imperical College London, UK2000.

[26] Sunghwan Sohn, Term-Centric Active Learning for Naïve Bayes Document Classification, The Open Information Systems Journal, 2009, 3, pp 54-67.

[27] Xiaodan Zhang, Medical Document Clustering Using Ontology-Based Term Similarity Measures, International Journal of Data Warehousing & Mining, 4(1), 62-73, January-March 2008.

[28] Yunjae Jung, Haesun 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.