International Journal of Mathematical Archive-2(6), June - 2011, Page: 876-879
Available online through www.ijma.info ISSN 2229 – 5046

INFORMATION EXTRACTION AND DOCUMENT CLASSIFICATION OF MEDLINE ABSTRACTS USING TEXT MINING

[1]S. Sagar Imambi* and [2]T. Sudha**

[1]Dept of Computer Science, TJPS College (P.G), Guntur
[2]Dept of Computer Science Vikram Simhapuri University, Nellore

E-mail: simambi@gmail.com

(Received on: 03-06-11; Accepted on: 15-06-11)

ABSTRACT

B*iomedical literature indexing and classification is time-consuming process that is prone to inconsistencies. Pubmed and Medline repositories are growing at the rate of 500000 articles per year. Therefore, an automated system which could correctly determine the relevant keywords of the abstracts retrieved from Pubmed , is needed. In this paper we developed classifier, to classify Medline documents published in between 2000-2010.*

We are restricted to Type 2 diabetes millets related literature only to decrease the size of corpora .Our objective was to investigate the benefits of using the MeSH controlled vocabulary as features to represent MEDLINE abstracts.. We developed and evaluated a classification model based on global relevant weighting schema to reduce the dimension Our algorithm out performed when compared to two popular Bayes and KNN classifiers and gives 20 % high accuracy than standard classifiers in worst cases.

Keywords Classification: *feature reduction and Medline*

---

1. INTRODUCTION:

Generally Pubmed articles are indexed by MESH terms. Mesh heading and sub heading are powerful tool indexing tools. As Pubmed repositories are growing at the rate of 5, 00,000 articles per year manual indexing becomes very difficult process.[1]. We require special text mining techniques like document classification. Document classification is a wide-spread problem with many applications, from organizing digital library to spam filtering. Indexing and Automatic extraction of useful information from these online sources remains a challenge because these documents are unstructured and expressed in a natural language form i.e in text format. Text mining tools are first developed in order to facilitate the automated searching of digital library material by users [13]. Due to advent of powerful computing facilities and widespread of www, text mining becomes a new and exiting research area. In this paper, we studied several classification techniques and tried to improve the Medline classification process by reducing the number of keywords .Our algorithm was tested and proved. It shows best results in worst case of document distribution.

We collected Pubmed abstracts, which are published between 2000 and 2010. The fig 1 shows the Mesh index for Diabetes complications.

2. SURVEY OF LITERATURE:

Several Document Classification algorithms are proposed and developed since 1960. Most frequently used algorithm in text mining include support vector machines (SVM), decision trees (DT), naïve Bayes (NB) and KNN.

Joachims [9] Found that SVM classifier is robust even at high dimensional document set.. But SVM degrades the performance when data set contain diversified vocabulary [10] .Sahani et al developed system to filter junk mails in 1998 .Humphrey et al. presented the technique of automatic indexing of documents .

Yang Jin [] developed MTag , which has been engineered to directly accept downloaded files from PubMed and formatted in MEDLINE format as input, and to output text and HTML file versions of the tagger results. MTag was tested with a randomly selected 1,010 (70%) training and 432 (30%) evaluation documents pertaining to cancer genomics and shows only 87 5 accuracy.

Corresponding author: [1]S. Sagar Imambi*, *E-mail: simambi@gmail.com

Donaldson et al. [2] used an SVM algorithm to locate PubMed® citations containinginformation on protein-protein interaction before they were curated into the Biomolecular Interaction Network Database. Dobrokhotov et al. [3] applied a combination of natural language processing and probabilistic classification to re-rank documents returned by PubMed according to their relevance to Swiss-Prot database curation. Beil et al. described an algorithm to hierarchically classify documents based on frequent term sets [8]. Their system provided an intuitive way for users to understand the contents of the clusters. Bernhardt et al. [4] developed an automated method for identifying prominent subdomains in medicine that relies on Journal Descriptor Indexing, an automated method for topical categorization of biomedical text.

Miotto et al. [5] tested the performance of DT and artificial neural networks to identify PubMed abstracts that contain allergen cross-reactivity information. McDonald et al. [7] exploited the maximum entropy classification principle to calculate the likelihood of MEDLINE® abstracts containing quotations of genomic variation data suitable for annotation in mutation databases.

Wang et al. [10] used an NB classifier to speed up the abstract selection process of the Immune Epitope Database reference curation. Classical statistical methods and machine learning algorithms have been shown highly useful for. However, they usually require long computational times and tedious manual preparation of training datasets. Our previous work shows that global relevant weighing schema improves the classification performance. [15] So we propose a novel Document Classification algorithm for saving manual indexing.

3. CONSTRUCTION OF CLASSIFIER:

We have a set of predefined categories and a set of documents .For each category, the document set is partitioned into two mutually exclusive sets of relevant and irrelevant documents. The goal of a text classification system is to determine whether a given document belongs to any of the predefined categories. The document can belong to any one of the predefined categories as we collected abstracts that belong to only on category. There are 4 steps in construction of classifier. First we preprocess the collected data, remove the stop words and generate stemmed mesh vocabulary. In second step these terms are given weight age, so that we can find the relevance of each term. Basing on the relevance of term, feature is selected. In third step the all the abstracts are represented in vector space model with reduced dimensionality. And Prototype vectors are generated for each class .The classification model was build by using the prototype vector in the 4^{th} step.66% of data is used as training data and remaining was test data.

To evaluate a text classification system, we use the *F*1 measure this measure combines recall and precision in the following way: Recall =number of correct positive predictions/number of positive examples Precision =number of correct positive predictions/number of positive predictions
*F*1 =2 *Recall *Precision /(Recall + Precision)

3.1 Dataset:

We collected 2800 document instances from pubmed online repository which are deposited between 2000-2010. The dataset includes documents related to Diabetes Mellitus Complications. I choose only cardiomypothesis, neuropathy and nephropathy according to Mesh tree structure fig 2.2. After removing nosiy documents the data set size becomes 2561 .The data set is labeled as (‘cardio, ‘neuro’,’nephro’).The distribution of abstracts under each category are cardio 744, neuro 934 and nephro 883.

3.2 Experimental Setup:

By using Matlab we developed software for generating features and use the novel approach for feature selection. As Matlab is very flexible in vector processing, we developed a program that generates vector space model of documents. The weka software is used to test the accuracy of various existing classification techniques available in weka and compared with our new method. The following results are obtained by the algorithm. Table1 shows the precision, table 2 shows recall and table 3 shows the F-measure. When compared with all the three classifiers, our algorithms produce more accuracy and recall is high. Though the data is un evenly distributed, algorithm classify data with 99.7% accuracy.

Classifier	Precesion			Total
	Category1	Category2	Category3	
Bayes	.837	.841	.86	0.846
DT	.882	.856	.978	.906
KNN	.893	.802	.892	0.86
Our algorithm	1	.995	1	.998

Table1: Precision of 4 classifiers

© 2011, IJMA. All Rights Reserved

Classifier	Precesion			Total
	Category1	Category2	Category3	
Bayes	.902	.835	.81	0.846
DT	.948	.914	.846	.9
KNN	.837	.889	.838	0.856
Our algorithm	0.993	1	1	.997

Table2: Recall of 4 classifiers

Classifier	Precesion			Total
	Category1	Category2	Category3	
Bayes	.0.868	.838	.834	0.845
DT	.914	..884	.907	.901
KNN	.864	.844	.864	0.857
Our algorithm	.996	.997	1	.997

Table3: Fmeasure of 4 classifiers

classifier	
Bayes	84.57%
DT	90.04%
KNN	85.6%
Our algorithm	99.7%

Table 4: Accuracy of 4 classifiers.

4. CONCLUSION:

Medical abstract classification is very complex process. As the documents are distributed unevenly and has high dimensionality accurate classification is the biggest problem. Our Weighing schema and classification algorithm out performs with non linear distributed data and reduced the dimensionality. Our algorithm is compared with 3 popular classification models like Bayes, Decision tree and KNN.

5. REFERENCES:

[1] Medical Subject Headings Home Page [homepage on the Internet]. [Cited 2009 Oct 22]. Available from: http://www.nlm.nih.gov/mesh/.

[2] Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, et al. PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 2003; 4: 11.

[3] Dobrokhotov PB, Goutte C, Veuthey AL, Gaussier E. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. Bioinformatics 2003; 19 (Suppl 1): i91-i94.

[4] Bernhardt PJ, Humphrey SM, Rindflesch TC. Determining prominent subdomains in medicine. AMIA Annu Symp Proc 2005; 46-50.

[5] Miotto O, Tan TW, Brusic V. Supporting the curation of biological databases with reusable text mining. Genome Inform 2005; 16 (2): 32-44.

[6] Chen D, Müller HM, Sternberg PW. Automatic document classification of biological literature. BMC Bioinformatics 2006; 7: 370.

[7] McDonald R, Scott Winters R, Ankuda CK, Murphy JA, Rogers AE, Pereira F, et al. An automated procedure to identify biomedical articles that contain cancer-associated gene variants. Hum Mutat 2006; 27 (9): 957-964.

[8] Beil F, Ester M, Xu X: Frequent term-based text clustering. Proceedings of the Eighth ACM SIGKDD 2002:436-442.

[9] Joachims .T (1998) Text categorization with support vector machines: learning with many relevant features ECML'98(pp.137-142).

© 2011, IJMA. All Rights Reserved

[10] Beccerman R. (2003) Distributional clustering of words for text categorization, Masyer's thesis CS department, Technion-Israel Inst of Technology.

[11] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inform Syst 2008; 14 (1): 1-37.

[12] Yetisgen-Yildiz M, Pratt, W. The effect of feature representation on MEDLINE document classification. AMIA Annu Symp Proc 2005; 849-853.

[13] 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.

[14] Muller HM, Kenny EE, Sternberg PW: Textpresso: An ontologybased information retrieval and extraction system for biological literature. PLoS Biol 2004, 2:e309.

[15] S.Sagar Imambi, T.Sudha- Classification of Medline documents using Global Relevant Weighing Schema', International Journal of computer Applications February 2011, pp 45–48.

****

© 2011, IJMA. All Rights Reserved