Classification of Medline documents using Global Relevant Weighing Schema

S.Sagar Imambi
Asst. Professor
TJPS College(PG)
Guntur, Andhra Pradesh

T.Sudha
Prof. & HOD
Vikram simhapuri university
Nellore,Andhra pradesh

ABSTRACT

Medline and Pubmed repositories are rich in medical literature .Once the documents are retrieved from PUBMED, they need further analysis. This paper describes new model for text classification by estimating terms weights and shows how the classification accuracy is improved with this method. The method uses global relevant weight as term weighing schema. Experiments performed with different weighing schemas shows that the new global relevant weighing method outperforms the traditional term weighing approaches.

Keywords

Classification- Global weight-Text mining

1. INTRODUCTION

We ask Text Mining is the process to extract meaningful data from the text, and, thus, make the information contained in the text accessible to the various data mining algorithms. Text mining is different from web search. In search, the user is typically looking for something that is already known and has stored by some one. In text mining, the goal is to discover unknown information, something that is not known. Text mining is a variation on a field called data mining [2] that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text.[17].

Medline and Pubmed repositories are rich in medical literature. 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. It is virtually impossible for researchers to obtain all the information that is important and available for their work. Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted.

There are several text mining approaches for handling the vast amount of textual domain-specific information available, some of them are Document clustering, Text classification.

Document clustering is defined as the automatic discovery of document clusters/groups in a document collection, where the formed clusters have a high degree of association (with regard to a given similarity measure) between its members. Members from different clusters will have a low degree of association

Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Most traditional clustering methods do not satisfy the special requirements for document clustering. An unified frame work for document clustering technique which can be implemented in searching text from digital libraries was proposed by T.Sudha et.al. The novelty of this approach is that it exploits top ranked documents from digital libraries; organize the cluster hierarchy, and reducing the dimensionality of document sets[8].

The goal of Text classification is to build a set of models that can correctly predict the class of the different text documents. The input to these methods is a set of documents (i.e., training data), the classes which these documents belong to, and a set of terms describing different characteristics of the documents.

One of the applications of text classification is E-mail classifier. An email classifier is one of the critical tools needed for the effective management of information in the Internet age. We [15] employed Bayesian classification approach to detect threats in e-mails and classify them to predefined classes.

There are several classification systems, which will analyze structured data from biomedical databases and unstructured data from open access abstracts and full text documents and provide the voluble knowledge to doctors.[4].

Text classification process involves following steps

- Representation of text Documents
- Term weighing
- Classifier learning.
- Calculating accuracy

Term weighing plays an important role in both classification and clustering. We propose new Term weighing approach which improves accuracy of analysis (classification or clustering).

2. REPRESENTATION OF TEXT DOCUMENT

All material Vector space model is used to represent Documents in n-dimensional space. Each individual document Di is represented as a term vector. The representation of documents in this model is as follows

D=[d1,d2,d3,d4……dm]

Di=(ti1,ti2,ti3,ti4,….tin)

Where D indicates Total document set with m elements and Di is set of n terms. Dij represents some information about the occurrence of jth term in the i th document. In the vector space model a document is located as point in an n- dimensional vector space.

Terms are selected from MESH related to Diabetes mellitus. Dimension is equal to the number of features terms, we retrieved from the MESH. MESH is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MESH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. The occurrence of term represents its proportional significance in representing the document.

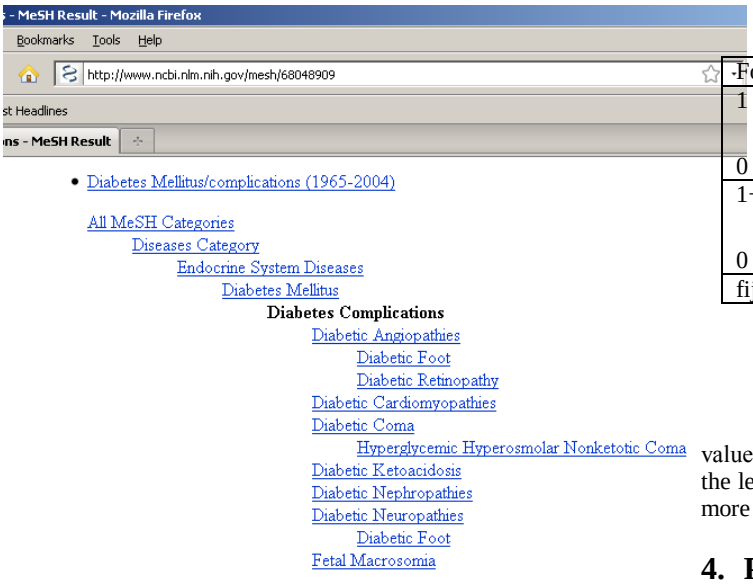

Fig 1. Diabetes Complications from MESH

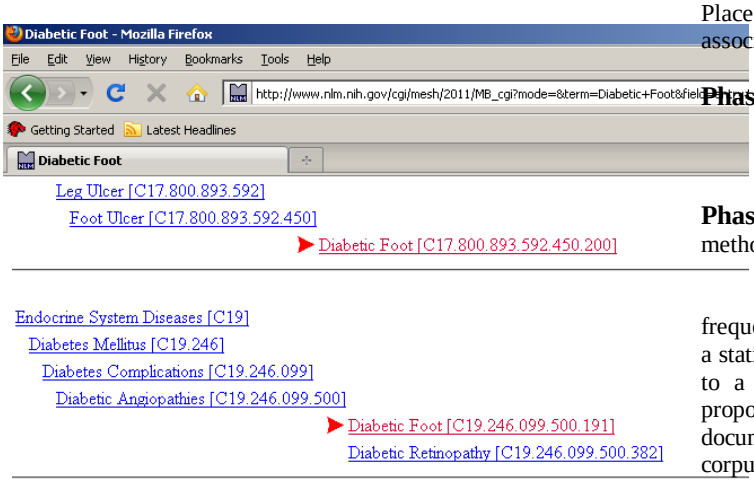

Fig 2: Diabetes complications Mesh tree Structure for Diabetic Angiopathies

3. TERM WEIGHING

Please Term weighting is an important factor in the performance of information retrieval systems. Many weighting methods have been developed within text search, and their variety is astounding.. Term weighing is the process of computing weight of every term in the document set. A weighting scheme is composed of three different types of term weighting: local, global, and normalization. Local weights are functions of how many times each term appears in a document, global weights are functions of how many times each term appears in the entire collection, and the normalization factor compensates for discrepancies in the lengths of the documents[9].

There are many different local weight schemas available. Some of them listed in Table 1.

Formula	NAME
1 if fij >0 0 if if fij =0	Binary
1+log fij if fij >0 0 if if fij =0	LOG
fij	Document frequency

Table 1 : Local weights

Global weighting tries to give a "discrimination value" to each term. Many schemes are based on the idea that the less frequently a term appears in the whole collection, the more discriminating it is [6].

4. PROPOSED GLOBAL WEGHING METHOD

Place Following steps are applied to derive the global weight associated with the terms in the each documents

Phase 1: The document is tokenized for punctuation, special symbols and word abbreviations. Common words are also removed.

Phase 2: Calculate term frequency using any traditional methods

The tf–idf weight (term frequency–inverse document frequency) is a weight often used text mining. This weight is a statistical measure used to evaluate how important a word is to a document collection The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

$$tf_{t,d} \times \log \frac{D}{\left|\left\{d \in D | t \in d\right\}\right|}$$

where tft,d is term frequency of term t in document d , | D | is the total number of documents in the document set; $\left\{d \in D | t \in d\right\}$ is the number of documents containing the term 't.'

Phase 3 Calculate global weight using the formula proposed by us.

Global weight of term tij = local weight of term tij * max (Pi) where Pi is the probability of the term tij belongs to class Ci.

5. EXPERIMENTAL RESULT:

In our experiment we used 1000 documents, collected from the Pubmed. We partitioned the dataset into a training set of 600 and a test set of 400 documents. Sample document is showed in fig3. Pmid is the identification number from the Pubmed. The documents are labeled by 4 categories, which represent the complications of Type 2 diabetes.

```
PMID- 20582892
OWN - NLM
STAT- MEDLINE
DA  - 20100628
DCOM- 20100913
IS  - 0041-4131 (Print)
IS  - 0041-4131 (Linking)
VI  - 88
IP  - 7
DP  - 2010 Jul
TI  - Cutaneous, pulmonary and sinusal aspergillosis in a diabetic patient.
PG  - 519-22
AB  - BACKGROUND: Cutaneous aspergillosis is rarely reported in diabetic patients. AIM:
      The objective of our study is to report a case of lethal disseminated
      aspergillosis revealed by multiples skin necroses, with pulmonary and sinusal
      involvement in a diabetic patient. CASE REPORT: A 60-year-old diabetic woman,
      presented with one month -rapidly -extensive, 1 to 10 cm skin necroses of the
      trunk, limbs and eyelids. Few days after her admission, she developed dyspnoea.
      Chest X-ray showed an interstitial and alveolar syndrome with multiple excavated
      anfractuous-edged-opacities. Facial CT scan showed a right orbital cellulitis
      with Pansinusitis. The methaminesilver stains on a cutaneous biopsy showed
      filamentous septate fungal hyphae with branches at right angles. The
      immunofluorescence with an anti-aspergillus serum was positive. The diagnosis of
      secondary disseminated aspergillosis to a primary pulmonary focus with cutaneous,
      sinusal, and upper airway's dissemination was made. The patient died despite an
      intravenous amphotericin B therapy. CONCLUSION: This report emphasizes the
      importance of evoking and seeking for a mycosis in every skin necrotic and
      ulcerative lesions occurring in an immunocompromised patient. The prognosis
      depends on the diagnosis and treatment institution delay.
AD  - Department of Dermatology, Charles Nicolle Hospital, Tunis, Tunisia.
FAU - Khaled, Aida
AU  - Khaled A
FAU - Fazaa, Becima
AU  - Fazaa B
FAU - Ammar, Donia
AU  - Ammar D
FAU - Bouzgarrou, Alya
AU  - Bouzgarrou A
FAU - Boubaker, Samir
AU  - Boubaker S
FAU - Kamoun, Mohamed Ridha
AU  - Kamoun MR
LA  - eng
PT  - Case Reports
PT  - Journal Article
PL  - Tunisia
TA  - Tunis Med
JT  - La Tunisie medicale
JID - 0413766
```

Fig 3: Pubmed document

We used both local global weights to represent the documents in vector space. After applying the classification algorithm we measured the effectiveness in terms of precision. Precision is calculated using the formula TP/(TP+FP). Precision is the fraction of the documents retrieved that are relevant to the user's information need.

Classified	Relevant - True	Not relevant - False
True	308	12
False	59	21

Table 2. Confusion matrix for the test data

Precision=308/(308+12)=0.9625.

The results are tabulated in table 3.

Schema	Weighing method	Precision
1	Local –Binary	0.9363
2.	Local-Log	0.9452
3.	Local -DF	0.91
4	Global Relevant	0.9625

Table 3. Precision for various weighing methods

The local weight combined with the global weight makes the difference. Our global weight works well in document classification as its precision rate is high compared to other local weighting schemas.

6. CONCLUSION

The Term weighing schema plays an important role in document classification and in Text mining applications. We proposed the global relevant weight schema based on the probability of term relevance. 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 PUBMED.

7. REFERENCES

[1] Berry Michael W, "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43,(2004).

[2] Navathe, Shamkant B., and Elmasri Ramez, "Data Warehousing And Data Mining", in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, (2000), 841-872.

[3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg (2005).

[4] Sergio Bolasco , Alessio Canzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, "Understanding Text Mining:A Pragmatic Approach", Roam, Italy (2002).

[5] Liu Lizhen, and Chen Junjie, China " Research of Web Mining", Proceedings of the 4th World Congress on Intelligent Control and Automation, IEEE, 2333-2337 (2002).

[6] Haralampos Karanikas and Babis Theodoulidis Manchester, "Knowledge Discovery in Text and Text Mining Software", Centre, (2001).

[7] Dhillon I., Mallela S., Kumar R.,A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification, *Journal of Machine Learning Research 3*, 1265-1287, (2003).

[8] 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

[9] Nordiannah et.al-Term weighting Schemes Experiment Based on SVD for Malay Text retrieval- International journal of Computer science and Network security , Vol 8.No.10, (2008).

[10] Srinivasa K.G et.al –Feature Extraction using Fuzzy C-Means Clustering for Data mining systems - International journal of Computer science and Network security Vol 6 No 3A (2006).

[11] 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

[12] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. , J. Documentation, 35(4)(1979) , pp.285-295

[13] 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).

[14] Christian Borgelt and Andreas Nurnberger-Experiments in Document clustering using Cluster Specific Term weighing, C*iteseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.4757&rep.*

[15] D.V. Chandra Shekar --- S.Sagar Imambi -Classifying and Identifying of Threats in E-mails - Using Data Mining Techniques -Lecture Notes in Engineering and Computer Science Vol: 2168 Issue: 1 (2008), pp: 562-566

[16] Ronen Feldman, James Sange, The Text mining Handbook, Cambridge University Press(2007).