Proceedings of the 2nd International Conference on Computer and Information Science and Technology (CIST'16)
Ottawa, Canada – May 11 – 12, 2016
Paper No. 124

Automatic Topic Spotting in Biomedical Literature of diseases using Fuzzy Logic

Sundus Ayyaz, Usman Qamar
Department of Computer Engineering, CEME
NUST, Islamabad, Pakistan
sundus.ayyaz@ceme.nust.edu.pk; usmaq@ceme.nust.edu.pk

Abstract - There are enormous number of biomedical publications being made every day. These publications hold great importance for prevention, treatment and diagnosis of diseases. The rapid growth of biomedical text has directed the emergence of many text mining techniques for extracting valuable information from huge text corpus. This paper proposes a novel approach for automatically assigning a topic to unlabelled text documents using fuzzy logic according to Medical Subject Headings (MeSH) and the Online Mendelian Inheritance in Man (OMIM) vocabulary. The proposed approach consists of five phases, the pre-processing phase for document cleaning and noise removal, the term weight calculation phase for calculating the frequency/weight of each word in the document, the Fuzzy logic phase which categorizes the frequent terms into three fuzzy sets 'low', 'medium' and 'high' using fuzzy rules for extracting a set of significant keywords. The extracted keywords are then used for topic spotting the unlabelled biomedical documents. Our work will be evaluated on the PubMed Dataset.

Keywords: Text Mining, Text Categorization, Topic Spotting, Fuzzy Logic, Extracted Keywords, Medical Subject Headings, Biomedical

1. Introduction

The rapid growth of digital data has granted a significant role to the field of text mining. Mining of the huge mountains of text has become a necessity because of the overwhelming corpus of data as it can ease the problem of information overload and discover valuable information [1]. Text Mining is a process of extracting only required hidden information from massive amount of available text. It can be metaphorically portrayed as a practice of mining precious chunks of pebbles from huge size caves; therefore text mining is a process of extracting the important hidden information from large text source [2].

There are many different types of Text Mining applications that can be used to mine the required information. One of the active research area of Text mining is Text Classification or Text Categorization. It is very important to classify documents into a set of predefined categories. There is a possibility that one document can belong to multiple classes. Text categorization can be applied to many useful applications e.g. emails, news stories, web pages, research articles, journal papers or customer blogs. Text Classification/Categorization is further divided into three classes: supervised text classification, in which the information is provided by some external source such as humans for training and testing of documents, unsupervised text classification also known as document clustering where no training data is available. The third is the semi-supervised text classification in which some documents are pre-labelled by external source [3]. One of the important application of text categorization is called topic spotting. Topic spotting automatically categorizes unlabelled documents from a set of predefined topics by extracting information from document contents [4]. Different types of documents (books, news stories, emails, journal papers, reports) can be categorized into various types of categories, mostly thematic or subject oriented such as the MeSH (Medical Subject Headings) thesaurus covering the medical field [6].

Automatic Topic Spotting ease the problem of manually organizing the huge data corpus, which is tremendously complex, time consuming, unaffordable, error prone and insufficient [5]. In this paper we have used Topic Spotting for automatically categorizing the medical journal papers to the predefined set of medical categories using MeSH and OMIM vocabulary. By extracting the significant medical information many diseases could be prevented and diagnosed earlier and accurately and thus can be cured better. Therefore, the information hidden in the large medical corpus is first extracted and then categorized according to their symptoms and diseases.

Our main objective is to develop a novel approach for finding the disease categories by analyzing the relation between symptoms and disease of unlabelled biomedical text documents using topic spotting. Our work consists of five phases, through which the document passes to get assigned to its appropriate category. This includes, the pre-processing phase, term weight calculation, fuzzy logic, keyword extraction and topic spotting. In the first phase, the pre-processing phase, the text document is transformed into a compact format which can easily be recognized by the classifying algorithm. Pre-processing includes; tokenization, stop word removal, case folding and stemming. The second phase identifies the feature terms by calculating the frequency of the remaining words in the document and selecting the words with high frequency having a specified threshold. In the third phase, Fuzzy logic is applied to assign a membership value to each frequent word through which we can extract the keywords having a particular membership value which is the fourth phase of our work. Using these keywords, the document is categorized by finding a relation between the symptoms and its diseases.

Section two of this paper shows the applications and usage of Text Categorization. Section three demonstrates the proposed solution, section four describes the results, section five describes the dataset used and section six shows the conclusion and future work.

2. Applications of Text Categorization

Text Categorization has a significant role in various areas of document management [5]. Applications of Text Categorization vary along several dimensions. The various areas in which it can be used are:

- *Text Filtering*- in which categorization is based on filtering the text document containing different keywords. e.g. email filters and news filters [3].
- *Automatic document Indexing*- indexes different versions of the same document into the same category to avoid the explosion of index size e.g. if one document has 50 versions then it should not be indexed 50 times. Many features of this area are still unknown [7].
- *Document Organization*- Automatic categorization techniques tackle the issues related to document organization or management e.g. the incoming ads in a newspaper office are categorized under the predefined categories before their publication in paper. The categories can be of Real estate or Sales etc [8].
- *Document Clustering*- This is one of the type of unsupervised learning in which the documents holding some form of similarity between them are collected in one cluster. In this application there is no external source to provide any information for testing the correct clustering of documents [3].
- *Word Sense Disambiguation (WSD)*- it is an activity of finding sense of an ambiguous word in the given text document, for example, a word 'bank' gives two meanings either a bank in which we submit our money or the bank of river. WSD decides the sense to which the given word points to. WSD is considered very much important for a number of applications including document indexing by word senses instead of words for information retrieval [8].
- *Automatic Document Distribution*- it allows automatic distribution of documents through email and fax while saving the time and effort used up by manual processing of emailing and faxing. The documents are categorized according to the message type and sender information. [3]
- *Hierarchical Web Page Categorization*- Several topic hierarchies in the form of web directories are available on web which are used for the classification of web pages as well as for categorization of web data bases e.g. web accessible databases are Medline or ZDNet product review [8].

Other applications of text categorization are, language guessing; that automatically determines the language of a given document, cyber terrorism investigation, spam filtering, segmenting hand written text, email routing, topic spotting; for automatically determining the topic of a text document and many more. Our work is focused on topic spotting of textual documents.

3. Proposed Solution

The proposed text categorization system for biomedical text documents is shown in Fig. 1. The main purpose of the system is to categorize the given collection of biomedical journal papers to the name of the disease available in MeSH vocabulary by interpreting the symptoms specified in the document.

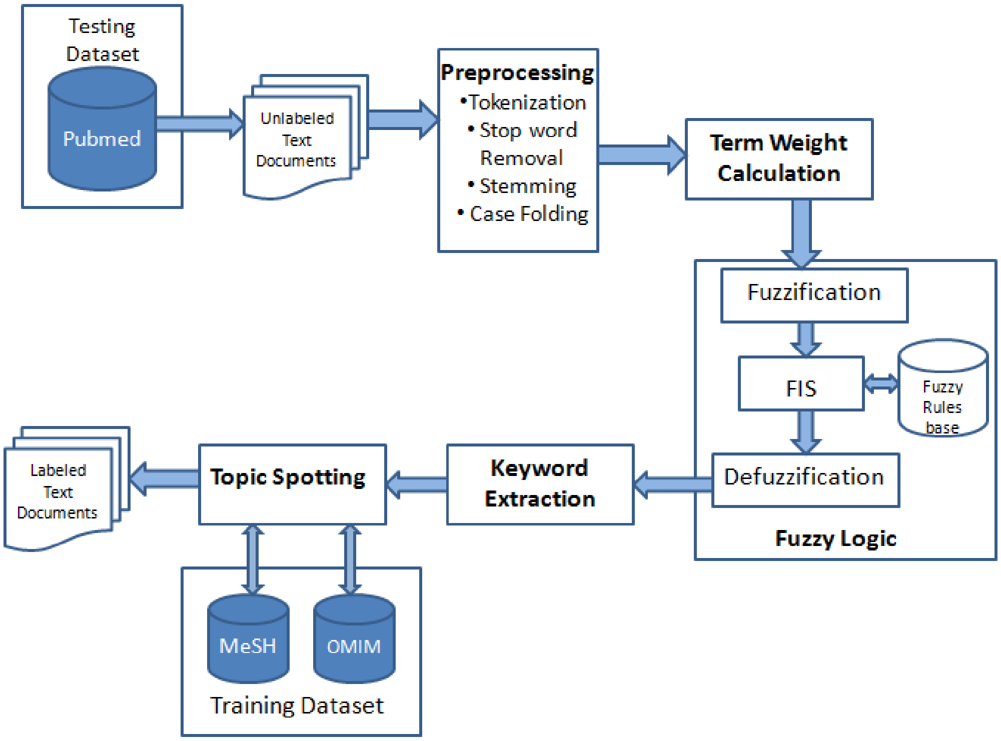

Fig. 1: The Text Categorization System : The proposed system consists of five phases; the Pre-processing Phase, Term Weight Calculation, Fuzzy Logic, Keyword Extraction, Topic Spotting.

The proposed system consists of five phases.

- The Preprocessing Phase
- Term Weight Calculation
- Fuzzy Logic
- Keyword Extraction
- Topic Spotting

The overall detail of the proposed system is given in the following section:

3.1. The Pre-processing Phase

After the text document is loaded in to the proposed system, it enters the pre-processing phase which transforms the original document into a comprehensive form that is easily readable by the classifying algorithm. The pre-processing operation involves the following steps:

a. ***Tokenization***
The input document is separated into a list of words or terms called tokens.

b. ***Stop words removal*** A list of less significant words called stop word list is present in every text document that includes conjunctions and pronouns such as 'is', 'and', 'as', 'be' etc are required to be removed. The remaining terms in the document are called candidate words.

c. ***Stemming***
A process to remove the suffixes and prefixes from the candidate words left in the document. This reduces the number of words in the document hence improving the accuracy of the classifying algorithm. For example; (heat, heats, heated, heating), these terms are reduced to single term heat by removal of the different suffixes -s, -ed, -ing.

d. ***Case Folding***
All the candidate words in the document are converted to lower case to avoid the replication of words in different cases, this helps to develop the accuracy of the system as well as to get the most frequently occurring words in the document.

3.2. Term Weight Calculation

At this phase, the document is transformed into a proper format containing only the frequent terms. To improve the performance of text categorization, weight of each term in the document is calculated and the terms having higher frequency are selected. This is achieved by setting a specific threshold value for each weighted term. Frequency of a term is calculated as follows:

Frequency of a term= no. of times it is repeated in the text document, e.g. if we set the threshold to 25 and check if a word 'blood' occurs 20 times in the document then it should not be selected for further processing according to the specified threshold value.

3.3. Fuzzy Logic

The learning technique used in our work for text categorization is Fuzzy Logic. We have used topic spotting with fuzzy as it provides sufficiently better means to categorize the documents with reasonable solutions and reduced efforts.

A fuzzy set is described as a set of measure which assigns a degree of membership for each element in the set. The membership value lies between 0 to 1. The larger value indicates that the element is a strong member of set and a 0 value means that the element does not belongs to the set. Fuzzy sets also handle the uncertainty that occurs in rule-based systems. For text categorization, we have used fuzzy rules.

3.3.1. Fuzzification

We have taken one input known as the 'frequent term'. Therefore, we have one crisp value to be converted into a fuzzy value. The high frequency terms selected in the previous phase are taken one by one and are fuzzified and given as the fuzzy input to fuzzy inference system/engine (FIS). As we get the fuzzy inputs, we compare it with the fuzzy rule set. The output is generated according to the fuzzy rule set defined in the knowledge/rule base. Next, the output enters the defuzzification method. The output is weighted and defuzzified to return a crisp output for measuring the importance of each frequent term as a keyword to be extracted. Fig 2 shows the working of fuzzy system.

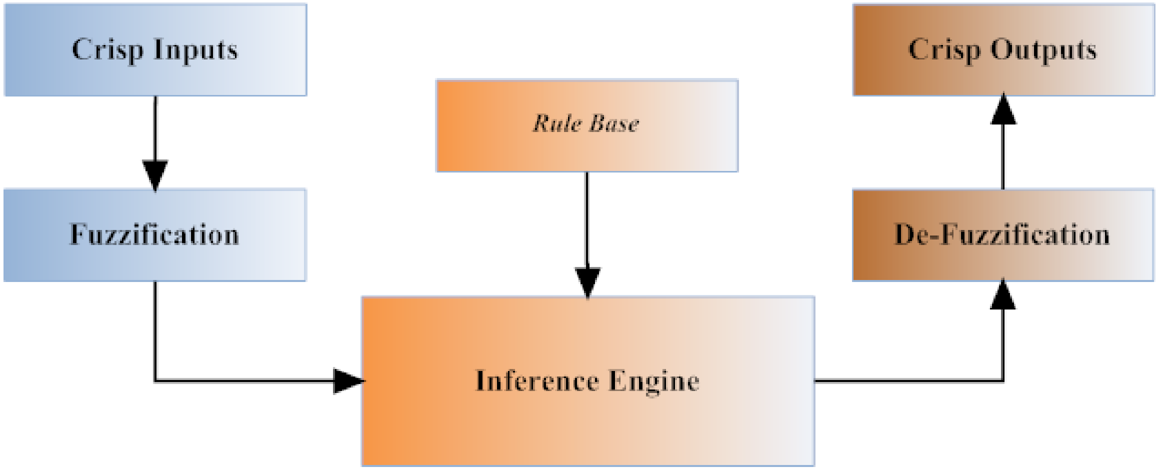

Fig. 2: Fuzzy Logic System: Working of the fuzzy system.

The sets defined for input feature 'frequent term' is *low, medium, high* and a range is assigned to each between 0 and 1. For example, each frequent term can be low , medium or high depending upon the numerical value assigned. Similarly the sets defined for output feature 'Keyword Extraction' are '*No'* and *'Yes'* with a specified range.

Fig 3 graphically visualizes the membership function of the fuzzy sets 'low', 'medium', 'high' with the respective membership values between 0 and 1.

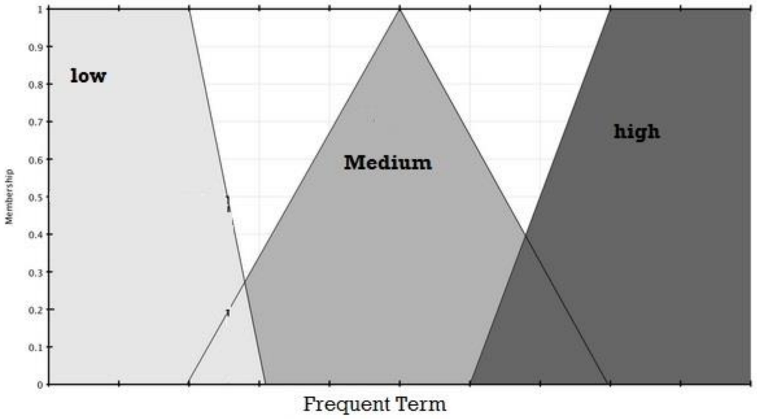

Fig. 3: Fuzzy sets for input feature 'frequent term' : the membership function of the fuzzy sets 'low', 'medium', 'high' with the relevant membership values between 0 and 1.

3.3.2. Fuzzy Inference System/Engine (FIS)

The fuzzy inference system (FIS) applies the fuzzy IF-THEN rules to the input feature 'frequent term' and gets the output feature 'keyword extraction' as 'No' or 'Yes' to be extracted for topic spotting. The rules can be defined as follows:

“IF FrequentTerm is Low THEN KeywordExtraction is No”
“IF FrequentTerm is High THEN KeywordExtraction is Yes”
“IF FrequentTerm is Medium THEN KeywordExtraction is Yes”
Only the frequent terms with a specified threshold values are selected to be extracted as keywords.

3.3.3. Rule Evaluation Result (Application of fuzzy rules)

Fig 4 shows the graphical visual representation of the output feature 'Keyword Extraction' with membership functions 'No' and 'Yes'. The fuzzy inference system (FIS) applies the fuzzy IF-THEN rules to the input feature 'Frequent Term' and gets the output feature 'Keyword Extraction' as either *'No'* or *'Yes'*, to make a decision about extracting the keyword or not.

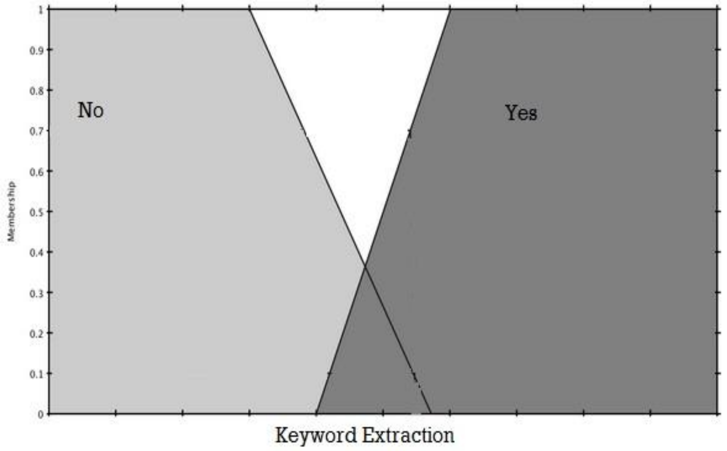

Fig. 4: Fuzzy sets for output feature 'keyword extraction' : representation of the output feature 'Keyword Extraction' with membership functions 'No' and 'Yes'.

3.3.4. Defuzzification

The defuzzification is performed using the most common method called Centroid method, which is used to calculate the centre of gravity under the curve to give the crisp output.

3.4. Keyword Extraction

After using the fuzzy sets for making the decision of extracting the weighted terms as keywords in the document, we have now a list of keywords that are extracted using the centre of gravity (COG) with the value greater than the specified threshold value as compared to others

3.5. Topic Spotting

The keywords obtained by implementing the fuzzy sets logic are then used by the classifying algorithm to find the relation between diseases and symptoms and categorize the text documents under the MeSH and OMIM vocabulary.

4. Results

Using the Fuzzy rules, a number of keywords are extracted from each document to be used for assigning topics to these documents.

The fuzzy rules makes a decision based on the list of frequent words left in the document by estimating the importance of each word through fuzzy inference engine which compute them as significant or not significant through *'Yes'* and *'No'*. The keywords in the document having a membership value as 'yes' are considered as candidate keywords and the remaining others with the membership value as 'no' are removed from the list. The membership function for *'Yes'* and *'No'* lies between the value 0 and 1. The candidate words from the document are eventually matched with the MeSH and OMIM vocabulary which contains the titles of different articles in biomedical domain.

5. Document Data Sets

In our work, we have used data sets from a biomedical knowledge base of diseases used for categorization of unlabelled biomedical documents. For testing, we have selected 5 journals in biomedicine from PubMed, each containing 45-80 scientific publications.

Since our focus is on extracting information and finding relation between the symptoms and its diseases for topic spotting, we first chose a vocabulary thesaurus Medical Subject Headings (MeSH) as training dataset. MeSH is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. The information in MeSH is continuously revised and updated. Each entity in MeSH has its own unique id, heading and its tree number and additional names as entry terms. There are 16 classes of entities defined in MeSH but our interest lies only in disease entity with its symptoms and signs [11].

Second, we have a comprehensive vocabulary called the Online Mendelian Inheritance in Man (OMIM), it includes the information on diseases such as human genes and genetic phenotypes. The dataset also contains significant information on symptoms caused by diseases [11].

By combining MeSH and OMIM, we can collect information regarding relation between symptoms and diseases and can be able to label the documents according to the headings in MeSH vocabulary.

6. Conclusion and Future work

In our work, we have presented a novel approach for solving the issues related to the huge body of biomedical text for extracting valuable information important for diagnosing the disease earlier and accurately using an application of text categorization known as topic spotting accompanied with fuzzy sets theory. We have introduced five phases in our approach for automatic topic spotting of biomedical publications and illustrated each phase in detail. We have also given a brief analysis of state-of-the-art applications of text categorization in various areas of document management. The dataset is provided by the PubMed database. We have used a controlled vocabulary thesaurus MeSH and OMIM for categorizing the documents according to their related topic. We have observed that different applications of text mining has been broadly used in the area of biomedicine as the rapid growth of the research work in this field has made it totally expensive and impossible to tackle the information manually.

Although huge work is done by applying different techniques of text mining in the area of biomedicine, it still requires more effort. We should develop applications that are helpful not only for extracting useful information from text documents but also making valuable decisions on their basis by creating relationships between them through semantic analysis of data.

Acknowledgements

College of Electrical & Mechanical Engineering (NUST) is responsible for funding my work and research.

References

[1] U. Y. Nahm and R. J. Mooney, "Text Mining with Information Extraction," in *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, Stanford, CA, March 2002, pp. 60-67.

[2] R. Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords," *International Arab Journal of e-Technology*, vol. 1, no. 4, pp. 164-168, 2010.

[3] Nidhi and V.Gupta, "Recent Trends in Text Classification Techniques," *International Journal of Computer Applications*, vol. 35, no. 6, 2011.

[4] C. J. Taeho, H. S. Jerry, and H. Kim, "Topic Spotting on News Articles with Topic Repository by Controlled Indexing," in *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineers, and Intelligent Agents*, December 2000, pp 386-391.

[5] A.T. Sadiq and S. M. Abdullah, "Hybrid Intelligent Techniques for Text Categorization," *International Journal of Advanced Computer Science and Information Technology*, vol. 2, no. 2, pp. 23-40, 2013.

[6] J. M. G. Hidalgo, J. C. Cortizo, E. P. Sanz, and M. E. Ruiz, "Concept Indexing for Automated Text Categorization," in *Proceedings of 9th International Conference on Applications of Natural Language to Information Systems, NLDB*, Salford, UK, 2004, pp. 195-206.

[7] J. He, H. Yan, and T. Suel, "Compact Full-Text Indexing of Versioned Document Collections," in *Proceedings of the 18th ACM conference on Information and Knowledge Management*, New York, NY, 2009, pp. 415-424.

[8] A. Addis, "Study and Development of Novel Techniques for Hierarchical Text Categorization," PhD Thesis, Electrical and Electronic Engineering Dept., University of Cagliari, Italy, 2010.

[9] FireLin. (2009, February). Fuzzinator- A Fuzzy Logic Controller [Online]. Available: http://www.codeproject.com/Articles/33214/Fuzzinator-A-Fuzzy-Logic-Controller, Feb 2009

[10] R. Khoury, F. Karray, and M. Kamel, "A Fuzzy Classifier for Natural Language Text using Automatically-Learned Fuzzy Rules," in *Proceedings of the 2nd International Conference on Artificial and Computational Intelligence for Decision, Control and Automation Tozeur*, Tunisia, November 2005.

[11] Min Ye, "Text Mining for Building a Biomedical Knowledge Base on Diseases, Risk Factors and Symptoms," Master's Thesis, Saarland University Center for Bioinformatics, Germany, 2011.