www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12374-12378

Systematic Approach to Extract Medical Relations using Machine Learning Approach

Megha M Khanikar, Ramesh Bhat
Mtech Student, dept of CSE
PESIT, Bangalore
Assistant Prof. of CSE
PESIT, Bangalore

Abstract: ***The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. Extracting medical relations is very trivial task since the medical information is stored in textual format and the database of medical information is also very large in size for example Medline is the medical database that contains 21 million records from 5000 selected publications. In addition to that web page containing medical information also contains some unrelated contents like advertisements, scroll bars, quick links, related searches etc., manually extracting only relevant information from such a huge database is very difficult task. To reduce user overhead of extracting useful information current approach is proposed.***

This approach presents the efficient machine learning algorithm and techniques used in extracting disease symptom and treatment related sentences from Medline. In this approach Multinomial Naive Bayes algorithm and several other techniques are used to extract semantic relation between disease symptom and their associated treatment. The proposed system gives the user exactly the Disease Symptom and Treatment related sentences by avoiding unnecessary information and this technique can be integrated with any medical management system to make better medical decisions.

Keywords: Machine Learning, Natural language processing, Medline.

I. INTRODUCTION

The MEDLINE database is a rich source of information for the biomedical sciences, providing bibliographic information and abstracts for more than nine million articles. A fundamental limitation of MEDLINE and similar sources, however, is that the information they contain is not represented in structured format, but instead in natural language text. The goal of our research is to develop methods that can inexpensively and accurately map information in scientific text sources, such as MEDLINE in, to a structured representation, such as a knowledge base or a database. Toward this end, we are investigating methods for automatically extracting key facts from scientific texts.

Medical Information database is very large in size and it is stored in textual format. Extraction of Relations from the text document is very trivial in natural language processing. Manually extracting the useful information from the medical text document is difficult task. In addition to that HTML pages containing medical information also includes some unwanted information like quick links, scroll bars, advertisements, related searches, suggestions etc.,

Researchers are faced with the difficulty of reading a lot of research papers, research articles, and medical thesis to gain knowledge in their field of interest due to the increase in the number of research articles, papers and thesis. Search engines like Pub Med reduces this constraint by retrieving the relevant document related to the user query. Though the relevant document is retrieved, the web page displaying it may contain many non-informative contents like advertisement, scroll bars, menus, citations, quick links, announcements, special credits, related searches, similar posts searched etc. This may cause difficulty in extracting only relevant information. All these unrelated contents are removed and text mining is performed on the extracted document from which information or sentences related to user specified disease are extracted. From the extracted file symptoms, causes, treatment of the particular disease is filtered and displayed to the user. Thus the user gets the required information alone which saves his time and improves the quality of the result.

In the proposed approach various machine learning algorithms are combined with Natural language processing in order to extract relation between the disease and treatment.

II. RELATED WORK

- Janani.R.M.S, Ramesh.V "Efficient Extraction of Medical Relations using Machine Learning Approach" International Journal of Advanced Research in Computer Science and Software Engineering, March 2013. In this author explained how to extract the text content from the html document and how to classify medical related information from the retrieved text document, and also explains formulas for calculating quality measures such as precision recall and F measure.
- Hersheeta Chandankar, Tanushree Chaubal, Rakshanda Bhat "Machine Learning And Data Mining Approach To Identify Disease Treatment Relations" Proceedings of 4th International Conference, 1st December 2013, New Delhi India. In this paper author explained a machine learning methodology for building an application that is capable of identifying and disseminating medical relations, it extracts sentences from published medical papers and extracts semantic relations that exists between diseases and treatment.
- The most relevant work is the work done by Rosario and Hearst in "Semantic Relation In Bioscience Text" where Hidden Markov models are used for entity recognition. This includes mapping biomedical information into structural representation. It involves converting natural language text into structural format. Their work uses machine learning for information extraction. The extraction of medical abstract is obtained through text classification. Semantic lexicons of words labeled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query.
- Khan Razik, Dhande Mayur, Patil Aniket, Gaikwad Namrata "To Identify Disease Treatment Relationship in Short Text Using Machine Learning & Natural Language Processing". In this author explained how machine learning and NLP can be used to extract knowledge from medical papers.
- P.Bhaskar, E.Madhusudhana Reddy "Efficient Machine Learning Approach for identifying Disease-Treatment Semantic Relations from Bio-Medical Sentences" International Journal Of Computational Engineering Research. In this paper author described ML based methodology for constructing an application that is capable of identifying and disseminating healthcare information.
- P.Menaka, Prof.D.Thilagavathy "Identifying Semantic Relations for Disease-Treatment in Medline" International Journal of Electronics and Computer Science Engineering. In this paper author describes which classification algorithms and representation techniques are suitable for identifying and classifying relevant medical information in short texts.
- Bharti E. Nerkar, Sanjay S. Gharde "Identifying Best Treatment For Disease Using Machine Learning Approach In Relation To Short Text" Iraj International Conference, 29th December 2013. In this author explained how to integrate the computer based systems into healthcare care fields and build an application that is capable of identifying and disseminating disease and treatment related information.

III. PROPOSED SYSTEM

In order to extract disease treatment relation, first save the html file that contains information about the user specified disease from Medline in the form of .html file in the user specified database then convert the html document into text document by removing all the unwanted tags, frames, images and extracting only the text content of the document and save it in .txt format.

Extracted text document processed by various classification algorithms, representation techniques. System architecture is shown in the fig.1.

Note that all the process must be followed in pipelined manner in order to achieve a high quality result. Now the extracted text file contains many stop words like a, an, is, for, of, and the words ending with ing, ed etc. These words can be removed to improve the quality of the result. Thus stop words are removed from the extracted text file.

Now the stop word removed text file is subjected to the combination of certain words in order to avoid repetition such as excessed, excessing, if both these words appear in the document we shall reduce the word count just by using stemming algorithm.

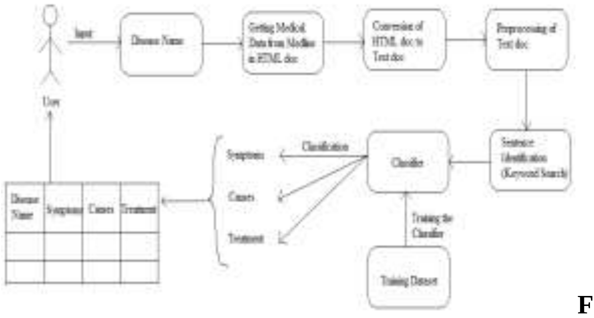
Input
User
Disease Name
Getting Medical Data from Medline in HTML doc
Conversion of HTML doc to Text doc
Preprocessing of Text doc
Sentence Identification (Keyword Search)
Classifier
Classification
Symptoms
Causes
Treatment
Training the Classifier
Training Dataset
Disease Name
Symptoms
Causes
Treatment

Fig.1

System Architecture

It is very common that all the documents will contains such repetition of words for user to get clear understanding of sentences or information. By removing such suffixes and combining these kind of sentences the content of the document is reduced but the quality of the document is increased by reducing the word count and describing the information in simpler form.

After applying stemming algorithm, the semantic relations should be extracted from the above processed text file. Here the semantic relation is the information related to Symptoms, Causes and Treatment of certain disease in the user uploaded html file. In order to extract this semantic relations a classification algorithm namely Multinomial Naïve Bayes classification algorithm is used in association with Bag-Of-Word representation technique.

A. Html to Text Conversion

The saved .html document is converted into a text file and is stored with .txt extension. The stored text file contains Disease-Treatment relation along with the details of the code involved in designing the page, forms, suggestion box, navigation menus, advertisement, feedback, etc. In this text file, the classification algorithm and representation techniques and other analysis algorithms are used to obtain disease-Treatment relation.

B. Extraction of Informative Data

Bag-Of-Word (BOW) representation is used for text classification where each of the word is used as feature for training the classifier. BOW represents a document as a histogram of word occurrences. Such representation is unable to maintain any sequential information. In the proposed work Weighted Bag-Of-Word representation is used which overcomes the above mentioned problem of BOW.

1) Weighted Bag-Of-Word Representation:

Weighted BOW uses local smoothing to embed documents as smooth curves in the multinomial simplex thereby preserving valuable sequential information. Weighted BOW is able to robustly capture medium and long range sequential trends in the document.

C. Sentence Identification and Relationship Extraction

The sentence identification and relation extraction task involves identifying Disease related sentences and its treatment relationship.

1) Multinomial Naïve Bayes Classification Algorithm:

To resolve the above problem and to result in efficient sentence identification Multinomial Naïve Bayes classification algorithm is used in the proposed system. Multinomial Naïve Bayes classification (MNB) algorithm adopts parameter learning method. Disease sentence are identified using MNB algorithm in which performance of the classifier is improved by adopting certain features of Compliment Naïve Bayes Classifiers.

D. Output Performance Evaluation

The performance of the proposed model is tested with the html page from MEDLINE containing information about user specified disease. The end result was a text file containing only the information about the disease mainly on symptoms, causes and treatment with increased rate of precision compared to the .html file which was given as input.

IV. VALIDATING THE PROPOSED SYSTEM

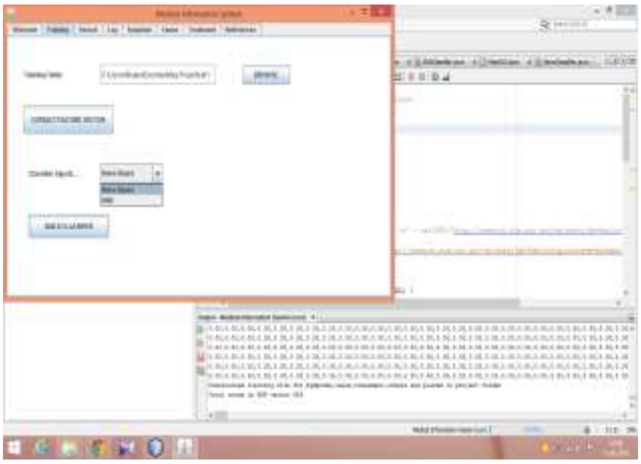

Fig 2 Preparing feature vector using training dataset and training the classifier.

Fig 3 Retrieving disease related information from Medline database, storing in html format and pre-processing obtained html document.

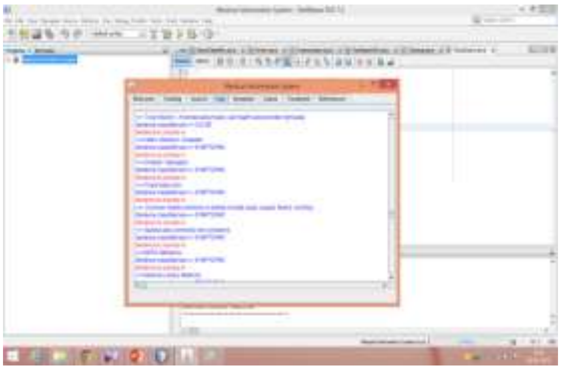

Fig 4 Information Retrieved from Medline database and classified to symptoms, causes and treatment.

V. CONCLUSION

The proposed system removes the unwanted contents from the HTML page from MEDLINE and result on a text document containing only the particular disease and its relevant Symptoms, Cause and Treatment. Experimental result shows that the technique used in the proposed work minimizes the time and the work load of the doctors in analysing information about certain disease and treatment in order to make decision about patient monitoring and treatment. This text mined document can be used in medical health care domain where a doctor can analyse various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies.

REFERENCES

[1] Janani.R.M.S, Ramesh.V "***Efficient Extraction of Medical Relations using Machine Learning Approach***" International Journal of Advanced Research in Computer Science and Software Engineering, March 2013.

[2] Hersheeta Chandankar, Tanushree Chaubal, Rakshanda Bhat "***Machine Learning And Data Mining Approach To Identify Disease Treatment Relations***" Proceedings of 4th

International Conference, 1st December 2013, New Delhi India.

[3] B.Rosario And M.A.Hearst**,** ***"Semantic Relation In Bioscience Text"***, Proc. 42nd Ann. Meeting On Assoc For Computational Linguistics, Vol.430,2004.

[4] Khan Razik, Dhande Mayur, Patil Aniket, Gaikwad Namrata "***To Identify Disease Treatment Relationship in Short Text Using Machine Learning & Natural Language Processing***" Journal of Engineering, Computers & Applied Sciences, April 2013.

[5] P.Bhaskar, E.Madhusudhana Reddy "***Efficient Machine Learning Approach for identifying Disease-Treatment Semantic Relations from Bio-Medical Sentences***" International Journal Of Computational Engineering Research vol.2 September 2012.

[6] P.Menaka, Prof.D.Thilagavathy "***Identifying Semantic Relations for Disease-Treatment in Medline***" International Journal of Electronics and Computer Science Engineering vol 1 number 2 Sep 2010.

[7] Bharti E. Nerkar, Sanjay S. Gharde "***Identifying Best Treatment For Disease Using Machine Learning Approach In Relation To Short Text***" Iraj International Conference, 29th December 2013.