International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 07 , July 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

CLUSTERING OF MEDICAL DOCUMENTS USING SYMPTOM/MEDICATION NAMES

Farhat Mulla[1], Prakash H.Unki[2]

[1]M.Tech (CSE) Student, [2]Associate Professor, Department of CSE,
BLDEA's Dr.P.G.Halakatti College of Engineering and Technology, Vijayapur, Karnataka, (India)

ABSTRACT

Clinical documents contain huge amount of medical data. This data can be used for medical treatment of various diseases and their symptoms along with their recommended medications. Data mining techniques are applied on this clinical data, which is an important source to improve the current healthcare system by making it more proficient. In this paper, we have developed a method for clustering of the clinical documents. The clinical documents are obtained from different hospitals and websites. Pre-processing of textual data is done to amplify the performance of Clustering. We have used different tools like MedEx and MetaMap to fetch essential data from clinical narratives. MedEx is used to fetch medication data and MetaMap is used to get symptom names which are related to patient. The multi-view Non-negative matrix factorization (NMF) is used to cluster clinical documents. Our experimental result shows that, as the number of clinical documents increases processing speed decreases to get the result.

Keywords: Clinical document, clustering, Non-negative matrix factorization (NMF), data mining.

I. INTRODUCTION

Data mining [1] is a well-known approach for knowledge discovery in database systems. It is an efficient way of finding useful information from huge amount of data. It is a branch of artificial intelligence (AI) method, which is used to extract vital data from enormous amount of data. The knowledge that we get through this technique can be used for further innovation and collaboration. There are many applications of data mining in medical field, as it has wide spread use in medical area. It is getting great pace in medical research as well as in clinical practice. Clinical data mining is nothing but mining clinical data, so as to get essential data based on our requirement. Clinical documents contain textual data. By applying data mining technique on these data, we can fetch key information like medication names and symptom names from clinical narratives. Information extraction is important task in case of machine learning (ML) and natural language processing (NLP), as it involves significant data extraction from natural language text.

Extraction of these essential data helps health care provider to advance health care system. Clinical document plays a vital role in analysis and diagnosis of disease [2]. Mining of vital data in medical field involves, handling number of important tasks like recognition of medical related terms, recognition of attributes such as negation, severity, uncertainty and mapping words in document to concept in domain specific ontologies. The entire procedure depends on many different types of NLP processes such as tokenization, parsing, and part of speech

tagging. We can also make use of many specific resources like dictionaries and ontologies such as the unified medical language system (UMLs) [3]. In our proposed method, our interest is only in medication names and symptom names. Therefore, to extract medication names and symptom names, we are using two annotators, namely medication annotator and symptom annotator. Medication annotator is an annotator which is based on MedEx [4] and symptom annotator is based on MetaMap.

Recently, Non-Negative Matrix Factorization (NMF) is getting a lot of attention in document clustering as well as in information retrieval [5]. Document clustering is a fundamental technique for grouping data based on similarity and dissimilarity and then dividing theses data into subsets. Each subset is having similar kind of data or we can say that each item present in subset is having same characteristic. Document clustering provides an intelligible summary of the collection, which can be used to provide random vision and to locate important patterns [6, 7]. Document clustering is a fundamental technique for content summarization [8], cluster-based information retrieval [9] and automatic topic extraction [10]. Clustering has shown remarkable progress in past decade [11, 12, 13, 14]. In our proposed method, we have used multi-view NMF for clustering. Multiv-view NMF provides more efficient result than simple NMF [15].

II. LITERATURE SURVEY

In [16], authors have proposed a method for data extraction from online medical forums. Lexico-syntactic pattern from annotated information with seed vocabularies is used to extract two entity types, namely, treatments and drugs. To get the efficient result, symptoms & conditions (SC) and drugs & treatments (DT) term are used from compiled online dictionary. In order to extract SC and DT term lexico-syntactic pattern are iteratively brought strongly to each entity. This proposed system extracts symptom names and the treatments, which are absent from original vocabulary dictionary.

In [17], structured data is extracted from clinical narratives using rule based method along with machine learning technique and feature engineering. Conversion of unstructured clinical narratives into structured narratives is of great demand. This proposed system performs three tasks such as, relation identification, assertion classification and concept extraction. In the case of concept extraction, switching model is verified so that extraction of treatment information can be improved.

In [18], authors have proposed a scheme to extract entity extraction using local grammar. In this method, medical related information from French clinical notes is extracted using rule based local grammar. There is difficulty in identifying medical data because of high terminological variations in medical domain. The disadvantage of this method is that a great human effort and more time are required. The experimental result of this proposed approach allows a good precision in complex structure of clinical document.

In [19], authors have worked on a method called automated de-identification and large scale evaluation of clinical notes. A NLP tool is used for automatic identification of large set of different clinical sets. The performance of human annotator is challenged by a NLP based de-identification method. This proposed approach helps in de-identification of millions of clinical documents.

In [20], authors have proposed a scheme to extract essential data from clinical notes using natural language information extraction method. This proposed method is found to be very effective to identify essential data from clinical narratives and relations in text by doing research in information extraction.

In [21], authors have proposed a method to recognize named entity in clinical notes using cascading classifier. In order to reduce misclassification, a support vector machine (SVM) and maximum entropy (ME) is used to reclassify the identified entities using a proposed cascading system.

III. METHODOLOGY

In our proposed system, we have used different tools like medication annotator, symptom annotator, section annotator and negation annotator. Medication annotator is based on MedEx, which is used to extract medication names. Symptom annotator is based on MetaMap, which is used to extract symptom names. We have used section annotator to identify different sections of clinical documents. We have used negation annotator to remove negation words such as avoid, deny etc.

RECORD #1

505128233 | RH | 36002733 | | 8399692 | 10/18/2005 12:00:00 AM | SMALL BOWEL OBSTRUCTION | Signed | DIS | Admission Date: 10/18/2005 Report Status: Signed
Discharge Date: 11/14/2006
ATTENDING: DEMMER , ROBT MD
PRIMARY CARE PHYSICIAN: Dustin Theurer , MD
CARDIOLOGIST: Hassan Gowins , MD
GASTROENTEROLOGIST: Coy Spurgers , MD
PRINCIPAL DIAGNOSES:
1. Small-bowel obstruction.
2. Congestive heart failure.
3. Dilated idiopathic cardiomyopathy.
HISTORY OF PRESENT ILLNESS: Mrs. Aswegan is a 74-year-old woman , recently admitted to the Kimonte on 7/8/05 for CHF exacerbation and UTI (Gram-negative rods Klebsiella) who presents to the Vide Tutenoke Rahbrier Healthcare with one day of abdominal pain , nausea , vomiting , and decreased ostomy output. The patient has a history of multiple abdominal surgeries including a total colectomy for ulcerative colitis with colostomy , ventral hernia repairs in 1998 , 2003 , and 2004 , revision of her colostomy in 1996 , 1997 , and 2003. As a result , she has had a chronic left lower quadrant hernia. The patient denies any recent fevers , chills , melena , or hematochezia from the ostomy. She has stable dyspnea on exertion after 12 feet, she uses a walker at home. She denies PND. She has two-pillow orthopnea. She has had stable lower extremity edema in the past day and no dysuria.

Fig.1: An example of clinical document

The initial phase of our methodology is pre-processing of clinical document. These clinical records contain crucial information about patient, i.e. patient's medications, symptoms, history of present illness, history of past illness, hospital course history. An example of such a clinical document is shown in Fig.1.

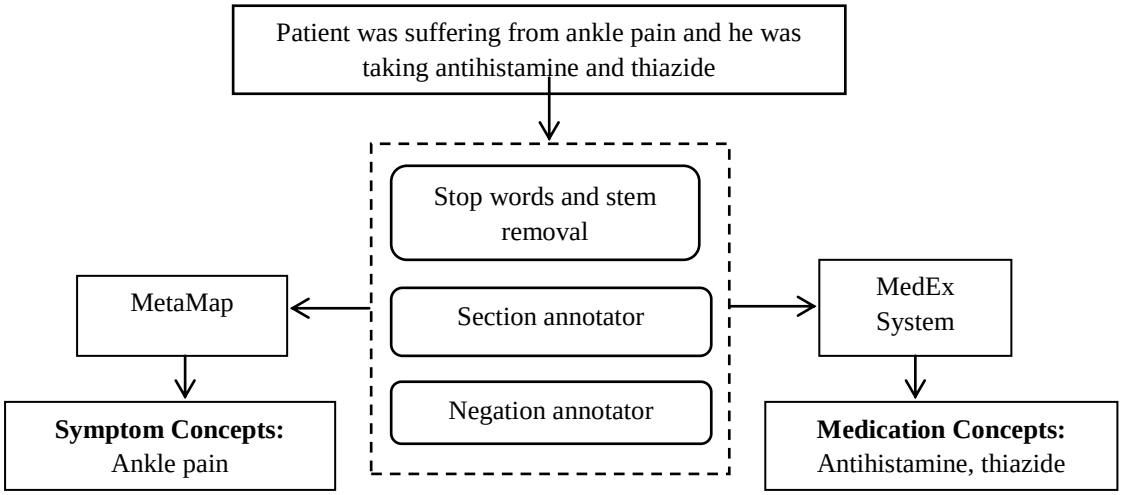

Fig.2: Block diagram of symptom and medication names extraction from clinical notes

The Fig.2 shows the block diagram of a symptom and medication names extraction from clinical notes. We have clinical text "Patient was suffering from ankle pain and he was taking antihistamine and thiazide". In this clinical text, symptom name is ankle pain. There are two medicines, namely, antihistamine and thiazide. Initially, Section annotator is used to find different sections in clinical document. In order to fetch these medication and symptom names, first we need to remove irrelevant words, using pre-processing. The pre-processing also improves quality of data. StandfordCore NLP tool is used to isolate words and sentences from clinical document. During pre-processing, we have to take out stop words and stem words which are nothing but most common words in English language, such as, this, that, she, he, etc. The output of stop words and stem words removal module is the medical terms along with the negation words. Negation annotator module is used to eliminate negation words like avoided, denies, ruled out etc. The output of negation annotator is only medical related term. The output of pre-processed data is fed into MedEx and MetaMap systems to get medical related terms. The medical related terms include medicines and symptoms.

After removing unnecessary content from clinical notes, we are clustering medical related data. We have used multi-view NMF for clustering. Multi-view NMF finds latent components in sub matrices. When a user enters the problem statement containing symptom name, proposed method provides medication names using MedEx system and symptom names using MetaMap. Finally, system provides Ankle pain as symptom name and medication names, such as Antihistamine and thiazide.

IV. EXPERIMENTAL RESULTS

We have carried out our experiments on 50 clinical documents of different patients suffering from different diseases. We have stored these clinical documents at the back end by creating one folder in our programming package. The programming package that we have used is NetBeans IDE 7.2.1.

Initially, we have pre-processed clinical notes to remove unnecessary content and we have stored clustered data in one file. We have extracted medication names using MedEx and symptom names using MetaMap. Whenever user enters the query related to symptom, our proposed system extracts related medication and symptom names

from stored file. During pre-processing, timer was set and processing speed for different clinical data set was recorded.

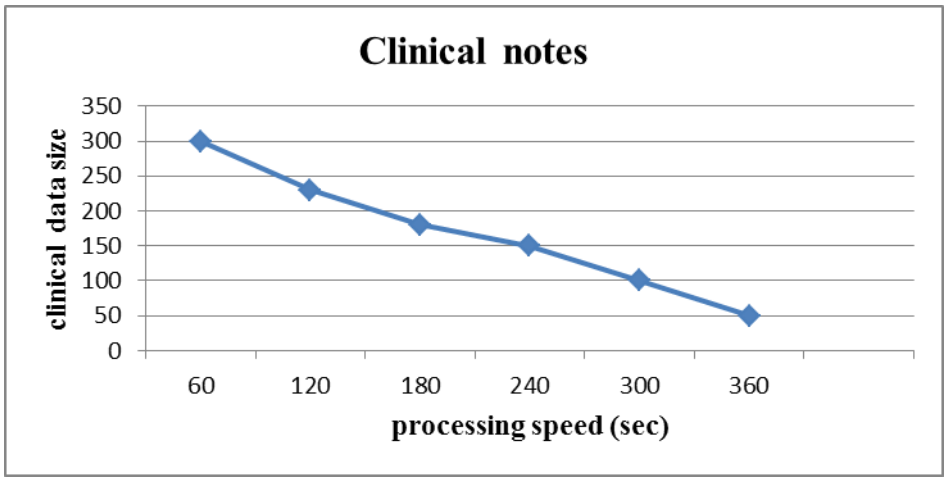

Fig.3: Processing speed with respect to clinical data set

Finally, we conclude that the processing speed increases as the number of clinical documents decreases, which is shown in Fig.3. If we have less number of clinical documents then we get less efficient results. In order to get more efficient results, we need to have more number of clinical documents.

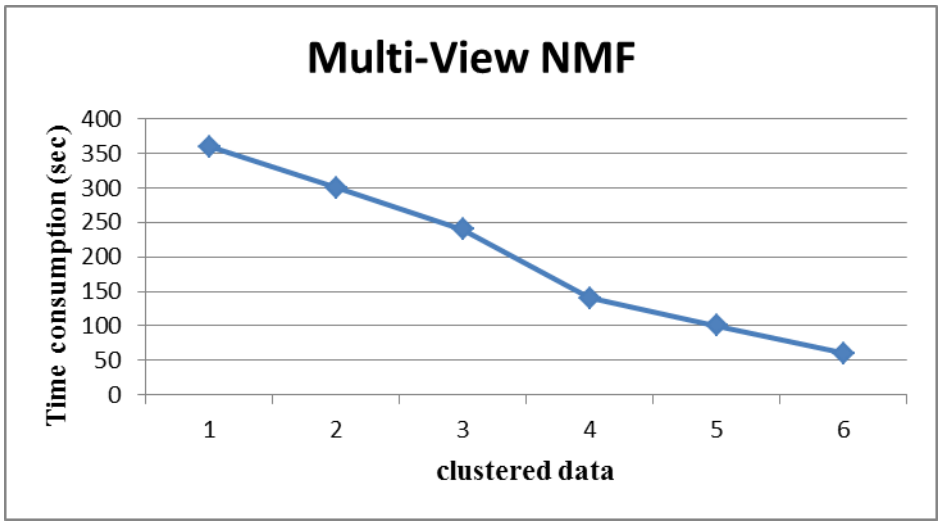

Fig.4: Time consumption with respect to clustered data

Time consumption to get the multi-view NMF results increases, as clustered data decreases. Our proposed system shows 80-85% accuracy. To get more accurate results, we need to have more number of clustered data, which is shown in Fig.4.

V. CONCLUSION AND FUTURE WORK

This proposed system extracts medication and symptom names from clinical documents. We have used section annotator, negation annotator, symptom annotator, and medication annotator tools to get different sections of clinical documents as well as medication names and symptom names. Proposed system also shows the accuracy of medication association with each symptom. Pre-processing before clustering provides an efficient result. Multi-view NMF is used to cluster clinical documents, which provides better performance than simple NMF. In

future, we may consider extraction of patient age/sex/gender information. We will also try to improve processing speed of our proposed system.

REFERENCES

[1] Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal, and Donald C. Trost, "Information Mining Over Heterogeneous And High-Dimensional Time-Series Data In Clinical Trials Databases", IEEE Transactions On Information Technology In Biomedicine, 2005, Vol. 10, no. 2.

[2] G. Hripcsak Et Al., "Mining Complex Clinical Data for Patient Safety Research: A Framework For Event Discovery", J. Biomed. Informat., 2003, Vol.36, no. 1, pp. 120–130.

[3] Jacquemin, C., "Spotting and discovering terms through natural language processing", MIT Press, 2001, http://books.google.com/books?id=W6AB06SBAGMC.

[4] Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., and Denny, J.C., "MedEx: a medication information extraction system for clinical narratives", Journal of the American Medical Informatics Association, 2010, pp. 19-24.

[5] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization", In Proceedings of the 26th Annual International ACM SI-GIR Conference on Research and Development in Informaion Retrieval (SIGIR), 2003, pp. 267-273.

[6] Hung Chim and Xiaotie Deng, IEEE "Efficient Phrase-Based Document Similarity For Clustering", IEEE Transactions On Knowledge And Data Engineering, 2008, Vol. 20, no. 9.

[7] F. H. Saad, B. D. L. Iglesia, and D. G. Bell, "A Comparison Of Two Document Clustering Approaches For Clustering Medical Documents", In Proc. Conf. Data Mining (DMIN), 2006.

[8] X. Wan and J. Yang, "Multi-document summarization using cluster-based link analysis", In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2007, pp. 299-306.

[9] X. Liu and W. B. Croft, "Cluster-based retrieval using language models", In Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2004, pp. 186-193.

[10] J. Silva, J. Mexia, A. Coelho, and G. Lopes, "Document clustering and cluster topic extraction in multilingual corpora", In Proceedings of the 1st IEEE International Conference on Data Mining (ICDM), 2001, pp. 513-520.

[11] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, pp. 45–55.

[12] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang, "Graph regularized non-negative matrix factorization for data representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, pp. 1548–1560.

[13] R. Arora, M. Gupta, A. Kapila, and M. Fazel, "Clustering by left stochastic matrix factorization", In ICML, 2011.

[14] I. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts", In KDD, 2004.

[15] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," in Proc. 16th Comput. Vis. Winter Workshop, 2011.

[16] Sonal Gupta, Diana L MacLean, Jeffrey Heer, and Christopher D Manning, "Induced lexico-syntactic patterns improve information extraction from online medical forums", J American Medical Informatics Association, 2014, pp.902-909.

[17] Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang, "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries", American Medical Informatics Association, 2012, pp. 1-9.

[18] Aicha Ghoulam, Fatiha Barigou, Ghalem Belalem ,and Farid Meziane, "Using local grammar for entity extraction from clinical reports", International Journal of Artificial Intelligence and Interactive Multimedia, 2015, vol. 3, no. 3, pp.16-24.

[19] Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti, "Large-scale evaluation of automated clinical note de-identification and its impact on information extraction", American Medical Informatics Association, 2012, pp. 1-11.

[20] Raymond J. Mooney and Razvan Bunescu, "Mining knowledge from text using information extraction", SIGKDD Explorations, Volume 7, Issue, pp. 1-3.

[21] Yefeng Wang and Jon Patrick, "Cascading classifiers for named entity recognition in clinical notes", Workshop Biomedical Information Extraction, 2009, pp. 42-49.