Available online at www.ijpcr.com
International Journal of Pharmaceutical and Clinical Research 2016; 8(1): 60-64

ISSN- 0975 1556

Research Article

Enhanced K-means Clustering Approach for Health Care Analysis Using Clinical Documents

Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M

SCSE Department of VIT, Vellore Institute of Technology, Vellore, Tamil Nadu, 632014

Available Online:31st December, 2015

ABSTRACT

Clinical documents contain enormous amount of medical information. These documents are gold mine of information for medical treatment of various diseases and their symptoms along with their prescribed medications. Data mining techniques when applied on this clinical data is vital source to improve the current healthcare system by making it more efficient. We define an approach to build a system that firstly pre-processes the clinical documents. Pre-processing of textual data will amplify the performance of Clustering. Then we apply the K-means clustering on the pre-processed notes. Extraction of symptoms and medication names on the clustered data results in improved medication recommendation. Our experiments show that K-means clustering is a favored approach for clustering of clinical documents.

Keywords: Clinical note; clinical document; document clustering; K-means Clustering; Machine learning

INTRODUCTION

Data mining[1] is well known to be a major source of knowledge discovery in databases. It is considered as an artificial intelligence method that allows finding useful information residing in bulk of data. It has shown a great potential to extract useful data from an exhaustive collection of data and facts. The knowledge obtained via this process might contain additional information that can be utilized for further discovery and collaboration. Application of data mining concepts to the medical domain has undoubtedly achieved great pace in the region of medical research and clinical practice ; thus saving time, money and life. Clinical data mining is the process of applying the data mining techniques on the obtained textual clinical documents. Rich text data sources of clinical documents contain information about medication and symptoms. Extracting this information has proved beneficial so as to help refine the health care system. Clinical documents[2] are widely used for future analysis and diagnosis of the disease.

Clustering is the procedure of grouping the similar objects into clusters. Clustering[3] these documents provides an intelligible summary of the collection, which can be used to provide arbitrary vision.

The clinical notes have a great use in pharmacy store so to reduce forgery and prevent drug abuse. Documents clustering for clinical notes is been to research for grouping them into relevant clusters. This is done primarily to locate important patterns[4]. This has proved beneficial by increasing speed, efficiency and accuracy of managing information in the area of medical diagnosis.

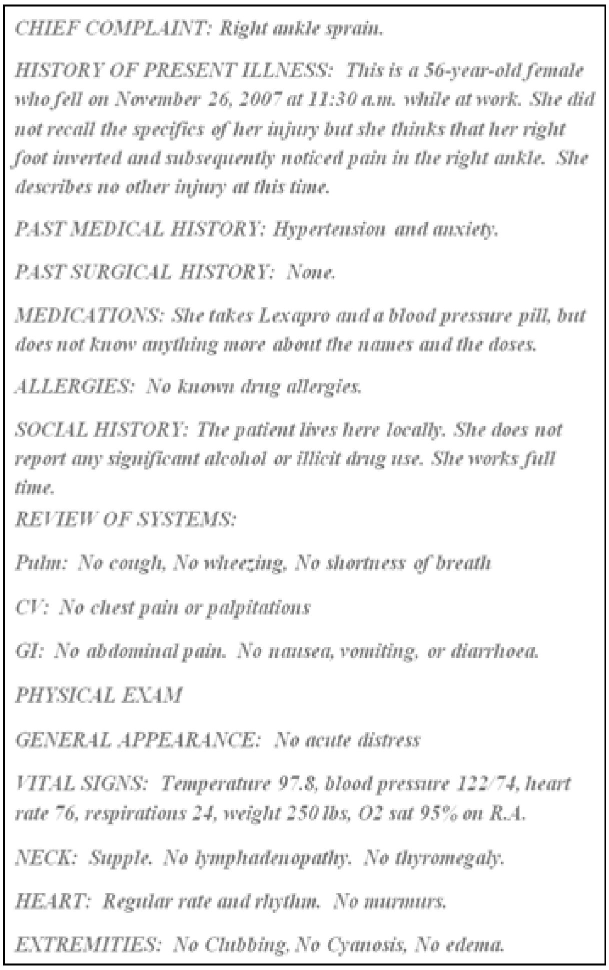
K-means[5] clustering is well known widespread clustering technique. It is mostly used to cluster the numerical data. Using this technique we have performed clustering on textual data corpus[6] that is in unstructured and semi structured format. This is done by considering various factors which is described in methodology section.

The following paper is organized as follows: Section II describes about related work done. Section III describes our methodology and Section IV describes the experimental results and Section IV concludes our experiments.

Literature survey

Health care information via Clinical source has seen a significant increase in both volume and variety. A significant portion of the relevant data is located in an unstructured or else semi-structured clinical text of documents stored in disparate repositories. Clinical documents containing records of patients' condition during the discharge, prescriptions for the treatment, lab reports, and free-form physician notes are filled with various abbreviations, acronyms, misspelled words and grammatically incorrect phrases. Furthermore, they have been superseded by systems such that the clinical concepts in these texts are extracted using the Natural Language Processing (NLP). In order to address the growing need for an efficient NLP solutions which can deal with the large volume and variety of clinical text, they have developed a clinical concept extractor that works on optimized Rules-based system, called TRACE[7] (Tactical Rules-based AQL Clinical Extractor) that uses Annotation Query Language (AQL) an open-source clinical text miner, on a set of prescription documents.

**Author for Correspondenc*



CHIEF COMPLAINT: Right ankle sprain.

HISTORY OF PRESENT ILLNESS: This is a 56-year-old female who fell on November 26, 2007 at 11:30 a.m. while at work. She did not recall the specifics of her injury but she thinks that her right foot inverted and subsequently noticed pain in the right ankle. She describes no other injury at this time.

PAST MEDICAL HISTORY: Hypertension and anxiety.

PAST SURGICAL HISTORY: None.

MEDICATIONS: She takes Lexapro and a blood pressure pill, but does not know anything more about the names and the doses.

ALLERGIES: No known drug allergies.

SOCIAL HISTORY: The patient lives here locally. She does not report any significant alcohol or illicit drug use. She works full time.

REVIEW OF SYSTEMS:

Pulm: No cough, No wheezing, No shortness of breath

CV: No chest pain or palpitations

GI: No abdominal pain. No nausea, vomiting, or diarrhoea.

PHYSICAL EXAM

GENERAL APPEARANCE: No acute distress

VITAL SIGNS: Temperature 97.8, blood pressure 122/74, heart rate 76, respirations 24, weight 250 lbs, O2 sat 95% on R.A.

NECK: Supple. No lymphadenopathy. No thyromegaly.

HEART: Regular rate and rhythm. No murmurs.

EXTREMITIES: No Clubbing, No Cyanosis, No edema.

Figure 1: An Example of Clinical Note

Another work carried out in this field is to apply this information (clinical notes) to direct the medical decisions to guide medical decisions is a challenging task. One issue is in presenting information to the practitioner: displaying certain extraneous information in the given clinical notes that often leads to information overload. They hypothesize that, knowledge (e.g. variables, relationships) can be used to annotate and contextualize information from the patients' record, that speeds up the process of accessing the desired parts of the record and thus improving medical decision making. To achieve the stated goal, they described a framework that aggregates as well as extracts information given in free-text clinical reports, which later maps this information to conceptualize in available source of knowledge, and provide a tailored view of the information needs of the user based on the records. A system called Adaptive EHR[8,9], has implemented this framework that demonstrates its capabilities to synthesize and present the information obtained from Neuro-oncology patients' records.

One of the experimental researches is carried out to improve the quality of clinical notes by extraction of medication/symptom names. Medication and Symptoms names are two important types of information that can be retrieved from textual clinical data. This information when extracted will greatly helpful in medication recommendation. This helps in knowing the accuracy of medication related to the particular symptoms. They applied the natural language processing tools to improve the quality of prescription by removing the irrelevant data. They build an integrating system which consists of following modules:

1) System for extraction of Symptom/Medication names from clinical notes

2) Apply multi-view NMF[10] to estimate the results of using medication/symptom names in improving the clinical notes clustering. This helps in analyzing the class of features of patterns.

3) Compared the experimental result of multi-view NMF and NMF.

METHODOLOGY

The following are the phases in our methodology:

Clinical Documents

In Clinical Notes[2] which are vital source for patient's prior information are gathered. They contain important information such as patient's past history of medications, diseases, present medications, symptoms etc. These documents are in unstructured format. An example of such clinical note is shown in figure 1. These measures when put together can greatly enhance the health care. We have performed our experiments on public corpus[11]. Collection of this corpus contains the different types of patients suffering from various diseases. They contain enormous information of medication which is primary concern of any prescription.

Pre-processing of clinical notes

The Collection of enormous textual documents contains considerable amount of structured data which is hidden in unstructured data. Pre processing[12] improves the quality of data. Removal of unnecessary words will help greatly in clustering. Here in pre-processing we are removing irrelevant words such as stop words and stem words. Also to improve the quality of data in terms of relevancy we are using the section annotator[13] to distinguish different sections in the clinical note. This annotator depends on the header information. Negation sections are also included in the clinical note and these are to be excluded. Illustration of this: "She is allergic to Diclofenac" from the Section allergies is to be discarded. Here the medication is Diclofenac is excluded since it is a Negative medication. Proceeding, we use the negation annotator to exclude the negative indicated medication and symptom. Example: "The patient was instructed to avoid the usage of ibuprofen". Here "ibuprofen" is eliminated since the word "avoid" is a pre-negation. We used NegEx to remove negative medication. In this tool pre-negation and post-negation are considered and according to that the names are removed. An overview of the pre-processing of clinical text is seen in fig 2. Here in an example "The patient was kept off aspirin given his rectal bleeding. He had hypertention and was given antihistamine and thiazide diuretics." Aspirin is removed, since it emphasis the negative medication due to the word

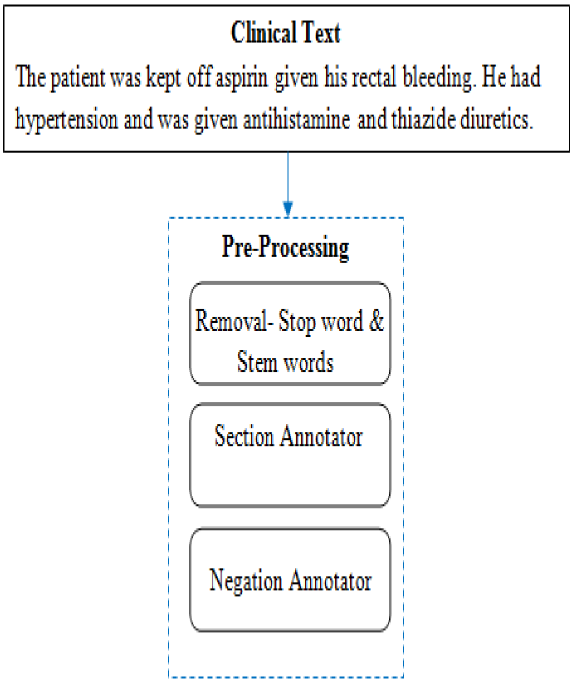

Clinical Text
The patient was kept off aspirin given his rectal bleeding. He had hypertension and was given antihistamine and thiazide diuretics.
Pre-Processing
Removal- Stop word & Stem words
Section Annotator
Negation Annotator

Figure 2: An overview of the pre-processing of the clinical note

'kept off'. The medications "antihistamine" and "thiazide diuretics" and symptom of it "hypertension" are to be extracted as mentioned in sub section d.

Clustering

Clustering[3] is a widely studied data mining technique in the area of text domain. It has diverse applications in real world scenarios. It is one of the main experimental researches in the field of data mining. Clustering of textual data is a way of directing and summarizing the content present in the document which paid heeds. In our statistical analysis we have chosen dataset which contains 100 such medical prescriptions.

After pre-processing the documents, we apply K-means Clustering on the processed documents in the following manner. K-means[4] clustering is usually applied on numerical data which does not need considerations of other computations. But to apply K-means algorithm on the textual data which is in unstructured/semi structured format[14] is to be converted to the numerical form. This can be done by converting the documents as vectors. To do this we compute the term frequency-inverse document frequency and creation of document vectors is to be done. This will help in mapping the most frequent words in the documents and indicates the how essential the word is in corpus. These numerical data is considered in K-means algorithm.

Before applying k-means on the text documents, these documents are represented as mutually comparable vectors using the tf-idf (term frequency-inverse document frequency) value[15]. It ranks the importance of a term in the textual document corpus. Term frequency is calculated as a normalized frequency i.e. it is a ratio of the frequency (number of occurrences) of a word in the document to the total number of words in that document. The inverse document frequency is the log of the ratio of the number of documents in the corpus to the number of textual documents holding that term. These two metrics when multiplied together gives tf-idf value, stating the importance of a term frequent and rare in the corpus.

Then tf–idf is computed as

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

Cosine similarity[16] is the calculation of the similarity between two documents. After converting the documents into document vectors by the previous calculation (tf-idf step) we can determine the similarity metric based on the cosine of the angle between the two document vectors. Each term in the document has its own axis. The formula given below finds the similarity between any two documents.

$$\mathit{Cosine\ Similarity}\ (d1,d2) = \frac{(d1,d2)}{\|d1\| * \|d2\|}$$

After obtaining the two vectors we divide them by the product of their magnitudes. The angle so calculated is a good indicator for the measure of similarity. The cosine of 0° is 1, and for any other angles it is less than 1.

After computing this, K-means clustering is performed. Initially the number of clusters (K) to be formed is decided. The value of K can be chosen as per the requirement. In our experiment we have chosen the value of K as 3. In a wide sense, K-means works by randomly initializing the k number of clusters as centroids. Here we apply an iteration of K-means on the dataset. The following method is employed here, by picking up the K objects and placing the centroid in the same place as those objects. Assignment of each object to it's closely cluster centroid. In the final step we need to update the average value of the cluster to the cluster centroid. Updating and allotment step is done is done periodically until the optimum solution is achieved hereby reducing fitting error.

Extraction of Symptom and Medication Names

Based on the clusters formed in previous steps we are extracting the medication and symptom names from the documents in each cluster. Application of clustering algorithm on extracted data to improve the visibility of huge clinical documents. Extraction of medication names such as Motrin, Tylenol from clinical text fig 1 is done. The patient is suffering from the right ankle sprain that is considered as a symptom. This can be done by using MedEx[17] tool on clinical note to extract the medication name. This tool is used to obtain the medication related details such as Dose volume, Drug name, Intake time of medicine etc. Drug name is considered as medication name. Simultaneously we are applying the MetaMap[18] to get the name of the symptom. MetaMap tool is used to obtain concepts associated to symptoms. The overall system is shown in fig 3. These concepts are helpful in mapping to biomedical texts in the UMLS[19] Meta-thesaurus. Extraction of Symptom names and medication names from clinical documents[7] which are in similar

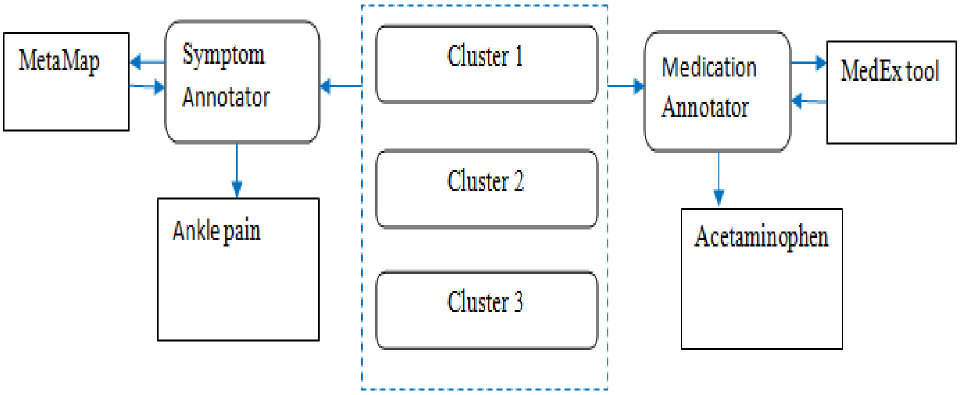
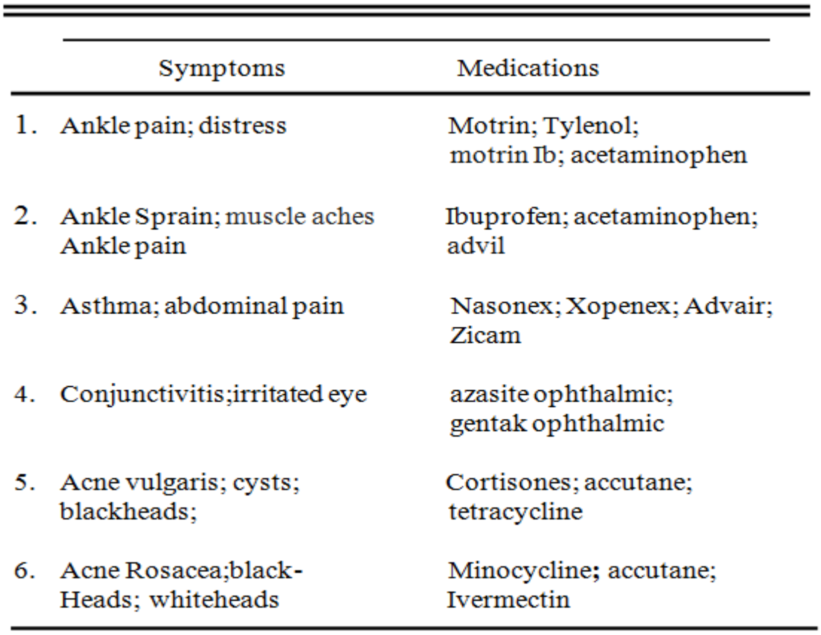


Figure 3: Application of symptom annotator and medication annotator on the cluster of clinical documents



	Symptoms	Medications
1.	Ankle pain; distress	Motrin; Tylenol; motrin Ib; acetaminophen
2.	Ankle Sprain; muscle aches Ankle pain	Ibuprofen; acetaminophen; advil
3.	Asthma; abdominal pain	Nasonex; Xopenex; Advair; Zicam
4.	Conjunctivitis;irritated eye	azasite ophthalmic; gentak ophthalmic

5.	Acne vulgaris; cysts;
	blackheads;
		Cortisones; accutane;
		tetracycline

6.	Acne Rosacea;black- Heads; whiteheads	Minocycline; accutane; Ivermectin

Figure 4: Extraction of symptom names and their related medication names from clusters of documents

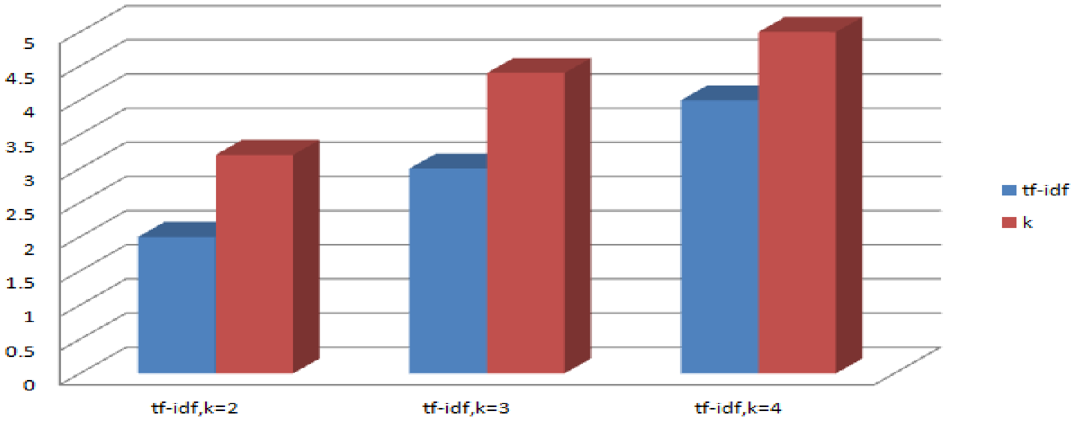


Figure 5: Accuracy of medications to the related symptoms on the basis of clusters.

clusters helps in knowing the exact prescription for particular symptoms[20] as shown in fig.4.

Experimental Result

The dataset we have chosen consist of the public corpus prescription[9]. The dataset consists of 2673 prescriptions of different patients suffering from different things.

We choose different k values to see the results of clustering. On opting k=3, the three clusters of most similar objects are formed taking the lesser time. The value of k can be incremented on our requirements. The results can be seen in fig 5 where for each k value the tf-idf factor is also increasing. The implementation is done in java platform.

CONCLUSION

In this paper we have build a system to know the accuracy of medication associated with each symptom. To do this we have applied K-means Clustering on the clinical note corpus.The document clustering results in improving the medication recommendation. Our experimental results show that pre-processing before clustering results in efficient process of clustering. To this we have used tools such as section annotator, negation annotator, symptom annotator and medication annotator to get different views of clinical notes which improves the visibility of clinical note. This result in increase of the accuracy of medications associated with the symptoms.

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, Vol. 10, No. 2, April 2005*.

2. *G. Hripcsak Et Al.*, "Mining Complex Clinical Data For Patient Safety Research: A Framework For Event Discovery*," J. Biomed. Informat., Vol.36, No. 1, Pp. 120–130, 2003.*

3. Hung Chim And Xiaotie Deng, Senior Member, IEEE "Efficient Phrase-Based Document Similarity For Clustering", *IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 9, September 2008.*

4. 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.

5. Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, And Angela Y. Wu, Senior Member, "An Efficient K-Means Clustering Algorithm: Analysis And Implementation" *IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002.*

6. Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, And Mohamed S. Kamel, Fellow, IEEE,"An Efficient Concept-Based Mining Model For Enhancing Text Clustering", *IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.*

7. Heather Champion, Nick Pizzi, Raja Krishnamoorthy Imt Winnipeg, Canada, "Tactical Clinical Text Mining For Improved Patient Characterization", *2014 IEEE International Congress On Big Data.*

8. William Hsu, Member, IEEE, Ricky K. Taira, Suzie El-Saden, Hooshang Kangarloo, And Alex A. T. Bui, Member, IEEE," Context-Based Electronic Health Record: Toward Patient Specific Healthcare", *IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 2, March 2012.*

9. Andrew Dalley, John Fulcher, David Bomba, Ken Lynch, and Peter Feltham*,"* A Technological Model to Define Access to Electronic Clinical Records", *IEEE Transactions On Information Technology In Biomedicine, Vol. 9, No. 2, June 2005.*

10. *Yuan Ling, Xuelian Pan, Guangrong Li*, and Xiaohua Hu, Member, IEEE,* "Clinical Documents Clustering Based on Medication/Symptom Names Using Multi-View Nonnegative Matrix Factorization", *IEEE Transactions On Nanobioscience, Vol. 14, No. 5, July 2015.*

11. Dataset- (https://idash-data.ucsd.edu/community/45)

12. Todd.Lingren, Louise.Deleger, Haijun.Zhai, Jareen.Meinzen-Derr, Megan.Kaiser, Laura.Stoutenborough, Qi.Li, "Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias."*2012 IEEE Second Conference On Healthcare Informatics, Imaging And Systems Biology.*

13. Emilia Apostolova, David S. Channin MD*, Dina Demner-Fushman MD, PhD, Jacob Furst PhD, Steven Lytinen PhD, Daniela Raicu PhD "Automatic Segmentation of Clinical Texts", *31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, September 2-6, 2009.*

14. Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text,"*J. Amer. Med. Informat. Assoc.*, vol. 17, no. 5, pp. 514–518, 2010.

15. Wen Zhang, Taketoshi Yoshida, Xijin Tang, "TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization", *IEEE International Conference on Systems, Man and Cybernetics, 2008.SMC 2008.*

16. Tengfei Xue, Yuyu Yuan, Qunchao Fu, Heng Gu, Siyue Zhang, Cong Wang, "The Application Of Text Similarity Computing In The Clinical Decision Support System", Proceedings of CCIS2014.

17. H. Xu *et al.*, "MedEx: a medication information extraction system for clinical narratives," *J. Amer. Med. Informat. Assoc.*, vol. 17, no. 1, pp. 19–24, 2010.

18. R. Aronson, "Metamap: Mapping text to the UMLS metathesaurus," pp. 1–26, 2006 [Online]. Available: http://skr.nlm.nih.gov/papers/references/metamap06.pdf

19. R. Aronson, "Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program ," in *Proc. AMIA Symp. Amer. Medical. Informat. Assoc.*, 2001.

20. Y. Ling, Y. An, And X. Hu, "A Matching Framework For Modelling Symptom And Medication Relationships From Clinical Notes," *In Proc. IEEE Int. Conf. IEEE Bioinformat. Biomed.(Bibm), 2014.*