

Enhanced Multi-View Point Non-Negative Matrix Factorization Clustering for Clinical Documents Analysis

DIVYA SHARMA, V. VIJAYARAJAN*, MOHAN KUBENDIRAN and R. PADMAPRIYA

SCOPE (School of Computing Science and Engineering), VIT University, Vellore, India.

*Corresponding author E-mail: vijayarajan.v@vit.ac.in

<http://dx.doi.org/10.13005/bpj/1338>

(Received: November 11, 2017; accepted: December 18, 2017)

ABSTRACT

Clustering of clinical documents is the major research area in the field of machine learning and artificial intelligence which aims to acquaint some type of association with the information that helps to highlight huge examples and patterns. The rich corpus of clinical notes consists of several unprocessed data which needs to be mined with appropriate technique to improvise the existing healthcare system. Biomedical information mining is a research strategy to recover, break down and analyze clinical information from a collection of medicinal records. This paper presents a novel approach that utilizes Non-Negative Matrix Factorization Clustering approach to mine the medication names based on age of the patients. Pharmaceutical data from clinical notes is regularly communicated with prescription names and other medication information which needs to be mined based on the similarity between documents so that more accurate extraction of similarity could be accomplished. Even in the wake of being an exceptionally effective solution, clustering is yet not deployed in the major search engines. The basic issue with it is to determine a fast and accurate cluster values even after reducing the complexity of the technique. This paper presents an enhanced multi-viewpoint similarity measure that utilizes many distinct viewpoints to measure similarity between documents so that more accurate extraction of similarity could be accomplished.

Keywords: Clinical Documents, Clustering Algorithms, Cluster Analysis, Heuristic, Text Mining; Semantics.


INTRODUCTION

Information mining is known to be a notable wellspring of knowledge retrieval techniques from database. It is considered as a counterfeit strategy that allows us to find helpful information dwelling in important parts of data sources. It has been demonstrated to have good potential to separate valuable data from a comprehensive gathering of information. Clustering has been one of the most proficient approaches that can provide the researchers a way to extract information from the

grouped clinical notes. Various researches have been made in this domain that have doubtlessly accomplished incredible pace in the locale of restorative exploration and clinical practice. Clinical data mining is the way toward applying the information mining strategies on the acquired literary clinical reports.

Rich content information of clinical reports contains data about drugs and syndromes. Extricating this data has proven useful to refine the medical system. Many studies have proposed



This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted Non Commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

various proficient techniques for diagnosing diseases to extricate the right medical information from the unprocessed data. Clinical documents are broadly utilized for future investigation and determination of the sickness. The clinical notes have an incredible use in drug store so to lessen imitation and avoid drug misuse. Record clustering helps gather them into pertinent groups. This is done basically to find critical patterns to put a collection of the comparable articles into groups. By utilizing this method, we can perform grouping on literary information corpus that is in unstructured and semi organized configuration. These sorts of important data separated from biomedical prescriptions are very useful to build a consolidated summary for patients^{1,2} Pharmaceutical data is an integration of prescription names and other medication information, for example drug name, dosage, course, etc. A significant amount of work has been done in the field of clinical research for clustering of pharmaceutical records for mining useful text. In 2008 Doing- Zhang et al. worked on ontology-based learning and proposed nine similarity measures of words in clinical document clustering³. The ontology based work discussed in⁴ can help to build an ontology based clinical retrieval system. In 2011 Patterson et al. showed their research by clustering a number of 17 biomedical notes with the help of an unsupervised clustering algorithm using various lexical and semantic patterns for proper extraction⁵. The work in⁶ ontology based access control mechanism is used to secure clinical data.

In 2010 Han et al. clustered clinical notes using latent semantic indexing method and found it as an effective method for measuring the similarity⁷. Harris et al. compared the linguistic features of clinical documents of various institutions according to their respective medical specialty by using clustering methods. Nonnegative matrix factorization (NMF) framework significantly outperforms the various classical clustering algorithms such as that of bisecting K-means, and hierarchical clustering⁵. Our contribution to the research paper is summarized as: 1) An enhanced framework system for extricating symptom/medication names from clinical documents; 2) Application of multi-view NMF to effectively use medication/symptom names to improvise the clustering results; 3) Extraction based on age 4) Computing accuracy on the basis of word count and tf-idf factor.

Related Work

Over the past few years, researchers have carried out various works relating to clustering of biomedical notes this section of document presents some of them. Xiaodi Huang et.al⁸ proposed the ensemble non-negative matrix factorization technique for effective clustering of clinical notes. A combination of ensemble non-negative matrix factorization and Hybrid Bipartite Graph formulation (HBGF) is proposed for clustering the prescriptions. Yuan Linget. al⁹ Build a combined approach for extricating drug names and their relating symptom names from a collection of clinical documents. Their work showed a comparison of two approaches namely, non-negative matrix factorization technique and multi-view NMF to produce clusters out of a set of pharmaceutical prescriptions based on different sample-feature matrices generated. DucThang Nguyen et. al¹⁰ Introduced an efficient multi viewpoint-based similarity measure for clustering clinical notes. This concept was applied on clustering which the authors named as MVSC. Subsequently, they represented it as $MVSC-I_x$ and $MVSC-I_y$ i.e. MVSC as a criterion function and respectively. The main goal is to perform document clustering by optimizing and.

Hung Chim at. al¹² Presented their work using an expression (phrase) based record clustering. They focused their work to compute the similar documents by the usage of a method called as Suffix Tree Documents model. The authors considered three kinds of suffix namely the root node, leaf nodes, and internal nodes for every document. STC Algorithm was then applied to obtain better clustering results than the conventional algorithms. Tsang-Hsiang Cheng et. al¹³ A technique named as clustering-based Category-Hierarchy Integration (CHI) was proposed which is an improvement in the technique of clustering-based Category Integration Approach abbreviate as CCI. Their execution of category-hierarchy integration showed improvement in the results as compared to the accuracy of that obtained by non-hierarchical category-integration techniques. William Hsu et. al¹⁴ Stated that knowledge from the clinical sources can be utilized to mine the useful data and also analyze clinical data present in the dataset improving the accessibility to various portions of the record. They extricated the features from the biomedical records

which were mapped to the concepts available in the text data thereby computing results based on concepts mentioned in the knowledge bases. Shady Shehata *et. al*¹⁵ Presented a model that computed the similarity by calculating the similarity between the sentences present in the documents by analyzing their meaning. Concept-based Analysis was the proposed algorithm that used the concept of likeliness to measure the values of term level, sentence level (*ctf*), document (*tf*), and corpus levels (*df*) for a set of documents. Concept based similarity was applied on various datasets and the results proved that the proposed work outperformed the conventional analysis methods. Jiayue Zhang *et. al*¹⁶ They devised a novel approach to enhance the search results for electronic search records by calculating temporal similarity.

The authors proposed an algorithm to combine textual and temporal relevance with the help of adapted hierarchical clustering method for the purpose of re-ranking of healthcare records. This is used for re-ranking and re-positioning of records. Adil M. Bagirov *et. al*¹⁷. The authors further modified the existing modified K-means algorithm by calculating the cluster in a stepwise increment manner. For this they generated the starting points with the help of the auxiliary functions. The minimization of the function is achieved by applying k-means algorithm on it.

Taxiarchis Botsis *et. al*¹⁸ Aimed to apply various clustering algorithm to document networks on the respective values of threshold to obtain document clusters. The authors applied three clustering algorithms namely *k*-means, visualization of similarities and Louvain to the obtained networks and calculated the performance to determine cluster values. Arthur, D *et. al*¹⁹ They applied their technique to achieve a better running time for the k-means approach. The authors discovered that the running time of k-means clustering algorithm was limited by a polynomial of $1/\sqrt{n}$ and n . The authors concluded that they would like to evaluate the quality of the local optimum obtained using *k*-means approach and whether the values achieved could be considered as global optimum or not.

Atanaz Babashzadeh *et. al*²⁰. They utilized semantic data to enhance the performance results

of clinical IR framework by defining queries in a representative and significant manner. A dataset namely TREC was used for validation of their approach. Results demonstrated the devised approach incomparably improved the performance values of the retrieval of information as compared to conventional keyword-based IR model.

Nan Cao *et. al*²¹ Presented their technique named as FacetAtlas, which is a multifaceted visualization approach used for visually evaluate rich text datasets. In order to extract the local as well as global values the authors devised an integrated approach by the application of searching technique on advanced visual analytical device. Edward Omiecinski *et. al*²² The authors presented an efficient parallel approach for record clustering. This technique was run on a SIMD machine. The authors proved that there does not exist any difference between the SplitMerge algorithm and their performance results. Honigman *et al*²³ Their work focussed on the detection of the ADEs in the patients' record using a lexicon device. This device was applied on clinical documents to extract Adverse Drug Events (ADEs). Their approach determined various problems in outpatients in an efficient and economic manner. Hripacsak *et al*²⁴ Focused their work on extraction of information regarding patients with tuberculosis using an NLP method. Also there is an NLP based generic framework discussed in²⁵ can help to retrieve the clinical data efficiently. They devised a clinical policy to ensure immediate solution to the problem and to isolate these patients. When combined with automated protocols the Clinical protocols produced an extraordinary good results rather than using it alone.

Emilia Apostolova *et. al*²⁶ who determined a technique that automatically segmented medical documents into semantic sections. They used Hand-crafted rules to develop a scalable biomedical documents segmentation approach that required very less user effort for efficient extraction of information from clinical texts. They for automatically identifying high confidence training set. Renchu Guan *et. al*²⁷ An Affinity Propagation (AP) based approach namely approach named as Seeds Affinity Propagation (SAP) to improvise the semi-supervised clustering approach. A dataset namely

Reuters-21578 was selected for comparison of the results obtained with the two clustering algorithms, namely, AP algorithm and k-means algorithm.

words etc which needs to be removed to a structured format of text. An example of free text format of clinical document is shown in figure 1.

METHODOLOGY

Clinical Documents

Clinical documents consist of medical data about patients such as demographic, symptoms, medicines, billing, gender, age or other historical information. The information present is in the unstructured text form. They consist of tags, stem

Pre- processing

The clinical prescriptions consist of the structured sections which consist of unorganized pharmaceutical information. Enormous amount of hidden medical information can be mined from them by applying right technique. In this research an efficacious approach is employed to pre-process documents in order to extract the valuable words and

CHIEF COMPLAIN: Chest pain
 She is a very pleasant 31-year-old mother of two children with ADD. she has a **partial hysterectomy** from January of 2009. In this setting, she has been having multiple cardiovascular complaints including **chest pains**, which feel **"like cramps"**; and sometimes like a **dull ache**, which will last all day long. She is also tender in the left breast area and gets **numbness** in her left hand. She has also had three spells of **"falling"**; she is not really clear on whether these are syncopal, but they sound like they could be as she sees spots before her eyes.
MEDICATIONS: Naprosyn, which she takes up to six a day.
ALLERGIES: Sulfa.
MEDICAL HISTORY: She does not smoke or drink. She has no remote history of syncope. She suffered no trauma.
BP: 130/70 without orthostatic changes.
PR: 72. **WT:** 206 pounds. She is a healthy young woman. No JVD. No carotid bruit. No thyromegaly.
Cardiac: Regular rate. There is no significant murmur, gallop, or rub.
Chest: Mildly tender in the upper pectoral areas bilaterally (breast exam was not performed).
Lungs: Clear.
Abdomen: Soft. Moderately **overweight**.
Extremities: No edema and good distal pulses.
EKG: Normal sinus rhythm, normal EKG
ECHOCARDIOGRAM (FOR SYNCOPES): Essentially normal study.
IMPRESSION:
 1. Syncopal spells **–** These do sound, in fact, to be syncopal. I suspect it is simple orthostasis/vasovagal, as her EKG and echocardiogram looks good. I have asked her to drink plenty of fluids and to not to get up suddenly at night. I think this should take care of the problem.
 2. Chest pains **–** The **Naprosyn** is not helping that much, I gave her a prescription for **Flexeril** and instructed her in its use (not to drive after taking it).
RECOMMENDATIONS:
 1. Reassurance that her cardiac checkup looks excellent, which it does.
 2. Drink plenty of fluids and arise slowly from bed.
 3. **Flexeril** 10 mg q 6 p.r.n. I have asked her to return should the syncopal spells continue.

Fig. 1: An example of textual clinical prescription

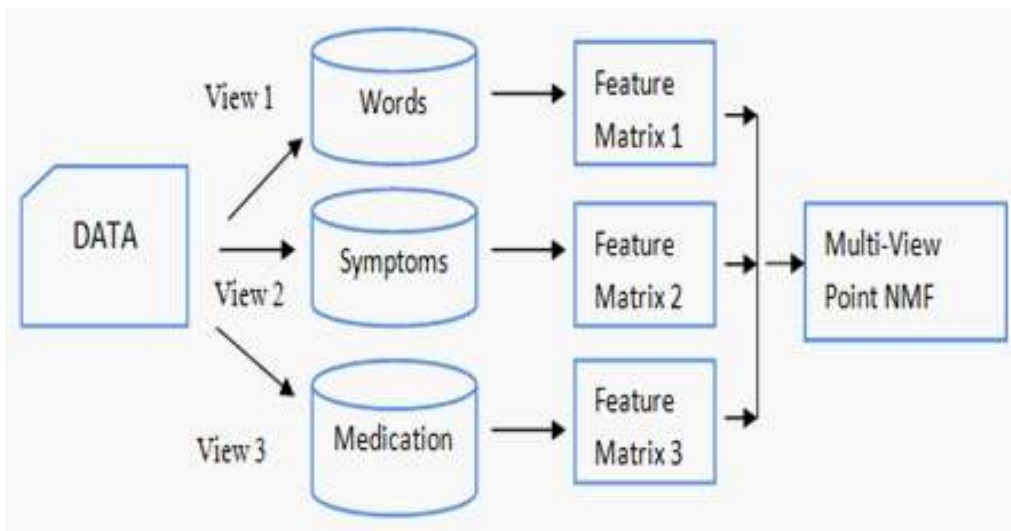


Fig. 2: An Example of the Pre-Processing Framework

proper sections from the clinical notes. This results into improvement of the quality of prescriptions present in the dataset. An overview of the pre-processing approach for the extraction of symptoms and medication is described in figure 1.

Pre-processing of the clinical prescriptions includes the removal of unnecessary words such as stem or stop words, eliminating unnecessary sections from the clinical notes. For this the clinical notes were first pre-processed the text document using the Standard CoreNLP Tool (<http://stanfordnlp.github.io/CoreNLP/>) which is a java based tool that helps to parse, identify the entities, sentimentally analyse the text.

A number of sections are present in the biomedical documents, symptoms and medication names needs to be extracted from these sections. Majorly the symptom names are contained in sections such as Chief Complain, History of illness, Assessment plan, Physical Exam, Specimen, Review of systems. Likewise, the drug names can be found in the following sections Medication,

Impression, Recommendations, Past medical history, Assessment plan, medication on discharge. For such computations section annotator is used in order to differentiate between the sections present in the textual prescription. The header information for the respective sections is computed and based on that the necessary sections are retrieved from the document. The sections which provide the negation recommendations are excluded using the negation annotator. This is done using the NegEx (<https://healthinformatics.wikispaces.com/NegEx+Algorithm>), a Natural Language Processing (NLP) tool which detects the negative terms present in the clinical text and removes the corresponding medicine name so that right medication is retrieved. This tool identifies the trigger terms and works according to the scope of the terms identified. This tool recognizes the pre-negation and post-negation words such as keptoff, avoid, away, without and was ruled out, free respectively. The clinical note shown in the figure 1 consists of chief complaint of chest pain. The statements “The naprosyn is not helping that much” and “keptoff Protonix” have negation words “not helping much” and “keptoff” respectively so the negation medicine relating these medicines are removed. Likewise the negation terms such as ‘avoid’, ‘allergic’ are annotated in the document and the corresponding medication is discarded for correct medical recommendation.

Table 1: Size of the Sample Matrices from the Dataset

Attributes	#Size
Symptom names	216
Medicine names	156

After the pre-processing process is over the sample matrices is obtained which has the attributes as shown in table 1.

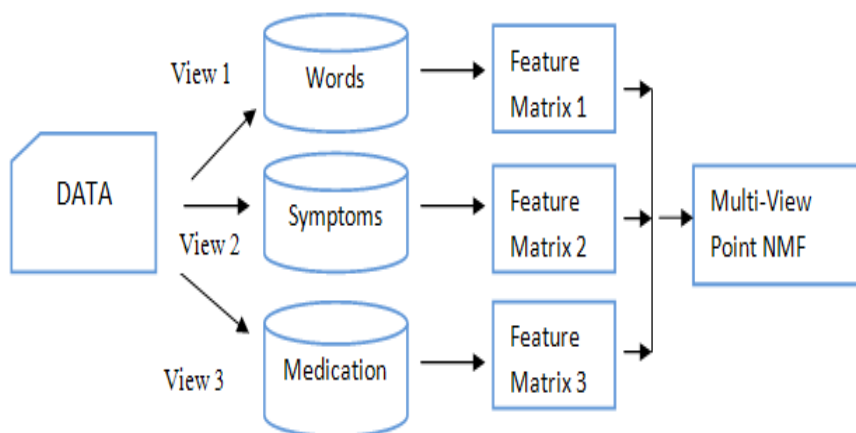


Fig. 3: An overview of Multi-View Point Non-Negative Matrix Factorization

Symptom/Medication Names Extraction

The medication and symptom names are present in the different sections of the clinical notes. After pre-processing of documents MedEx [28] is used for identifying and extricating the medicine names and MetaMap^{8,27} is used for symptom names extraction from these sections. Figure 2 shows the extraction of the same using MedEx and MetaMap after pre-processing the clinical notes. MedEx (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995636/figure/fig1/>)³⁰ is a java based open source tool that helps to identify the medicine terms such as dosage, intake time, drug names, duration and amount in the clinical text. It maps the medication information found to the UMLS thesaurus concepts to find the most precise match for accurate medication extraction. The symptoms extraction is done using MetaMap (<https://metamap.nlm.nih.gov/>) which is a configurable program that helps to relate the biomedical data to Meta-thesaurus concepts. It uses semantic knowledge representation NLP based approach for mapping the concepts such as “aapp” , “clna”, “clnd”, “nnon” which means amino

acid, clinical attribute, clinical drug and nucleic acid respectively. Some other are bact, bodm, enzy, impo, vita etc (<https://mmtx.nlm.nih.gov/MMTx/semanticTypes.shtml>).

Matrix Factorization Technique (M-NMF)

Non-Negative Matrix Factorization

NMF is an efficient method to factorize a given non-negative matrix says, $N \times M$ into the product of two non-negative matrices of lower dimension in the forms as matrix $N \times K$ and another matrix, $K \times M$ which can be efficiently expressed using the Euclidean distance formula 1:

$$\min \|A - WH\|^2 \dots(1)$$

The NMF aims at finding a low rank estimate of a matrix V (by considering V as a product of two-dimensional decreased matrices W and H. Each of the columns of W represents a basis vector and that of H contains an encoding of the linear composition of the basis vectors that approximates

Table 2: Extraction of symptom and medicine names

Symptoms	Medicines
1. Chest pain, wheezing, haemoptysis, quot, cramps, moderately overweigh, orthostasis	Flexeril; Albuterol; hcl
2. Abdominal pain, obstipation, sinus, abdominal scars, Vault prolapsed	Nasonex; Xopenex; Advair
3. Heart failure, fatigue, fever, coronary artery disease, vomiting	Methyl dopa; thera; Dipyridamole

Table 3: Dataset Results for the Age Group below 30

Type	Views	Accuracy
Count	Words	70.9090909090909
	Symptoms/ Medicine	80.7272727272727
	All views	83.7272727272727
TF-IDF	Words	73.4545454545454
	Symptoms/ Medicine	78.7272727272727
	All views	80.9272727272727

Table 4: Dataset Results for the Age Group Above 30

Type	Views	Accuracy
Count	Words	74.9090909090909
	Symptoms/ Medicine	72.7272727272727
	All views	75.7272727272727
TF-IDF	Words	71.4545454545454
	Symptoms/ Medicine	76.7272727272727
	All Views	78.9272727272727

the respective column of V . Measurements of W and H are and separately, where k is the decreased rank matrix⁹. Lee and Seung^{7,29} proposed an efficient technique to update rules of W and H multiplicatively which is known as multiplicative method (MM). The algorithm is stated below as⁹:

(1) Initialization of H and W with non-negative values.

(2) Iterate for each variable c , i , and j until convergence or after l iterations:

$$(a) H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T W H)_{\alpha\mu}} \quad \dots(2)$$

$$(b) W_{i\alpha} \leftarrow W_{i\alpha} \frac{(A H^T A)_{i\alpha}}{(W H H^T)_{i\alpha}} \quad \dots(3)$$

The optimization of $A^i \approx W^i H^i$ for each view l can be achieved by the equation:

$$\min \sum_{i=1}^p \|A^i - W^i H^i\|^2 + \sum_{i=1}^p \|H^i - H^*\|^2 \quad \dots(4)$$

Multi-View Point NMF

NMF has been effectively utilized in multi-view learning. Multi-view technique helps to identify latent components in various distinct sub-matrices in a simultaneous manner. Z. Akata *et al.* [9] proposed the extension of the basic NMF in an optimized form using different views.

$$\min_{W^i H \geq 0} \sum_{i=1}^p \lambda_i \|A^i - W^i H\|^2$$

$$\sum_{i=1}^p \lambda_i = 1, \lambda_i \geq 0 \quad \dots(5)$$

Figure 3 shows the M-NMF technique that utilizes three views i.e. words, symptoms and medication to compute the feature matrix. The matrix is used for the computation of multi-view point similarity computation. As more than one view are utilized the accuracy of the computations made

becomes high. Also, the extraction is based on the age of the patients', so the accuracy value is much higher.

RESULTS AND DISCUSSION

Dataset Result

The result for the medicines and symptoms extraction is shown in table 2. The idea is to extricate the medicine and symptoms based on the age of the patients' using the proficient clustering technique i.e. Multi-View Point Non-negative matrix factorization. The value of k (cluster value) is taken as three ($k=3$) for clustering the documents into three clusters.

Evaluation Metrics

Accuracy based on three views is computed in this research work. It is the measurement of the fraction of documents that are labelled correctly. A one-to-one correspondence exists between the true classes and the assigned clusters. Accuracy is calculated using formula 6.

$$Accuracy = \frac{1}{n} \max_q \sum_{i=1}^k n_{i,q(i)} \quad \dots(6)$$

Where q is the possible permutation from 1 to k .

Two age groups are considered for the extraction of symptoms and medication names. Table 4 shows the results for the accuracy based on patients' age for the age group less than 30. As shown in Table 3 the count based on words in the clinical notes is calculated using the feature matrix thereby calculating the accuracy using the formula specified in equation 6. The accuracy based on count and TF-IDF factor as per the respective view is computed.

Table 4 shows the results for the accuracy based on patients' age for the age group above 30. As shown in Table 3 the count based on words in the clinical notes is calculated using the feature matrix thereby calculating the accuracy using the formula specified in equation 6. The accuracy based on count and TF-IDF factor as per the respective view is computed.

CONCLUSION

In this research work a medical framework is implemented for the extraction of symptom and medicine names from the free-text documents. The system uses pre-processing units i.e. negation annotator, section annotator, word annotator. From the extraction of medicines and symptoms three views from the clinical notes. Multi-View NMF is applied for clustering the clinical notes. The accuracy

with respect to three views is computed for extraction of the medicines and symptoms based on the age of the patients' mentioned in the clinical notes. The age based result obtained for both the techniques i.e. NMF and M-NMF showed improved accuracy for M-NMF as compared to NMF technique. It indicates that the M-NMF technique is an improvised version of the NMF and has the capability to perform faster to obtain better results.

REFERENCES

1. F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. S  by, S. Bredkj  r, A. Juul, T. Werge, and others, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Comput Biol*, **7**(8): p. e1002141 (2011).
2. G. Hripcsak, S. Bakken, P. D. Stetson, and V. L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery," *J. Biomed. Inform.*, **36**(1): pp. 120–130 (2003).
3. X. Zhang, L. Jing, X. Hu, M. Ng, J. Xia, and X. Zhou, "Medical document clustering using ontology-based term similarity measures," (2008).
4. V. Vijayarajan, M. Dinakaran, and M. Lohani, "Ontology based object-attribute-value information extraction from web pages in search engine result retrieval," *Smart Innov. Syst. Technol.*, **27**(1): pp. 611–620 (2014).
5. O. Patterson and J. F. Hurdle, "Document clustering of clinical narratives: a systematic study of clinical sublanguages," in *AMIA Annu Symp Proc*, pp. 1099–1107 (2011).
6. K. Mohan and M. Aramudhan, "Ontology based access control model for healthcare system in cloud computing," *Indian J. Sci. Technol.*, **8**(S9): pp. 218–222 (2015).
7. C. Han and J. Choi, "Effect of latent semantic indexing for clustering clinical documents," in *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, 2010, pp. 561–566.
8. X. Huang, X. Zheng, W. Yuan, F. Wang, and S. Zhu, "Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization," *Inf. Sci. (Ny)*, **181**(11): pp. 2293–2302 (2011).
9. Y. Ling, X. Pan, G. Li, and X. Hu, "Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization," *IEEE Trans. Nanobioscience*, **14**(5): pp. 500–504 (2015).
10. A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD NLM, NIH, DHHS*, pp. 1–26 (2006).
11. Y. Yan, L. Chen, and D. T. Nguyen, "Semi-supervised clustering with multi-viewpoint based similarity measure," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, 2012, pp. 1–8.
12. H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Trans. Knowl. Data Eng.*, **20**(9): pp. 1217–1229 (2008).
13. T.-H. Cheng and C.-P. Wei, "A clustering-based approach for integrating document-category hierarchies," *IEEE Trans. Syst. Man, Cybern. A Syst. Humans*, **38**(2), pp. 410–424 (2008).
14. W. Hsu, R. K. Taira, S. El-Saden, H. Kangarloo, and A. A. T. Bui, "Context-based electronic health record: toward patient specific healthcare," *IEEE Trans. Inf. Technol. Biomed.*, **16**(2): pp. 228–234 (2012).
15. S. Shehata, F. Karray, and M. Kamel, "An efficient concept-based mining model for enhancing text clustering," *IEEE Trans. Knowl. Data Eng.*, **22**(10): pp. 1360–1371 (2010).
16. J. Zhang, J. X. Huang, J. Guo, and W. Xu, "Promoting electronic health record search through a time-aware approach," in *Bioinformatics and Biomedicine (BIBM), 2013*

- IEEE International Conference on*, 2013, pp. 593–596.
17. A. M. Bagirov, J. Ugon, and D. Webb, "Fast modified global k-means algorithm for incremental cluster construction," *Pattern Recognit.*, **44**(4): pp. 866–876 (2011).
 18. T. Botsis, J. Scott, E. J. Woo, and R. Ball, "Identifying Similar Cases in Document Networks Using Cross-Reference Structures," *IEEE J. Biomed. Heal. Informatics*, **19**(6): pp. 1906–1917 (2015).
 19. D. Arthur, B. Manthey, and H. Röglin, "Smoothed analysis of the k-means method," *J. ACM*, **58**(5): p. 19 (2011).
 20. A. Babashzadeh, J. Huang, and M. Daoud, "Exploiting semantics for improving clinical information retrieval," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 801–804.
 21. N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted visualization for rich text corpora," *IEEE Trans. Vis. Comput. Graph.*, **16**(6): pp. 1172–1181 (2010).
 22. E. Omiecinski and P. Scheuermann, "A parallel algorithm for record clustering," *ACM Trans. Database Syst.*, **15**(4): pp. 599–624 (1990).
 23. B. Honigman, P. Light, R. M. Pulling, and D. W. Bates, "A computerized method for identifying incidents associated with adverse drug events in outpatients," *Int. J. Med. Inform.*, **61**(1): pp. 21–32 (2001).
 24. C. A. Knirsch, N. L. Jain, A. Pablos-Mendez, C. Friedman, and G. Hripcsak, "Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system," *Infect. Control Hosp. Epidemiol.*, pp. 94–100 (1998).
 25. V. Vijayarajan, M. Dinakaran, P. Tejaswin, and M. Lohani, "A generic framework for ontology based information retrieval and image retrieval in web data," *Human-centric Comput. Inf. Sci.*, **6**(1); (2016).
 26. E. Apostolova, D. S. Channin, D. Demner-Fushman, J. Furst, S. Lytinen, and D. Raicu, "Automatic segmentation of clinical texts," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 5905–5908.
 27. R. Guan, X. Shi, M. Marchese, C. Yang, and Y. Liang, "Text clustering with seeds affinity propagation," *IEEE Trans. Knowl. Data Eng.*, **23**(4): pp. 627–637 (2011).
 28. H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives," *J. Am. Med. Informatics Assoc.*, **17**(1): pp. 19–24 (2010).
 29. M. Doyle, "The Metamap Process: A New Approach To The Creation Of Object Oriented Image Databases For Medical Education," in *Engineering in Medicine and Biology Society, 1991. Vol. 13: 1991., Proceedings of the Annual International Conference of the IEEE*, 1991, pp. 1046–1047.
 30. S. Sohn, C. Clark, S. R. Halgrim, S. P. Murphy, C. G. Chute, and H. Liu, "MedXN: an open source medication extraction and normalization tool for clinical text," *J. Am. Med. Informatics Assoc.*, **21**(5): pp. 858–865 (2014).
 31. J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013, pp. 252–260.
 30. Dataset :- (<https://idashdata.ucsd.edu/community/45>).