A Survey on Chemical Text Mining Techniques for Identifying Relationship Network between Drug Disease Genes and Molecules

Mita A. Landge
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune-44

K. Rajeswari, PhD
Professor, Department of Computer Engineering,
Pimpri Chinchwad Collage of Engineering
Pune-44

ABSTRACT
The Text mining plays essential roles in the field of Chemoinformatics to reveal unknown information. The enormous amount of biomedical information is available on internet and resides in the form of published articles, files, patents etc. As the rich source of data is growing massively, it is widely contributing to the scientific researchers. Text mining is the most widely used in field of Natural Language Processing. The Text Pre-processing and data analysis techniques applied on biomedical literature allows us to identify ad investigate new theories. Finding the association between the chemical entities like drug, disease, genes and molecules is the new area of focus for researchers. This paper presents the study on several approaches and techniques proposed for chemical text-mining to identify relationship network for drug-disease, disease-gene associations.

In this paper, we focus on comparative analysis of various Text mining techniques used for chemical literature with their results evaluations as well as observations.

Keywords
Chemical text mining, data analysis, text mining techniques, Natural Language Processing (NLP)

1. INTRODUCTION
The increasing amount of scientific publications, articles, patents and journal has largely contributed to the era of new discoveries using text mining techniques. Text Mining allows us to identify the previously unknown information, find interesting patterns from large data sources and repositories and helps in proposing new theories in the field of bioinformatics. Text mining also known as Intelligent Text Analysis [1] is the process of extracting non-trivial and interesting information and knowledge from unstructured text. The chemical text mining allows us to extract information of various chemical molecules, genes and drugs for available for several disease from the rapidly growing scientific literature [2]. The range of biomedical applications has been proposed till now to detect biomedical entities from the huge literature and to identify relationship network between those entities. Various Chemical text mining tools are available for processing the chemical literature. These tools work for identifying the chemical structural elements and behavior of those chemical elements to new molecules [3]. Text mining comprises of data mining, information retrieval, natural language processing and machine learning methods [4]. Text mining algorithms are available finding the relevant document from the huge data source, calculating the term document matrix for the given data. The term frequency helps in analyzing the different patterns in the data and its occurrence [5]. The relevance of a term in a document with the other terms helps in generating certain rules for association between those terms. Name entity recognition (NER) and natural language processing (NLP), machine learning are the text analysis tools [6]. The text mining work flow consist of text preprocessing , text categorization, text classification, text clustering, concept or entity extraction, document summarization, sentiment analysis, and entity relation modeling [7]. The data analysis is the part of text mining process which allows for retrieving different patterns from the given data. The algorithms like classification, clustering, pattern matching, rule mining are widely used to get unknown and new knowledge. The standard chemical data sources available are PubMed, PubChem, Chemspider, and SciFinder [8]. Chemical literature databases associate structures or other chemical information with the relevant documents.

Chemoinformatics is the big area for solving computational and scientific problems in the field of chemistry. Performing text mining operations on this huge literature is time consuming and complex in the real life scenario. Various algorithmic strategies are focusing towards parallel approach for text mining. The GPU computation for parallel algorithms is the future of high performance computing for mining the huge data source. This paper focuses on various text mining techniques and tools used for relationship identification between chemical entities and outline the comparative algorithms used for text mining.

2. LITERATURE SURVEY
In the era of big data analytic the increasing amount of literature data is providing an open platform for several researchers to reveal new theories and knowledge. There are several text mining techniques for knowledge discovery proposed till present. The study of different approaches for identifying relationship network using text mining was identified.

1. In the work, ***EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature, 2000***[9], presents a natural language processing approach for extracting drugs and genes relationship for cancer disease. The automatic extraction and and document clustering techniques were used. The text preprocessing included POStagging, parsing of data using syntactic parser the data extraction from MedLine was done to retrieve abstracts for biomedical literature terms. The drugs and genes association for Cancer used two ways

1. The Impact of gene expression on the drug sensitivity of a cell,

2. Result after drug treatment resulting the changes in the cell's gene.

It uses existing syntactic NLP tools in combination with new semantic and pragmatic analyses. It is beneficial for identify well-characterized genes, drugs and cell lines can be immediately useful to biologists. The Hierarchical Clustering approach is applied for document clustering to retrieve better results. The case study was solved for genes and drugs involved in cancer therapeutics. The rule based system is used along with syntactic parsing of each sentence.

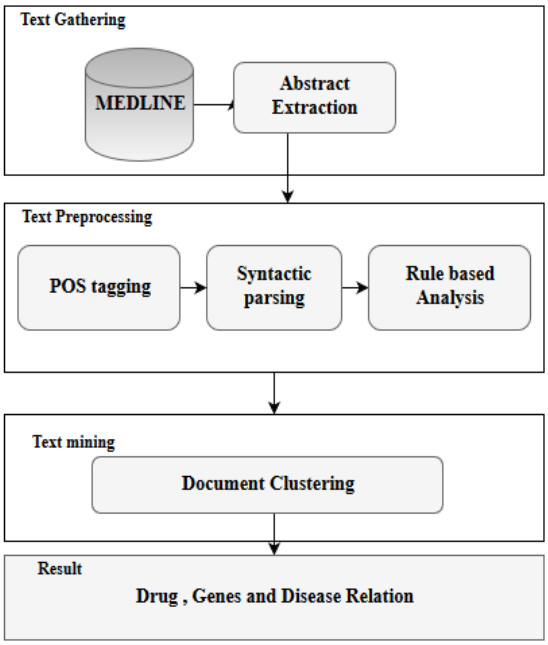

Fig 1. Workflow of Text mining using Hierarchical Clustering

2. In the work, ***Classifier Approach for Disease-Drug Relationships Identification for Tardive Dyskinesia (2012) [10],***

The identification and automatic extraction [11] of disease-drug relationships is done using sentence classification and annotations for tardive dyskinesia. The Naïve Bayes classification algorithm [12] is used to identify sentences pertaining to disease-drug relationships using 10-fold cross validation. This method helps to identify drug effects on tardive dyskinesia and extracts the disease-drug relationship. Weka tool for building Naïve Bayes classifier was used. The Naïve Bayes is simple, efficient and easy to implement [12, 13]. To evaluate the performance of the predictive modeling the Cross validation is used.

The above Fig1 shows the workflow pipeline of document retrieval and sentence classification using Naïve classifier. The steps included in this technique are:

1. Text Gathering from PubMed
2. Drug ontology and splitting using GENIA sentence splitter.
3. Training corpus on tardive dyskinesia with prior annotation
4. Naïve Bayes classification.
5. Disease – Drug relationship identification using software tools.

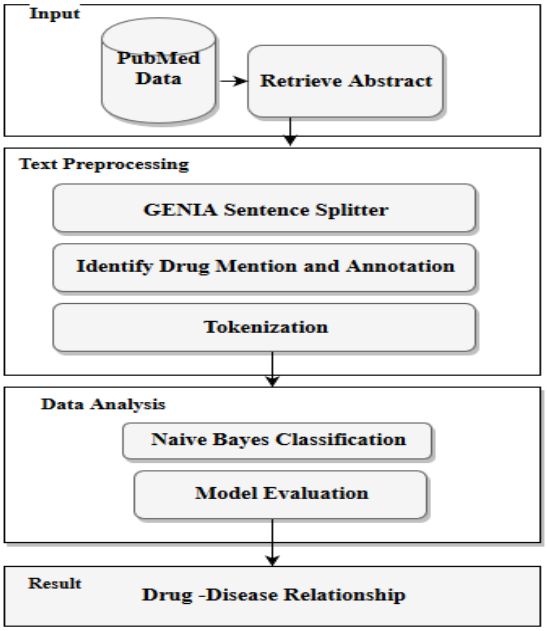

Fig 2. Workflow of Text mining using Text classification using Naïve Bayes

3. In the work, A Study and Analysis of Gene Drug Association for Diabetic Gene – A Text Mining Approach, 2014[14],

Drug and gene association is analyzed using appriori algorithm. The Dictionary based term identification approach is used for text mining and text preprocessing. In the work they analyzed the alternate chemical compositions similar to Terachlorodibenzodioxin drug .The chemical synonyms were verified biologically using MESH and PharmagKB. The Dictionary based approach is used to identify the drug names and gene name present in the extracted abstract terms. The literature study for various tools is also presented.

In the Fig 2, the system work flow divided into three phases is presented.

Phase 1: The first phase consist of creating of Dictionary based on corpus using MESH, MedLine.

Phase 2: Text gathering by extracting abstracts data from PubMed and corpus. The Text preprocessing techniques like tokenization, stopword removal, stemming, POStagging, and document term matrix creation is done.

Phase 3: Data Analysis using Association rule mining on Disease, genes and chemical.

The Association Rule mining algorithm allows association between the terms which occurs frequently. The term frequency generates the set of candidates for every iteration based upon the minimum support and confidence. This method helps to identify the association between Disease, Genes, and Chemical entities. To extract different patterns resulting in Gene - Disease, ad disease and Chemical association is done using Association Rule mining algorithm. The Inference rules satisfying the minimum support and confidence forms the rules for association between the terms. The accuracy of results is based on the total occurrence frequency. This work was experimented for Diabetic Gene.

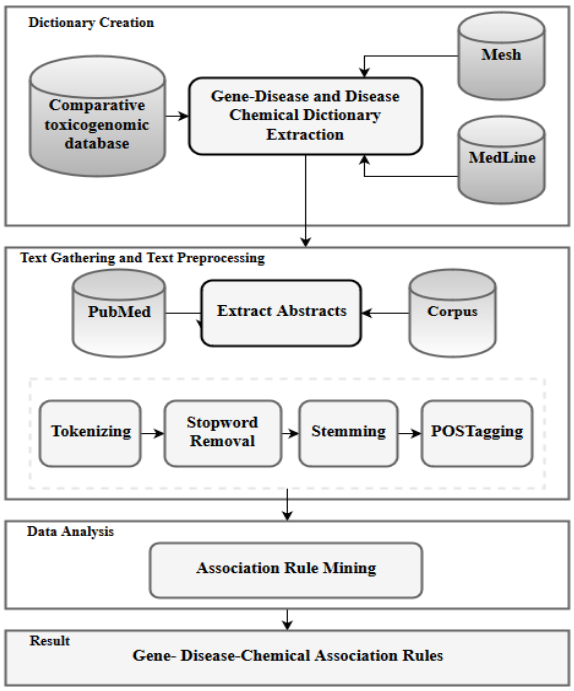

Fig 3. Workflow of Text mining using Association Rule mining

4. In the work, ***Learning the Structure of Biomedical Relationships from Unstructured Text.2015 [15],***

The Extraction of relationships from natural language sentences is performed. The text mining algorithms like Ensemble Biclustering for Classification (EBC) was combined with Hierarchical Clustering approach to demonstrate the better results for drug ad genes relationship.

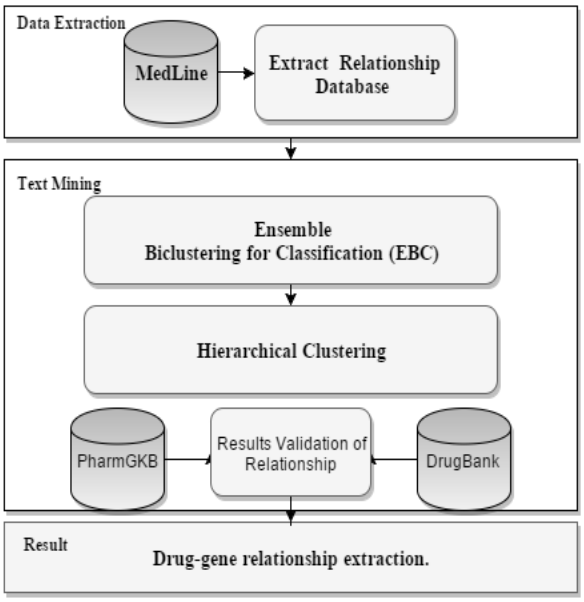

Fig 4. Workflow of Text mining Using Ensemble Biclustering and Hierarchical Clustering

The performance of this method was validated against PharmGKB and Drug Bank drug-gene relationship and drug target relationship [15].

The Preprocessing task for relationship extraction is:

1. Identification of drug-gene pairs co-occurrence n for all the sentences within a corpus of text.
2. Find total number of observed paths m by extracting all dependency paths connecting these drug-gene pairs in the corpus.
3. Form n x m matrix representing rows as drug-gene pairs and the columns as dependency paths.
4. EBC algorithm:
    a. Unsupervised step: Use Information-Theoretic Co-Clustering to bicluster the n x m matrix N times, recording the number of runs in which each row appears in a row cluster with each other row. The result is an n x n array, C, of co-occurrence values. Note that no information about the seed set is incorporated at this stage, so the unsupervised step need be run only once per data matrix.
    b. Supervised step: Identify a seed set, S, of rows that share some property of interest. Rank the entity pairs in a test set, T, based on a scoring function related to how often they co-cluster with members of S. Repeat this step as desired with different seed sets.
5. In work, Chemical named entities recognition: a review on approaches and applications 2014[16],

The extraction of molecules and their properties and activities is some by extracting the scientific literature data. The challenge of identifying and recognizing chemical entities from the instructed data is solved using Named Entity Recognition.

They proposed the dictionary-based, rule-based and machine learning, as well as hybrid chemical named entity recognition approaches with their applied solutions.

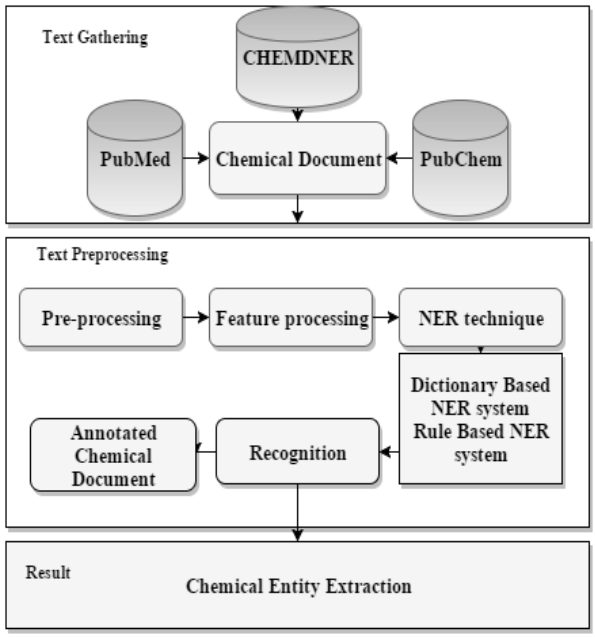

Fig 5. Workflow of Text mining using Named Entity Recognition

In fig 5, the NER step shows the flow of text mining approach to identify chemical entities in document.

3. OBSERVATIONS

For natural language sentences extracting these relationships from on such a large scale literature is complicated, however the appropriate text mining algorithms can be used to find the relationship information between chemical entities. The classification and association machine learning algorithms with supervised approach are better suitable such application with huge data size. The clustering provides better result after identification of relevant document cluster for accurate identification of drug and gene relationship. The Named Entity Recognition (NER) is also widely used method for recognizing and naming the entity within the unstructured data. Identifying the terms in the extracted data and then classifying it to the appropriate class is the task of NER. The high precision value and the recall and F-measure determine the accuracy of the results. The more the accuracy the better the results of extraction. For classification operation, the Naïve Bayes classification algorithm, SVM, Decision tree, KNN classifiers can be used. For the large data mining Naïve Bayes is widely used. The Association rule mining algorithm is also known for its machine learning capability of generating rules from the training dataset. The rules satisfying the minimum support and maximum confidence are stated as the inferred rules. The accuracy of this algorithm is based upon the support and confidence measure of those rules along with the f-measure. The clustering approach is widely used for clustering of document to retrieve faster results in text mining. Various clustering algorithms like Hierarchical clustering, EBC clustering, etc. can be used in text mining.

4. RESULTS AND INTERPRETATIONS

For Naïve Bayes classification on chemical data, the classification for sentences was done based upon three categories: i.e. Positive, Negative, and Neither category. The sentences that demonstrated benefit of a drug in relation to a disease were assigned to the Positive category, i.e. drug is used to treat the disease. Sentences that involve negative effects between a drug and a disease were assigned to the Negative category, i.e. the drug induces the disease or is associated with progression of the disease. Sentences that belong to neither the positive nor the negative effect category were assigned to the Neither category. This occurs when the drug has no relation to the biological disease or when the sentence is inconclusive or exploratory in nature [9].

Table 1. The Term probability for a given the class [11]

Term	Positive	Negative	Neither
therapeutic	0.0022	3.23E-04	9.16E-04
improvemer	0.0055	6.46E-04	4.58E-04
vacuous	5.49E-04	0.0036	2.29E-04
neurotoxic	2.75E-04	0.0013	2.29E-04

Results from the 10-fold cross-validation were measured in terms of precision, recall, F-measure, and area under the ROC curve, as shown in Table I. Precision, recall, and F-measure were calculated according to conventional definitions.

Table 2.Accuracy by classification [11]

Class	Precision	Recall	F-measure	ROC area
1	0.645	0.686	0.665	0.827
2	0.627	0.615	0.621	0.837
3	0.691	0.667	0.679	0.824
Weighted A	0.66	0.659	0.659	0.828

The experiment analysis found 50% occurrences of causing diabetic of Type 1 and Type2 with gene CAT and HNF1B. Tetrachlorodibenzodioxin drug for treating diabetic disease was identified. Thus the association between gene and disease and chemical is performed. The results were biologically verified to validate the results by comparing the chemical composition and the chemicals treated for diabetic with specific gene using pharmGKB and mesh data base.

The result of association with inference results of relational association between Gene and Disease and Disease and chemical is found using ABC model. The results of level1 and level2 were identified to retrieve associations between Gene (A) -> Chemical (C).

Table 3. Gene – Chemical-Disease Interaction [14]

Chemical	Disease	Gene
Tetrachlorodibenzodioxin	Diabetes Mellitus, Insulin-Dependent,21	TAGAP
Tetrachlorodibenzodioxin	Diabetes Mellitus, Insulin-Dependent,12	CTLA4
Tetrachlorodibenzodioxin	Wolfram Syndrome	WFS1

5. CONCLUSIONS

In this survey work, comparative analysis of various methods to find the association between chemical entities are studied along with the comparative analysis of various text mining techniques. The machine learning algorithms using supervised approach gives better results for text mining on chemical data. The identified relationship network is proved to be accurate based upon the comparisons with the scientific standard relational network of chemical entities. This study also reveals the new discovery for various drug and molecules associated with a particular disease.

6. FUTURE WORK

The text mining algorithms takes large amount of time for large dataset. The time can be minimized using the parallel approach of text mining. The conventional algorithms can be parallelized and applied for mining and extracting information and knowledge from huge dataset.

7. REFERENCES

[1] Vishal Gupta "A Survey of Text Mining Techniques and Applications" Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[2] Krallinger M, Rabal O, Leitner F, et al. "The CHEMDNER corpus of chemicals and drugs and its annotation principles." Journal of Cheminformatics. 2015;7(Suppl 1):S2. doi:10.1186/1758-2946-7-S1-S2.

[3] Shidha M.V, Dr.T.Mahalekshmi, "Chem Text Mining-An Outline", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 2, Feb 2014.

[4] Cano, C.; Monaghan, T.; Blanco, A.; Wall, D.P.; Peshkin, L.Collaborative text annotation resource for disease centered relation extraction from biomedical text. *J. Biomed. Inform.*, 2009, *42*, 966- 977.

[5] Na Wang, "An Improved TF-IDF Weights Function Based On Information Theory", IEEE 2010.

[6] Kao, A.; Poteet, S. R. "Natural Language Processing and Text Mining". Springer: 2007.

[7] Muthukumarasamy Karthikeyan, Yogesh Pandit," MegaMiner: A Tool for Lead Identification Through Text Mining Using Chemoinformatics Tools and Cloud Computing Environment", 2015 Bentham Science Publishers.

[8] Eltyeb, Safaa, and Naomie Salim. "Chemical Named Entities Recognition: A Review on Approaches and Applications." *Journal of Cheminformatics* 6.1 (2014): 17. *PMC*. Web. 9 Feb. 2016.

[9] Rindflesch, Thomas C. et al. "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2000): 517–528. Print.

[10] Xia Bi , Hongzhan Huang , Sherri Matis-Mitchell, "Building a Classifier for Identifying Sentences Pertaining to Disease-Drug Relationships in Tardive Dyskinesia", IEEE 2012.

[11] Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomics literature. Pharmacogenomics. 2010 Oct;11(10):1467-89.

[12] Eibe Frank, Remco R. Bouckaert. "Naive Bayes for text classification with unbalanced classes." Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases. 2006. Springer-Verlag Berlin, Heidelberg.

[13] Cheng BY, Carbonell JG, Klein-Seetharaman J. "Protein classification based on text document classification techniques". Proteins. 2005 Mar 1;58(4):955-70.

[14] Kirthika.B, Dr.V.Bhuvaneswari, "A Study and Analysis of Gene Drug Association for Diabetic Gene – A Text Mining Approach"IEEE 2014

[15] Bethany Percha, Russ B. Altman, "Learning the Structure of Biomedical Relationships from Unstructured Text". PLOS Computational Biology,PCBI journel, 2015.

[16] Safaa Eltyeb, Naomie Salim "Chemical named entities recognition: a review on approaches and applications", Eltyeb and Salim Journal of Cheminformatics 2014.