Integrated Intelligent Research (IIR)

International Journal of Computing Algorithm
Volume: 05 Issue: 01 June 2016, Page No. 46- 49
ISSN: 2278-2397

Relation Extraction using Hybrid Approach and an Ensemble Algorithm

R.Radha[1], S.Vijaya[2]
[1]Associate Professor, Department of Computer Science, S.D.N.B.Vaishnav College for Women, Chromepet, Chennai.
[2]Research Scholar, Department of Computer Science, S.D.N.B.Vaishnav College for Women, Chromepet, Chennai.
radhasundar1993@gmail.com,vijeshnisna@gmail.com

Abstract: The amount of information stored in the form of digital has a gigantic growth in this electronic era. Extraction of heterogeneous entities from biomedical documents is being a big challenge as achieving high f-score is still a problem. Text Mining plays a vital role for extracting different types of entities and finding relationship between the entities from huge collection of biomedical documents. We have proposed hybrid approach which includes a Dictionary based approach and a Rule based approach to reduce the number of false negatives and the Random Forest Algorithm to improve f-score measure.

Keywords -- Biomedical documents, Heterogeneous entities, Named Entity Recognition, Relation Extraction, Text Mining.

I. INTRODUCTION

Wide ranges of textual information are hidden in digital forms in the data reservoirs. To extract required information and discover new knowledge from these sources Text Mining plays a vital role in this electronic era. Data Mining is used to discover knowledge from structured data where as Text Mining offers retrieval of text from the unstructured data collection, extracts relation between the entities and leads to discover new, hidden knowledge. In Biomedical field due to the increased amount of data collection, covering all the information by a researcher is tedious task. Text mining tool makes this tedious task ease with its major activities such as Information Retrieval, Information Extraction, Finding association, Categorization and summarization etc.,[Hearst….][1].

Main focus of this paper is identifying relationship between gene and disease because getting high performance for the same is still a problem. The combination of Data and Text mining is referred to as Duo-Mining[Creese.G,][2]. Many researchers recommended duo-mining for good decision making. The tasks involved in this work are i) Collecting abstracts from PubMed database ii) Recognizing entities from biomedical abstracts and Finding associations between heterogeneous entities. PubMed abstracts are retrieved from http://www.ncbi.nlm.nih.gov/pubmed , and parsed by removing stop words. Biomedical Entities identified by processing the parsed data with the proposed methods Dictionary based approach, rule based approach and the associations are classified using ensemble classifier Random Forest Algorithm. This paper is organized as follows, Section 2 discusses related work based on Information Extraction, Named Entity Recognition, Section 3 presents proposed work and describes proposed methods. Section 4 discusses the experimental results of the proposed algorithm and analyses the performance of the proposed algorithm. Finally Section 5 includes the conclusion of this paper.

II. RELATED WORK

Alex Bravo et al.[6] developed BeFree system which is used to identify relationships between heterogeneous biomedical entities with a special focus on genes and the diseases associated with them. The need for tools becomes necessary to identify, gather relevant information and placing it in the context of biomedical knowledge. In identifying complex relationships between entities of biomedical domain several challenges remain still. They have presented BeFree - a supervised text mining system which composed of biomedical NER(Named Entity Recognition) and a kernel based RE(Relation Extraction) module , to identify relationships between depression and the gene associated with depression, from free text. They have developed a semi-automatic annotation corpus. [J-D Kim et al][3] used supervised learning approaches which require the annotated corpora for the development and evaluation of Relation Extraction and their results shown good performance . Maryam Khordad[4] presented a system that provides an important as well as an up-to-date data for database construction and updation, by solving the problems such as Finding genotype names , finding phenotype names and phenotype-genotype extraction.

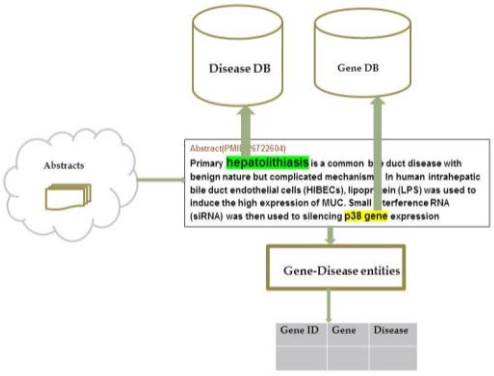
A.S.Foulkes[5] had used population-based genetic association studies to relate genetic sequence information derived from unrelated individuals to a measure of disease progression or disease status. Understanding the role of genetics in diseases is one of the major goals of the post-genome era. Changqin Quan et al[8] proposed Automatic gene-disease association approach based on text mining and network analysis. They combined information filtering , grammar parsing and network analysis for gene disease association extraction. They have taken Breast cancer as testing disease for system evaluation and checked with 31 top ranked genes and closeness centralities through NCBI database. The evaluation showed 83.9% accuracy for testing genes and diseases, 74.2% accuracy for testing genes. Hong-woo Chun et al.[9] described a system to extract disease-gene relation from Medline , constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching and used machine learning based Named Entity Recognition to filter out false recognitions of disease / gene names. Performance of Relation Extraction was dependent upon the performance of Named Entity Recognition filtering and improved the precision of Relation Extraction by 26.7%.Their future work encompassed increasing the size of the annotated corpus and enriching annotation. Krauthammer et al[10] used Dictionary based model and achieved 78.8% Precision and 71.1% Recall value. Yi-Feng Lin et al[11] used Maximum Entropy model based on features , combined with post processing and got the

International Journal of Computing Algorithm
Volume: 05 Issue: 01 June 2016, Page No. 46- 49
ISSN: 2278-2397

Precision value 72.1%, Recall value 71.1% and 78.5% f-score. D.Hanisch et al.[13] used Maximum Entropy method on protein data set and the precision, recall and f-score they achieved are 49.1%,62.1% and 54.8% respectively. Zhenfei Ju et al[12] developed a biomedical named entity recognition system with GENIA corpus data set and scored 84.24% Precision and 80.76% Recall. For efficient retrieval of data we have combined dictionary based approach and rule based approaches.To improve the f-score measure we have proposed an ensemble approach called Random Forest Algorithm in our work.

III. PROPOSED WORK

Alex Bravo et al.[6] developed BeFree system which discovered only a small proportion of the gene-disease associations from expert curated databases . Our proposed work aims to increase a large proportion of gene-disease associations with high performance. The block diagram of our proposed system is given below.



Block Diagram of Proposed work

3.1. Problem solving methodologies

The major task involved in the proposed work are as follows:
1.Information Extraction
2.Named Entity Recognition

The first task is concerned about the retrieval of abstracts from PubMed database, the second task dealt with the identification of Biomedical named entities and finding associations between entities.

3.1.1.Information Extraction

Wide range of information is available in the form of electronic documents. If the data is in tabular form it would be easier to answer queries straight forward. So, the first task to be done is converting unstructured data to structured data as extraction of structured data from electronically available biological literature has become important area of current research. In this task we extracted abstracts from PubMed database and pre-processed the abstract collections by segmenting words and removing stop words.

3.1.2. Named Entity Recognition(NER)

Named Entity Recognition refers to the ability of recognizing entities such as name of the person, location, date and time. In the field of biomedical it recognizes entities such as genes , diseases, protein and drug. It mines useful knowledge. The primary goal is identifying all textual mentions of the named entities. There are two major tasks involved in NER , they are i) Identifying the boundaries of the Named Entity ii) Identifying its type.

3.2. Proposed Method

3.2.1 Dictionary Based Approach

The first step in extracting information from the biomedical documents is Dictionary based entity name recognition. Dictionary based protein name recognition is used in extracting information from biomedical documents as it can provide ID information on recognized terms. To overcome false recognition mainly caused by short names Tsuruoka et al[14] have used machine learning to filter out false positives , to improve recall rate they have used appropriate string searching techniques and expanded the dictionary in advance with a probabilistic variant generator. As Dictionary based approaches have limitations such as false positive recognition and lack of a unified resource that covers newly published names they had addressed an approach with two-phase method where the first method scans text for protein name candidates and the second phase method filters irrelevant candidates by utilizing a Naïve Bayes Classifier. In our work we have proposed Rule based method to reduce the rate of false recognition caused by Dictionary based method and to increase the recall rate.

3.2.2.Rule Based Approach

Rule based approaches are used in Information Extraction to deal with a large database, broader range of variations and cover the word order variations. They utilize handmade rules and patterns. Fukuda et al [17]presented a method to identify core terms , feature terms and concatenates the terms by utilizing handcrafted rules, patterns and extend the boundaries to adjacent nouns and adjectives. Instead of Rule based approach Proux et al.[20] used a tagger with a nondeterministic finite automaton and scored 91.4% Precision, 94.4% recall and 92.9 % F-score.

3.2.3. Random Forest Algorithm

The Random Forest(RF) algorithm is an ensemble approach and popular machine learning algorithm with in statistical genetics. The main characteristics of RF are,i)Well adopted for both prediction and Variable Importance(VI), ii) Robust to the setting of tuning parameters. It has the capability of handling large amount of observations and predictors, and this method combines the “bagging” idea of Brieman and the random feature selection .[Benjamin A et al][16]. RF develops lots of decision tree based on random selection of data and random selection of variables. It provides the class of dependent variable based on many trees. As the trees are based on random selection of data as well as variables, these are called as random trees. Random Forest is an ensemble classifier that uses many decision tree models. Ensemble models combine the results from different models, the result from an ensemble model is usually better than the result from one of the individual models.

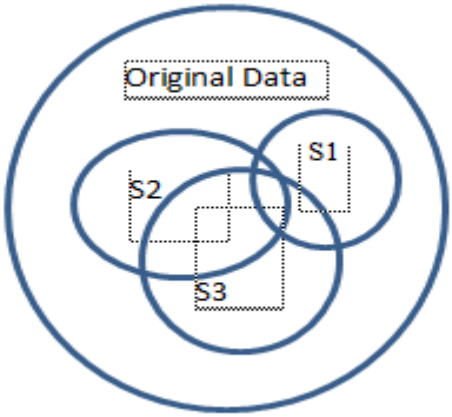
3.2.3.1: Steps involved in RF Algorithm

The following algorithm is used to construct each tree :

1. The number of training cases can be taken as N, and the number of variables in the classifier as M.

International Journal of Computing Algorithm
Volume: 05 Issue: 01 June 2016, Page No. 46- 49
ISSN: 2278-2397

2. 'm' should be much less than 'M' as the number of input variables(m) to be used to determine the decision at a node of the tree.
3. From all 'N' available training cases , choose a training set for this tree by choosing 'n' times. The rest of the cases can be used to estimate the error of the tree, by predicting their classes.
4. Randomly choose 'm' variables for each node of the tree on which to base the decision at that node. The best split can be calculated based on those 'm' variables in the training set.
5. 5.Each tree is not pruned but fully grown.
6. A new sample is pushed down the tree for prediction and is assigned the label of the training sample in the terminal node it ends up in. The average vote of all trees is reported as random forest prediction by iterating over all trees in the ensemble.



Random Selection of data

3.2.4 Implementation Steps

The steps involved in our work are as follows:

1.Dataset Collection.
2.Input Document.
3.Preprocessing the Document.
4.Comparing with genes and diseases.
5.Applying proposed algorithm.

3.2.4.1 Dataset Collection

We have used PubMed abtracts as dataset. The abstracts are retrieved from http://www.ncbi.nlm.nih.gov/pubmed and stored in a text file.

3.2.4.2 Input Document

The input document (text file in which retrieved abstracts are stored) is read using file reading commands.

3.2.4.3 Pre processing the Document

The file read is pre processed by segmenting and removing stop words.

3.2.4.4 Comparing genes and diseases

The terms (genes and diseases) are recognized by comparing each term with the respective genes file and disease file retrieved from genome database (www.genome.jp/kegg/genes.html) and Disease database (www.genome.jp/kegg/disease.html).

3.2.4.5 Applying proposed algorithm

Applied proposed methods (dictionary-based approach, Rule based approach and Random Forest Algorithm) on the pre processed data.

IV. EXPERIMENTAL RESULTS

The f-score value of biomedical named entity recognition system reached more than 85% but is less than general named entity recognition that can reach about 90%[Zhenfei Ju et al][12]. With the GENIA corpus data set they got 84.24% Precision and 80.76% recall , where on the same data set using HiddenMorkovModel[Zhou et al.][15]got 66.5% Precision and 66.6 Recall. With our proposed hybrid approaches (Combination of Dictionary based approach and Rule based approach) with Random Forest ensemble classifier we achieved 95.15% Precision and 92.64% Recall with 93.89% F-score on gene names and 96.75% Precision and 96.70% Recall with 96.73% F-score on disease names. Precision, an F-Score are calculated as,

$$\text{Precision(P)} = \frac{TP/}{TP + FP}$$

$$\text{Recall(R)} = \frac{TP}{TP + FN}$$

$$\text{F- Score} = \frac{(\beta^2+1)+(P*R)}{\beta^2\ (R + P)}$$

(As we weighted Precision and Recall equally, $\beta = 1$) . So we calculated the F-Score as 2*(P*R)/(P+R).

V. CONCLUSION

The need of Text Mining tool to discover and extract relation between entities is in a great need in the field of biomedical due to the rapid growth of biomedical documents reservoirs. To improve the performance in relation extraction between gene and disease we used a hybrid approach with Dictionary based approach , Rule based approach and used Random Forest algorithm to improve f-score. Our experimental results shown good results with 95% Precision, 92.64% Recall and 93.89% F-score on gene names . With disease names 96.75% Precision, 96.70% Recall and 96.73% F-score.

References

[1] Hearst,M.A., "Untangling Text Data Mining", Proc. 37th Annual Meeting of the Association for Computational Liguistics , College Park, MD,1999, pp.3-10.
[2] Creese, G.Duo-Mining:Combining data and text mining , DM Review,September,(2004).
[3] Kim J-D, Ohta T, Pyysala S, Kano Y, Tsujii, "Overview of BioNLP'09 shared task on event extraction ", In BioNLP'09 Proc Work Curr. Trends Biomed Nat Lang Process Shar Task. Association for Computational Liguistics;2009:1-9.
[4] Maryam Khordad, "Investing Genotype-Phenotype Relationship Extraction from BioMedical Text", The University of Western Ontario, London, Ontrio, Canada.
[5] A.S.Foulkes,"Applied Statistical Genetics with R : For population-based Association Studies, Use R",

International Journal of Computing Algorithm
Volume: 05 Issue: 01 June 2016, Page No. 46- 49
ISSN: 2278-2397

DOI:10.1007/978-0-387-89554-3-1.@Springer Science + Business Media LLC 2009.

[6] Alex Bravo,Janet Pinero, Nuria Queralt, Michael Rautschka and Laura l.Furlong.,"Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translationl research", bioRxiv preprint first posted online July 24, 2014; doi;http://dx.doi.org/10.1101/007443.

[7] Buyko, E..Beisswanger,E., & Hahn, U,(2012), "The extraction of pharmcogenetic and pharmacogenomic relations- a case study using PharmGKB, Pacific Symposium on Biocomputing, 376-87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22174293.

[8] Changqin Quan and Fujiren, "Gene-disease association extraction by text mining and network analysis ", Proceedings of the 5th International workshop on Health Text Mining and Information Analysis(Louhi)@ EACL 2014, pages 54-63, Gothenburg, Sweden, April 26-30 2014.(C)2014 Association for Computational Linguistics.

[9] Hong-woo Chun, Yoshimasa Tsoruoka, Jin-Dongkim, RieShiba, Noiki Nagata,TeruyoshiHishiki and Jun'ichi Tsujii,"Extraction of Gene-disease relation from Medline using domain dictionaries and machine learning",Pacific Symposium on Biocomputing 11:4-15(2006) October 12 ,2005,15:5 Proceedings.

[10] M.Krauthammer, A.Rzhetsky, P.Morozov, C.Friedman, "Using BLAST for Identifying Gene and Protein Names in Journal articles", Gene 259(2000) 245-252.

[11] Yi-Feng Lin, Tzong-Han Tsai et a., "A Maximum Entropy Approach to Biomedical Named Entity Recognition ", BIOKDD04: 4th Workshop on Data Mining in Bioinformatics(With SIGKDD Conference)

[12] Zhenfei Ju,Jain Wang , FeiZhu, "Named Entity Recognition from Biomedical Text using SVM", 978-1-4244-5089-3/11/ ©2011 IEEE.

[13] D.Hanisch, J.Fluck,H.Mevissen and R.Zimmer, "Playing Biology's Name Game: Identifying Protein Names in Scientific Text" , In PSB '03, 2003.

[14] Tsuruoka ,Y., Tsujii J, "Improving the Performance of Dictionary-based Approaches in Protein Name Recognition", J Biomed Inform 2004,37:461-70.

[15] Zhou G, et al., "Recognizing Names in Biomedical Texts : A Machine Learning Approach", Bioinformatics, Vol.20,No.4,2004,pp.1178-1190.

[16] Benjamin A,Goldstein, Eric C.Polley and Farren B.S. Briggs(2011)," Random Forests for Genetic Association Studies ", Statistical Applications in Genetics and Molecular Biology: Vol. 10 : ISS.1. Article 32.

[17] Fukuda K., et al., "Toward Information Extraction: Identifying Protein Names from Biomedical Papers", Proc. Pacific Symp.on Biocomputing, 1998, pp.707-718.

[18] Jong C.Park and Jung-jae Kim,"Named Entity Recognition", Text Mining for Biology and Biomedicine.

[19] www.youtube.com

[20] Proux D., et al., "Detecting Gene Symbols and Names in Biomedical Texts: A First Step Toward Pertinent Information.