Text Mining in Biosciences-A Review

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

Abstract – In this recent scenario, the research towards the area of text mining in biosciences are lacking. This paper is based on the survey made in biosciences using text mining in the document world. The data which do not have clear classification and answers can also be implemented and evaluated successfully by clustering. Now, analyzing the data plays a vital role which is done by information retrieval. Errors and evaluation has to be minimized and validated for accurate and immediate goal. This paper discusses various techniques that can be integrated with this multidisciplinary field of Text Mining. The different methods used by the researchers are discussed in this work.

Keywords : Natural Language Processing(NLP) , Information Extraction(IE) , Information Retrieval(IR),Topic Tracking, Categorization , Summarization, Clustering, Concept Linkage.

———————————— ◆ ————————————

1. Introduction

Text mining plays a vital role in the field of research as most of the information available in various sources is in the form of text and more than 80% of the data are in unstructured form. The main task of text mining starts with structuring the unstructured text, discovering the patterns from the structured text and then evaluating the output.The structured data stored in the database is then used for data mining [1]. Data mining tools are designed to work on structured data (data bases) or XML files where as Text mining tools can handle unstructured or semi-structured data sets like HTML files, emails , full-text documents. For text related research integration of Text mining and Data mining (“Duo-mining”) would be a better solution[2]. Text Mining involves extracting information from various sources automatically and discovering new previously unknown information[3]. The goal of IE is to find specific information from Natural language text. The entities, events, attributes that are extracted using IE approach are stored into the database on which Data Mining techniques can be applied.[4] Text is checked and make sure whether it is grammatically correct and fluent enough by Natural Language Generation(NLG). For this purpose NLG uses Syntactic realizer (Example application is Machine translation system). Natural Language Understanding (NLU) is used to compute meaning representation. Tokenization , morphological or lexical analysis , syntactic analysis and semantic analysis are the components that can be used with NLU[5].

The Natural Language Processing is used to design and build a computer system to analyze and generate structured text[6][7]. Information Extraction is used to extract structured information (entities, facts, relationships) automatically. The main focus of IE is collecting , organizing information and application of information to answer questions.[8]. Most of the IE are developed using the following steps[9][10]: Text Pre-processing, Rule Selection, Rule Application. IE uses Pattern Matching method to identify key phrases and relationship with in a text document. Pattern Matching is the process of matching predefined sequences of text with the text

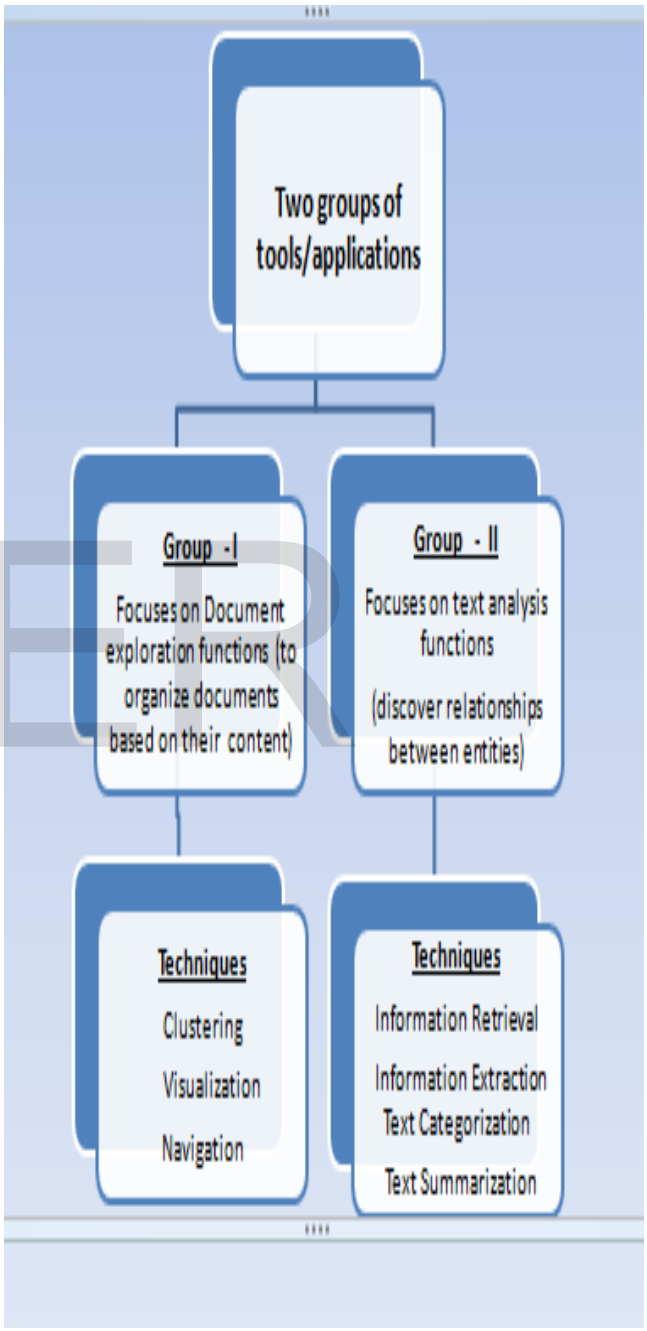
given by the user. To analyze large set of text data this technique can be used. The information extracted and stored in structured database need to be post-processed[11[12]. In Information Retrieval approach , based on user's query the relevant documents are located from a document collection. The two basic measures used for accessing the quality of text retrieval are Precision and Recall. Precision is the percentage of retrieved documents that are in fact relevant to the query. Recall is the percentage of documents that are relevant to the query and were in fact retrieved[13]. Topic tracking maintains the topic searched by user. The related documents to the previous documents are predicted effectively by the system next time [14] Companies can use this topic tracking to alert from a competitor. Some of the approaches that have been used for Topic tracking approach are VSM(Vector Space Model), Hierarchical clustering, Named Entity Recognition , etc., Topic tracking system can be used in the field of Medical industry to track patients condition and the treatment to be taken, in business to get the details of their company products and in News industry to check the articles consisting of same events.

2. Proposed System

2,1 Objectives

- To help biomedical researchers to extract knowledge from the biomedical literature to make new discovery in efficient way.
- To provide some ability to identify the relationship between entities
    a. Genes and other factors that cause disease
    b. Root cause and disease

2.2 Techniques of the Proposed System



2.3 Performance analysis table

The following table presents the data set , techniques/methods used and the precision and recall percentage obtained by different researchers.

Techniques/Methods , Precision and Recall Percentage analysis Table:

S.No	Author	Year	Method used	Dataset	Precision	Recall
1.	Chade-Meng Tan et al.	1998	Naïve Bayes	Yahoo-Science Reuters-21578	65.1% 90.8%	76.7% 57.6%
2.	Dr.Ahmed T.Shadiq et al.	2013	Rough set theory	Different categories of documents Comp.Science A I Database Security Mathematics Algebra Statistics	 100% 95.65% 83.33% 100% 95.23%	 100% 100% 95.23% 100% 90.90%
3.	Wongkot Sriurai	2011	LDA Topic-model Approach	Collectionof documents Topic-Model NB DTree SVM	 67.6% 72.5% 80%	 69.6% 71.2% 78%
4.	Alan Ritter et al.	2012	Information extraction TWICAL approach	Twitter	56%	74%
5.	Apte' et al.	1998	Decision Trees and Decision Rules	Reuters-21578	87.8%	87.8%
6.	Nitin Jindal	2006	Supervised learning approach	News articles, customer reviews on products, Internet forum postings	79%	81%
7.	Fukuda.K et al.	-	Information Extraction	Medical articles	94.70%	98.84%
8.	Johannes Furnkranz	-	Information Extraction	20 Newsgroups Reuters-21578	76.71%	83.42%
9.	Schutze.H	2003	Information Extraction	Tipster Corpus	36.78%	-

10.	Lewis.DD	-	Information Extraction	Reuters-21,450	65%	-
11.	Daniel M. McDonald et al.	2004	Semantic & syntax hybrid system	PubMed articles	89%	61%
12.	Mathieu Roche et al.	2008	Named Entity Recognition	Medline abstracts	66.7%	80%
13.	Zhou et al.	2004	Hidden Markov model	GENIA corpus	66.5%	66%
14.	Chang et al.	2002	Statistical learning algorithm	Medstract corpus	96%	84%
15.	Hanisch et al	2003	Information Extraction	Complete Dictionary	95%	90%
16.	Yu & Agichtein	2003	Machine learning techniques	Biological journal articles	90%	80%
17.	L. Tanabe et al.	2002	Rule–based approach	Medline articles	61%	35%
18.	Eskin & Agichtein	2004	Text & Sequence Mining with SVM	SWISS-PROT	87%	71%
19.	Khaled Khelif et al.	- 2007	Ontology based semantic annotations	Biological Documents	83.6%	60.8%
20.	David P.A Corney et al.	-	Information Extraction	BioMed articles	39%	48%
21.	Schwartz et al.	2003	Information Extraction	Medline abstracts	96%	82%
22.	Gaizauskas et al.	2003	Information Extraction	PASTA Corpus	82%	84%
23.	Chen and Friedman	2004	Named Entity Recognition	MedLee system	64.0%	77.1%
24.	Donaldson et al.	2003	Support Vector Machine	Medline abstracts	96%	84%
25.	K.Mythili et al.	2012	Pattern Taxonomy Model, D-Pattern Discovery	Document collection	70%	68%

3. Technologies that can be integrated with Text Mining

3.1 Document Classification

Text categorization is used wisely but maladroit , for document classification.

Standard documentation method organizes documents into folders, each folder for each concept.

3.2 Information Retrieval

Information Retrieval is the term commonly related with online documents. A basic concept is measuring similarity a comparison is made between two documents measuring how similar the documents are.

3.3 Clustering

Clustering is parallel to assigning the labels needed for text categorization. By implementing the different keywords that characterize a cluster by the clustering algorithm.

3.4 Information Extraction

To make the structured text, the proposed system implements a superficial representation that measures the single occurrence of words.IE is the subfield of text mining that attempts to move text mining onto an equal foothold with the organized world of data mining.

4.Prediction and Evaluation

The basic concept of prediction in our proposed system is the measurement of error. The system helps to determine whether the predicted answer is 'right or wrong' and also determines the measures of validity are primarily important to document analysis.

5.Conclusion

There is a great need of text analysis tool to discover and extract the massive wealth of hidden knowledge embedded in the digital data. Text Mining is an intelligent text analysis tool to extract the useful information from the natural language text and it is known as Knowledge Discovery in Text(KDT).This paper briefly discusses the concept of Text Mining in biosciences and various datasets involved with Text Mining. Techniques and methods used in this most popular research field by different researchers for different applications are discussed to get more information about this domain. In future study the proposed system will be extended to cluster the data using fuzzy clustering algorithm and to predict and evaluate the clustered data; the same can be implemented using neural networks.

References

1. Weiguo Fan, Linda Wallace , Stephanie Rich and Zhongju Zhang, "Tapping into the power of Text Mining", Journal of ACM , 2005.
2. Creese, G. Duo-Mining: combining data and text mining, DM Review, No. September, (2004), http://www.dmreview.com/article_sub.cfm?articleId=1010449.
3. Hearst.M , What is Text Mining http://www.sims.berkeley.edu/~hearst/text mining.html (2004).
4. Shaidah Jushoh and Hejab M.Alfawareh , "Techniques, Applications and Challanging Issue in Text Mining", International Journal of Computer Science Issues, Vol. 9, Issue6, No.2, November 2012.
5. S.Jusoh and H.M.Alfawareh, "Natural language interface for online sales", in Proceedings of the 2009 International Conference on Intelligent and Advanced System (ICIAS2007) Malaysia:IEEE, November 2007, pp.224-228.
6. Falguni N Patel and Neha R.Soni , "Text Miining : A Brief Survey", International Journal of Advanced Computer Research, Vol.2,No.4,Dec.2012 .
7. U.Ackermann, B,Angelini, F.Brugnara, M.Federico, D.Giuliani,

R.Gretter, G.Lazzari and H.Niemann, "Speed Data: Multilingual Spoken Data Entry", International Conference, IEEE, Trento,Italy, 2211-2214.

8. K.Prabhavathy and Dr.P.Sumathi,"Text Mining Interepreting Knowledge Discovery from Biomed Articles", The SIJ Transactions on Computer Science Engineering and its Applications , Col.1, No.2, May-June 2013.
9. R.Hobbs, D.Appelt, J.Bear, D.Israel, M.Kameyama, M.Stickel and M.Tyson, "FASTUS: A cascaded finite-state transducer for extraction information from natural Language text", in Finite States Devices for Natural Language Processing, E.Roche and Y.Schabes, Eds., 1997,pp.383-406.
10. J.Cowie and Y.Wilks ," Information Extraction" , New York, 2000.
11. N.Kanya and S.Geetha , "Information Extraction:A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr.M.G.R.University, Chennai, TamilNadu, India, 1111-1118.
12. Sergio Bolasco, Alessio Canzonetti, Fracesca Della, Ratta-Rinald and Bhupesh K.Singh, (2002), "Understanding Text Mining : a Pragmatic Approach", Roam, Italy.
13. R.Sagayam, S.Srinivasan and S.Roshni, "A Survey of Text Mining : Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research, Vol.2, Issue5.
14. Sungjick Lee and Han-joon Kim, "News Keyword Extraction for Topic Tracking ", 4th International conference on Networked Computing and Advanced Information Management, IEEE(2008), Korea.554-559.
15. Charusila Kadu, Praveen Bhanodia, Pritesh Jain , "A Review on Ontological- based Pattern Mining Tecniques", International Journal of Scientific & Engineering Research ", Volume 4, Issue 5, May-2013
16. Chade-Meng Tan ,Yuan-Fang Wang, Chan-Do Lee, "The use of Bigrams to enhance Text Categorization", Department of Computer Science , University of California.
17. Dr.Ahmed.T.Shadiq, Sura Mahmood Abdullah, "Hybrid Techniques for Text Categorization", Helvetic Editions 2013.
18. Wongkot Sriurai, "Improving Text categorization by using a topic model", Advanced:Computing:An International Journal (ACIJ), Vol.2,No.6,Nov.2011
19. Apte' .C, Dameran.F and Weiss.S, "Text Mining with Decision Trees and Decision Rules", Presented at the conference on Automated learning and discovery,Pittsburg, PA(1998).
20. Aaron M.Cohen and William R.Hersh, "A survey of current work in biomedical text mining", Henry Stewart publications 1467-5463,

Briefings in BioInformatics, Vol.6,No.1, 57-71, March 2005.

21. D.Hanisch, J.Fluck, H.T.Mevissen and R.Zimmer, "Playing Biology's Name Game: Identifying Protein Names in Scientific Text", Pacific Symposium on Biocomputing,3rd-7th January , Hawaii,pp.403-414,2003.
22. Johannes Furnkranz, "A study using n-gram features for Text Categorization", Technical Report OEFAI-TR-98-30., Austrian Research Institute for Artificial Intelligence, Vienna, Austria.
23. Schutze H, Hull.D and Pederson.J, "A comparison of classifiers and document representations for the Reuting problem ", in Croft et al.(Ed.), Proceedings of SIGIR-95, 15th ACM International conference and development in Information Retrieval(pp.215-223), NewYork:ACM Press.
24. David P.A, Corney, David T. Jones , Bernard F. Buxton, William B.Langdon, Joanne Charwood and Peter M. Woollard, "Extracting Biological Informatiom from Full-length papers ", Technical Report RN/03/17", Department of Computer Science , University College London, Gowerstreet, London,UK.
25. Mathieu Roche and Volaine Prince, "Managing the Acronym/Expansion Identification process for Text Mining Applications", manuscript published online 29th Dec. 2008.
26. Fukuda.K, Tsunoda.T, Tamura.A and Takagi.T, "Toward Information Extraction: Identifying protein names from biological papers", Proceedings of the Pacific Symposium on Biocomputing (PSB 98), pp.705-716,1998.
27. Lorraine Tanabe and W.John Wilbur , "Tagging gene and protein names in biomedical text", Bioinformatics, Vol.18, No.8, pgs.1124-1132.
28. Jeffrey.T.Chang, Hinrich Schutze and Russ B.Altman, "Creating an online dictionary of abbreations from medline", Journal of the American Medical Informatics Association, Vol.9,No.6, Nov/Dec 2002.
29. Yu.H and Agichtein E, "Extracting synonymous gene and protein terms from biological literature', Bioinformatics, Vol.19, Suppl.1,pp.340-349 , (2003).
30. Eskin E and Agichtein.E, "Combining text mining and sequence analysis to discover protein functional regions", in Proceedings of the 9th Pacific Symposium on Biocomputing,6th-10th January, Hawaii,pp.288-299,(2004).
31. Daniel M. McDonald, Hsinchun Chen Hua Su and Byron B.Marshall, "Extracting gene pathway relations using a hybrid grammar: The Arizona Relation Parser", BioInformatics, Vol.20 , issue 18, Oxford University Press, (2004).
32. Schwartz.A.S. and Hearst.M.A.,"A simple algorithm for identifying abbreviation definitions in biomedical text", in Proceedings of the 8th Pacific Symposium on Biocomputing, 3rd – 7th January, Hawaii,pp.451-462.

33. Gaizauskas.R, Demetriou.G, Artymiuk.P.J. and Willett.P , "Protein structures and information extraction from biological texts: The PASTA system", Bioinformatics, Vol.19(1),pp.135-143.
34. Chen .L and Friedman.C , "Extracting phenotypic information from the literature via natural language processing", in proceedings of the 11th World Congress on Medical Informatics, IMIA, San Francisco, CA,pp.758-762,(2004).
35. GuoDong Zhou, Jie Zhang , Jian Su , Dan Shen and ChewLim Tan , "Recognizing names in BioMedical Texts: A Machine Learning Approach", BioInformatics, Vol.20,pp.1178-1190.
36. K.Mythili and K.Yasodha, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining", International Journal of Science and Applied Information Technology, Vol.1, No.3, July-August 2012.
37. Khaled Khelif , Rose Dieng-Kuntz and Pascal Barbry, "An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain", Journal of Universal Computer Science, Vol.13, no.12(2007), 1881-1907.
38. Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, Tony Pawson and Christopher WV Hogue," PreBIND and Textomy – mining the biomedical literature for protein – protein interactions using a Support Vector Machine", BMC BioInformatics, Vol.4(1), p.11,(2003).
39. http://en.wikipedia.org/wiki/Documentclassification
40. David.D.Lewis, " Feature Selection and Feature Extraction for Text Categorization", Center for Information and Language Studies, University of Chicago, Chicago.