Impact Factor 5.8701 (ICI)

International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

Vol.5., Issue.3, 2017
May-June

RESEARCH ARTICLE

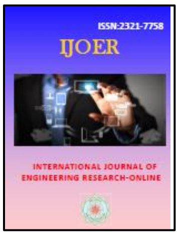
ISSN: 2321-7758

TEXT MINING TECHNIQUES AND TEXT MINING TOOLS IN BIOINFORMATICS

SARANGAM KODATI[1], Dr. R VIVEKANANDAM[2]

[1]Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore,Bhopal,Madhya Pradesh , (India)

[2]Professor, Department of Computer Science and Engineering,Sri Satya Sai University of Technology and Medical Science,Sehore,Bhopal, Madhya Pradesh , (India)



ABSTRACT

Recently, the development and application of digital data is increasing in various field, knowledge discovery and text mining have concerned great consideration with up coming requirement for turnoff such data into useful information and knowledge. The innovation of suitable patterns and movements to analyze the text documents from enormous volume of data is a big concern. Text mining is a process of extracting motivating and nontrivial patterns from huge extent of text documents. There are several methods and tools to mine the text and determine appreciated information for future expectation and decision making process. The assortment of accurate and correct text mining method assists to improve the speed and reduce the time and effort necessary to extract valuable information. The proposed systems concisely deliberate and evaluate the text mining technique, applications and text mining tools in bioinformatics field.

Keywords: , Text Mining, Text Mining Tools, Bioinformatics

1. INTRODUCTION

The size of available biology researches, and consequently the underlying knowledge, is growing at a high rate. After completion human genome sequencing in early 2000s, the adding of detailed genetic records to biomedical researchers, made the condition even more intense. An essential issue of research in biology informatics is how to use effectively of the great quantity of biological data to develop biological systems. Scientific literature in biology is the most important and reliable source of knowledge and information. At the moment, technical advances and specialized competitions with the aid of high-throughput technologies provided a huge amount of articles, which keeping up with them is practically impossible. Using computer aided automatic processing of texts, text mining, is a promising solution that helps experts with dramatically quick data processing within limited time consuming. In this field, the main aim of text mining exploitation is information retrieval and knowledge extraction from biology textual sources to promote new discoveries and help field experts in using them in realistic diagnosis, prevention and treatment. Mining is the method of inferring for patterns with among a structured or unstructured data. There are a number of mining techniques out about which they differ into the context and type about dataset that is applied. The process of extracting information and knowledge beyond unstructured textual content led to the need for various dig methods because useful pattern discovery. “Data Mining (DM) and Text Mining (TM) [2] is similar of that each methods “mine” huge amounts of data, searching because of meaningful patterns [3].”Some of the mining types are data, text, web, business Process and Bioinformatics.

2. TEXT MINING

Text mining is a burgeoning recent area that tries to extract meaningful information beside natural language text [6]. It may stay characterized as the method of analyzing text to extract information that is useful for a specific purpose. Compared along the kind on data stored on databases, text is unstructured, ambiguous, and difficult in conformity with process[5]. Nevertheless, between modern culture, text is the most communal way for the formal exchange of information. Text mining usually deals with texts whose function is the communication of actual information and opinions[9].

2.1 ELEMENTS OF TEXT MINING

As text mining mostly mimics the way database curator works, it starts with the determination of which resources to look at. These can be published papers, journal articles, patents, and even electronic medical records (EMRs). This step involves text classification and/or clustering, as well as information retrieval (IR)[1]. After selecting what to read, identification of important entities and relations between those entities in those selected documents are performed. These steps corresponds to name entity recognition (NER) and relation extraction in the context of text mining.

3. TEXT MINING TOOLS

A high level overview on text mining tools is stability according to provide a comparison involving textual content mining capabilities perceived strengths, potential limitations, relevant data sources or output results so applied within conformity with chemical biological after patent information. Examples on tools are given below consist of business enterprise name tool function output and website referenc.

3.1 DIFFERENT TYPES OF TEXT MINING TOOLS

We used the following search string to determine famous text mining tools [(Text) AND (Mining OR Analytics) AND (Tool)]. From the search results, we recognized 55 famous textual content mining tools but studied theirs features. Table 1. lists this tools along with theirs purposes and methods employed by using them. In the following sections, we analyze the popular strategies and features on textual content mining tools. durability Text Mining Tools be able lie classified into iii categories[4].

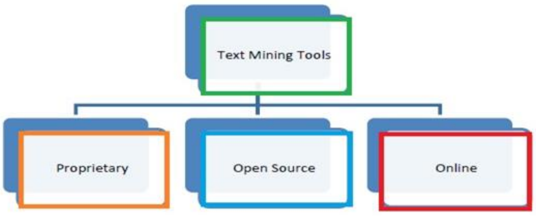


Fig 1: Different Types of Text Mining Tools

Proprietary Text Mining Tools: These tools are commercial text mining tools owned by a company. To use these tools purchase is required. Although demo/trial versions are available free of cost but have limited functionality. 39 out of these 55 tools are proprietary tools.

Open Source Text Mining Tools: These tools are available free of cost and also there source code and one can even contribute in their development.13 out of these55 text mining tools are open source.

Online Text Mining Tools: These tools can be run from the website itself. Only a web browser is required. These tools are generally simple and provide limited functionality. Three out of these 55 text mining tools are online web based tools[10].

Tool	Type	Techniques	Features/Uses	Website	Additional
Ranks.nl[10]	Online	Keyword analysis	Page Analysis, Article Analysis, Multi page analysis	Http://www.ranks.nl/	Website has been put together using Perl, Mysql, Javascript and HTML. Input Supported: Text/URL

Text Sentiment Visualizer	Online	Deep neural networks and D3.js.	Sentiment analysis	Http://sentiment.lucas estevam.com/	Input Supported: Text/URL
Textalyser	Online	Text Analysis, Keyword Analysis	Text analysis	Http://textalyser.net/	Input Supported: Text/URL
Alceste	Proprietary	Hierarchical descending classification, ascending classification, thematic classification	Textual data analysis, Multilingual analysis, temporal analysis	Http://www.image-zafar.com/Logicieluk.html	OS required-Win XP, VISTA, 7, 8 et Mac OS-X
Anderson Analytics odintext	Proprietary	Advanced statistics and other machine learning techniques	Text analytics	Http://odintext.com/#	
Ascribe	Proprietary	Hybrid technology approach, natural language processing, machine learning and semi-automated coding tools	Text analytics	Http://goascribe.com/	
Basis Technology Rosette	Proprietary	Linguistic analysis, statistical modeling, and machine learning	Text Analytics, multilingual text analytics	Http://www.rosette.c om/	Integrated with curl, Python, PHP, JAVA, C#, nodejs, Ruby
Buzzlogix text analysis api	Proprietary	Semantic Text Analysis using natural language processing	Text analysis, sentiment analysis, classification, keyword analysis	Https://www.buzzlog ix.com/text-analysis/	
Clarabridge	Proprietary	Linguistic and statistical algorithms, Natural Language	Text analytics	Http://www.clarabrid ge.com/text- analytics/	
Clustify	Proprietary	Classification	Categorization of documents	Http://www.cluster-text.com/	
Dataladder productmatch	Proprietary	Machine learning	Data cleansing, classification	Http://dataladder.com /products/productmat ch/	

Discovertext	Proprietary	Cloud-based text analytics, Active Learning machine classification engine	Text analytics	Http://discovertext.co m/	

Tool	**Type**	**Techniques supported**	**Features/Uses**	**Website**	**Additional Remarks**
Ranks.nl	Online	Keyword analysis	Page Analysis, Article Analysis, Multi page analysis	Http://www.ranks.nl/	Website has been put together using Perl, Mysql, Javascript and HTML. Input Supported: Text/URL
Text Sentiment Visualizer	Online	Deep neural networks and D3.js.	Sentiment analysis	Http://sentiment.lucasestevam.com/	Input Supported: Text/URL
Textalyser	Online	Text Analysis, Keyword Analysis	Text analysis	Http://textalyser.net/	Input Supported: Text/URL
Alceste	Proprietary	Hierarchical descending classification, ascending classification, thematic classification	Textual data analysis, Multilingual analysis, temporal analysis	Http://www.image-zafar.com/Logicieluk .html	OS required- Win XP, VISTA, 7, 8 et Mac OS-X
Anderson Analytics odintext	Proprietary	Advanced statistics and other machine learning techniques	Text analytics	Http://odintext.com/#	

Ascribe	Proprietary	Hybrid technology approach, natural language processing, machine learning and semi-automated coding tools	Text analytics	Http://goascribe.com/	
Basis Technology Rosette	Proprietary	Linguistic analysis, statistical modeling, and machine learning	Text Analytics, multilingual text analytics	Http://www.rosette.c om/	Integrated with curl, Python, PHP, JAVA, C#, nodejs, Ruby
Buzzlogix text analysis api	Proprietary	Semantic Text Analysis using natural language processing	Text analysis, sentiment analysis, classification, keyword analysis	Https://www.buzzlogix.com/text-analysis/	
Clarabridge	Proprietary	Linguistic and statistical algorithms, Natural Language Processing (NLP).	Text analytics	Http://www.clarabridge.com/text- analytics/	
Clustify	Proprietary	Classification	Categorization of documents	Http://www.cluster-text.com/	
Dataladder productmatch	Proprietary	Machine learning	Data cleansing, classification	Http://dataladder.com /products/productmat ch/	
Discovertext	Proprietary	Cloud-based text analytics, Active Learning machine classification engine	Text analytics	Http://discovertext.co m/	

Tool	Type	Techniques supported	Features/Uses	Website	Additional Remarks
		language processing.	social media analysis		
Megaputer Text Analyst	Proprietary	Linguistic, semantic, statistical and machine learning techniques.	Text analytics	Http://www.megaputer.com/site/textanalys t.php	
Monkeylearn	Proprietary	Machine learning, natural language processing, classification, extraction, clustering and regression	Text analysis	Http://monkeylearn.c om/	Integrated with php,python,.net,java ,ruby, javascript
Netowl (from SRA Internationl)	Proprietary	Advanced computational linguistics, natural language processing, machine learning	Multilingual text and entity analytics, document categorization, text mining	Https://www.netowl. com/text-analytics/	
Ontotext	Proprietary	Semantic graph database	Knowedge discovery, content managemnet, semantic search	Http://ontotext.com/	
Polyvista,	Proprietary	Pre-built recognition algorithms	Text analysis	Http://www.polyvista .com/	

Picture safe	Proprietary	Statistical methods, core linguistic principles,	Categorization, clustering, text analysis, audio video content analysis	Https://www.picturesafe.de/en/products/products-semantic- analysis/	
Power Text Solutions	Proprietary	Multi-document summarization technology, non-query-biased summarization of documents	Text analysis	Http://www.powertextsolutions.com/#/home	
Right find (tm) XML for Mining	Proprietary	Knowledge discovery techniques	Build a corpus of full-text articles in XML format useful for text mining	Http://www.copyright.com/business/xmlfor mining-2/	
SAS Text Miner	Proprietary	Predictive models, machine learning, natural language processing, data mining techniques	Text processing and analysis, Document theme discovery.	Http://www.sas.com/en_us/software/analyt ics/text-miner.html	OS Required :HP/UX on Itanium, IBM 64-Bit Enabled AIX, Linux (x86-64), Microsoft Windows (x86-64), 64-Bit Enabled Solaris on SPARC, Solaris on x64

SIFT	Proprietary	NLP, machine learning	Text analysis for customer feedback analysis process	Http://www.siftnlp.co m/	Browser must have scripts enabled.
Skyttle API	Proprietary	Sentiment analysis and keyword extraction, NLP	Text analytics	Http://www.skyttle.c om/	
Swapit, Fraunhofer-FIT text and data analysis tool (updated version of docminer)	Proprietary	Docminer text mining engine, state-of-the-art methodologies from statistics, retrieval, artificial intelligence and visualisation	Text and data analysis,	Https://www.fit.fraunhofer.de/en/fb/risk/projects/swapit.html	XML-based (SOAP protocol) Inter-service communication. Graphical user interface realised in Java technology
Textpipe Pro	Proprietary	Text processing	Text conversion,	Http://www.datamyst	

4. TEXT MINING IN BIOINFORMATICS

Bioinformatics is the area that combines computer science, information technology, and biology. Tools provided through bioinformatics help scientists analyze and explain various types regarding data, including sequences over amino acids, numerical or textual data[8]. Research areas in the area of bioinformatics[6] include sequence analysis, genome annotation, literature mining, and analysis of many other biological subjects. Beside others, literature mining is the key area to that amount deals with the analysis and interpretation about textual data and it is done by using the help of the text mining methods.

4.1 TEXT COLLECTIONS USING IN BIOINFORMATICS

The most prominent text collection for the text mining community in bioinformatics is, without a doubt, the MEDLINE which contains over 21 million paper abstracts. Even though abstracts are freely available, text mining community has no access after near of the full-text articles[7], especially on journal articles appropriate to the copyright issues. Considering that full-text variations contain a lot more data compared according to abstracts, that is the biggest setback because of the community. Nevertheless, into some fields, such as chemistry, the situation is even worse, where too article abstracts are inaccessible[8].

Full Text Corpora

URL

HighWire Press
http://highwire.stanford.edu
PubMed Central
http://pubmedcenral.org
Tagged Corpora
FetchProt
http://fetchpod.sics.se
GENETAG
ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe
GENIA
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
PennBioIE
http://bioie.ldc.upenn.edu
Yapex
http://www.sics.se/humle/projects/prothalt

Above lists the most known text collections in the world of the biomedical text mining.

5. CONCLUSION

Text mining generally refers to the process of extracting valuable information from unstructured text. In this survey of text mining, several text mining and text mining tools and applications in various fields have been discussed. For text mining in bioinformatics, the future is coming fast. Even though it is a very new field, them motivation behind this progress is very strong, fueled with the importance of health care in our lives, as well as prospective commercial aspects of the field. Following discussion elaborates on each issue that will come up in the future of the text mining using in Full Text implementation, Being more user focused, Data Integration in bioinformatics.

REFERENCES

[1]. Malhotra, Ruchika, et al. "Severity Assessment of Software Defect Reports using Text Classification." *International Journal of Computer Applications*83.11 (2013).

[2]. Vidhya. K. A and G. Aghila, "Text Mining Process, Techniques and Tools: an Overview", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 613-622, 2010.

[3]. Bragge, Johanna, and Jan Storgårds. "Profiling academic research on digital games using text mining tools." *Proceedings of DiGRA 2007 Conference*. 200

[4]. S. Niharika, V. Sneha Latha and D. R. Lavanya, "A Survey on Text Categorization", International Journal of Computer Trends and Technology, ISSN: 2231-2803, Volume 3, Issue 1, pp. 39-45, 2012.

[5]. K. Sarangam, P. Vijay pal Reddy, Vishnu Murthy.G, Dr. B. Vishnu Vardhan "A comparative study on term weighting methods for automated telugu text categorization with effective classifiers" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, November 2013

[6]. Daniel Waegel. —The Development of Text-Mining Tools and Algorithms.Ursinus College, 2006.

[7]. Qi, Y., Y. Zhang, et al. (2009). Text Mining for Bioinformatics: State of the Art Review, IEEE.

[8]. A. M. Cohen and W. R. Hersh, —A survey of current work in biomedical text mining,‖ (Briefings in bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.)

[9]. Lokesh Kumar and Parul Kalra Bhatia,"Text Mining:Concept,Process,Applications," Journal of Global Research in Computer Science Volume 4, No. 3, March 2013 .

[10]. Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.