PROCEEDINGS

978-9380686-77-6

Avinashilingam
Institute for Home Science and Higher Education for Women
University
(Estd. u/s 3 of UGC Act 1956)
Coimbatore, Tamil Nadu, India
(Deemed University under Category 'A' by MHRD)
Re-accredited with 'A' Grade by NAAC

UGC SPONSORED
NATIONAL SEMINAR ON

BIG DATA CHALLENGES AND OPPORTUNITIES

- A Perspective on Security & Social Media Analytics

Organized by

DEPARTMENT OF COMPUTER SCIENCE

19TH & 20TH FEBRUARY 2015

Shanlax Publications

BIG DATA CHALLENGES AND OPPORTUNITIES
– A Perspective on Security and Social Media Analytics

February 19-20, 2015

ISBN: 978-93-80686-77-6

Publisher

SHANLAX PUBLICATIONS
61, T.P.K. Main Road, Vasantha Nagar
Madurai – 625003
Ph: 0452-4208765 Mob: 9600303383
email: shanlaxpublications@gmail.com
web: www.shanlaxpublications.com

A Questionnaire for Analytic Data Collection in Child Education
Jitendra Kumar Jaiswal, Rita Samikannu, Dr.Ilango P. 77

Parallel PSOK Clustering Algorithm with Data Sampling for Large Data Sets
M Thangarasu, Dr. H Hannah Inbarani 82

Text Mining and Visualization Techniques - A Survey
K.T.Mathuna, Dr. I.Elizabeth Shanthi 87

Data Mining and Visual Analytics - The State-of-Art Tools and Techniques
Nandhini. K, Dr. I. Elizabeth Shanthi 93

Review on Educational Data Mining from Indian Authors perspective
Dhanalakshmi.D, Dr. J.Komala Lakshmi 99

Hadoop Based Diminution of Health Care Outlay Using Big Data Analytics
B.Meena Preethi, S.Subha Indu 102

Cloud enabled Big Data Analytics
R.Rama Prabha, R.Sujithra,Dr.Vasantha Kalyani David 105

Survey on Threat Challenges in Manet with the Support of Bigdata Analytics
S. Mohamed Imranul Hasan 109

Transitioning from Relational Databases to Big Data
M.Anitha, J.Sindhumathi 114

Big Data Infrastructure
R.Suganya, S.Thangam 118

Big Data's – Security and Privacy
P.Geethanjali, K.Kirthika 125

Analysis of Data towards Efficient Waste Management
Dr. Vasantha Kalyani David, S. Nandhini 131

Text Mining and Visualization Techniques - A Survey

K.T.Mathuna
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
mathuna.thangaraj@gmail.com

Dr. I.Elizabeth Shanthi
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
shanthianto@gmail.com

Abstract— *Data mining is the process of extracting hidden patterns from structured data. Text mining is an ongoing trend and has a wide range of applications where a huge volume of data is considered. The growing amount of data has lead to the concept of Big data. Data management is important in all fields for understanding the past, experimenting the present and for predicting the future. Text is the major component of all data and is generated everywhere and Knowledge Discovery is an interesting phenomenon. Text mining has a variety of applications and the techniques applied also vary with respect to them. Text mining algorithms play a vital role in providing a structured framework. Recent text visualization tools and techniques are to be reviewed in this paper. This paper analyses various state of the art algorithm tools and techniques for text mining and text visualization along with the commonly used dataset and their applications in various fields.*

Index Terms— Text mining, Text analytics, Visualization tools, Static and Dynamic visualization.

I. *INTRODUCTION*

Text mining is different from data mining. Data mining uses stored information in a structured manner, whereas text mining uses texts that are unstructured or semi-structured. Text mining has now become an important aspect because of the increased amount of information generated. There is a huge growth in the usage of public media like World Wide Web, internet, facebook, Linked in and so on[1]. Even the databases used in organizations, educational institutions, government organization and other industries have increased. Therefore the application of text mining has expanded in various domains.

Humans are able to explore the content and correlate with the patterns and have the ability to differentiate the context, spelling and the grammatical syntax which computers cannot. Making the system think like human brain is a challenging task and text analysis is trying to overcome these difficulties. Text analytics is used to extract high-quality information from text through deep analysis [2].

All the information and data that leads to knowledge are in many forms like documents, electronic files, tables, e-mail, information sheets and reports. Texts collected from these sources are preprocessed to clean and format the document [3]. Classification and clustering algorithms are applied to mine the data and finally effective visualization tools and techniques are required to extract user queries from those semi- structured and unstructured formats. Text Visualization is a key concept which projects the results in a more understanding and self explanatory way to the user [4].

The organization of the paper is as follows. Section 2 deals with the literature in text mining. Section 3 explains about the text mining process and the algorithms used. Section 4 lists the visualization tools and techniques available. Section 5 gives the application of text mining and Section 6 concludes the paper.

II. *TEXT MINING PROCESS AND ALGORITHMS*

A. Process in Text mining

Text mining starts with data collection from multiple sources. The gathered information will be in different formats like word, document, html, pdf, excel sheets, css, one note and so on[5]. The second step is data preprocessing, where the data gathered is cleaned to reduce noise, redundancy and missing values in the document. Preprocessing of text data is done through processes like tokenization, stop word removal and stemming. The former one is the process of splitting the sentence into single word called tokens. The following eliminates prepositions, conjunctions and articles from the splitted words. And the last one leaves the root word by removing past, present and future tens.

The third step in text mining process is data warehousing, in which data that is required are extracted and cleaned through feature generation models like Vector Space model and Bag of words[6]. Data transformation is the next step, where features generated are selected based on their counting and statistics.

The following step is text mining or pattern mining, where text mining merges with data mining.

Effective supervised and unsupervised data mining algorithms are used.

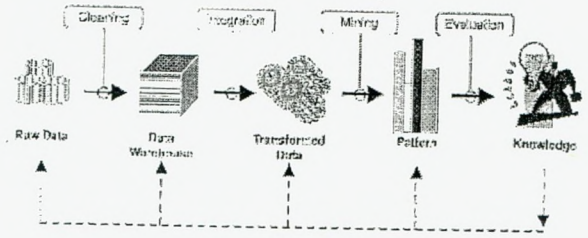


Figure 1: Text mining process

Knowledge discovery by evaluating the mined data is the final process and visualization techniques project the interpreted results in a user friendly manner. Fig 1 shows the text mining process.

B. Algorithms

Algorithms are essential to computers for processing information. They inform the system to do work according to steps it has to follow. Algorithms are framed based on the type of process it has to perform. Several algorithms are formulated based on their applications. Classification, Clustering and Association rules are some of the popular text mining algorithms.

Text classification is the method to build class models based on the selected attributes. These models are used to classify the new records based on the previously formed class labels [7]. Many algorithms are framed for training and testing the datasets based on the pre defined classes. The classification algorithms like C4.5, Decision tree, Support vector machine, k-nearest neighbor, Naïve Bayes, Linear Least squares, Voting, Associative classifiers, Neural networks and CART algorithms are applied for text mining [8].

Clustering is an unsupervised learning where no labeled data is available and it groups data into certain number of clusters. The objects in a cluster have similar

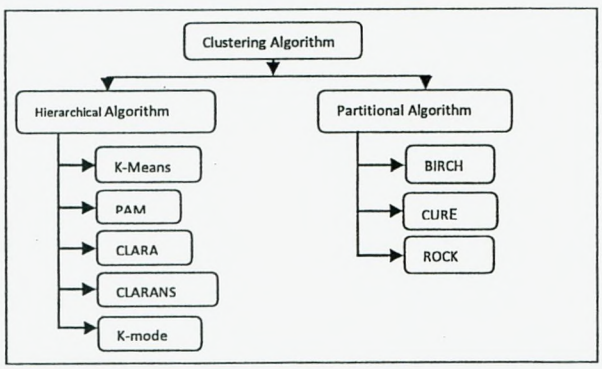


Figure 2: Clustering algorithms for text mining.

characteristics and high differences with other clusters. The text mining algorithms for clustering are categorized into hierarchical and partitional algorithms[9]. Fig 2 shows the clustering algorithms in text mining.

The algorithms indicated like K-Means algorithm, Partitioning Around Medoids(PAM), Clustering Large Application(CLARA), Clustering Large Applications based on Randomized Search (CLARANS), K-mode, Balanced Iterative Reducing and Clustering Using Hierarchies(BIRCH), Clustering Using Representatives(CURE) and Robust Clustering using linKs(ROCK) are commonly used in text mining .

Association is a powerful data analysis technique that is frequent in data mining. Text mining imports association rules which "aim to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other repositories"[10]. Some of the association rule algorithms used in text mining are Apriori, Direct Hashing and Pruning (DHP) Algorithm, Partitioning algorithm and Sampling algorithm[11].

III. Tools And Techniques For Text Visualization

A. Text Visualization Tools

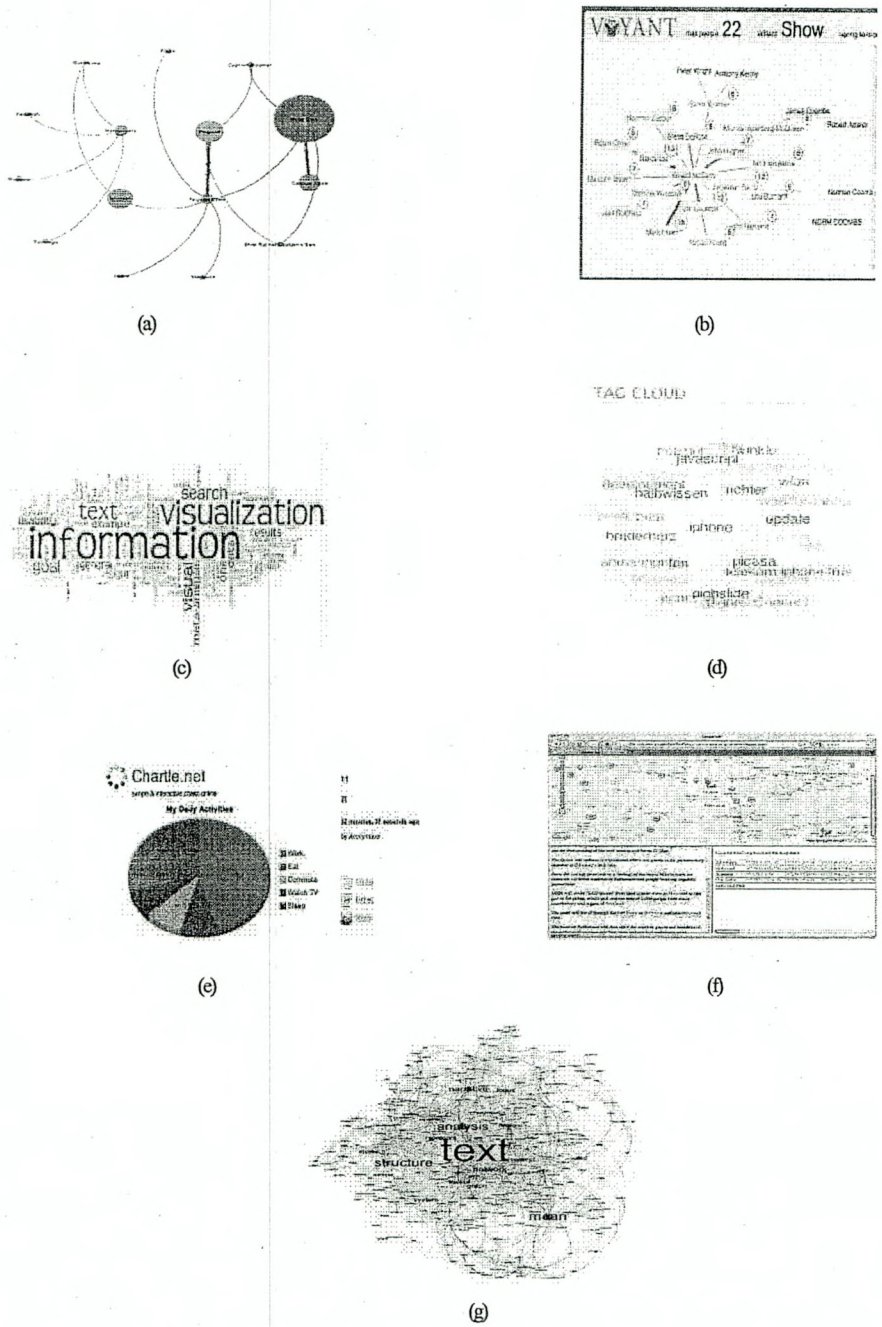
Text visualization is the final process in text mining. The text is processed, algorithms are applied and the refined knowledge is displayed in a better way. This provides simple and fast understanding about the result. Some of the text visualization tools are given in this chapter. Some of the recent tools for text visualization are Antconc, Voyant Tools, Wordle, TagCrowd, Chartle.net, Edinburg GeoParser and Gephi. Fig 3 shows the text visualization tools.

1) Antconc

Searches the keyword in the context and collocates from the document set. It is a cross platform for Mac,Windows, and Linux. Fig 3(a) shoes the Antconc tool.

2) Voyant Tools

Voyant is a free web based suite tool. It contains panels like Cirrus which gives frequent words and unique words. Corpus panel gives the summary of all the documents and highlights words. Words in a document are listed with counts and trends are shown as trellis chart. Word trends is a link graph having word frequencies. Keyword in context panel is similar to AntConc tool. Fig 3(b) shows the Voyant tool.



(a) (b) (c) (d) (e) (f) (g)

Figure 3: (a) Antconc tool (b) Voyant (c) Wordle (d) TagCloud (e) Charle.net (f) Edinburg GeoParser (g) Gephi.

3) Wordle

Wordle is a free web based word cloud generator. It displays the word important words bigger and the less important words smaller and can translate the text to other languages. The text are taken form web url, ATOM and RSS feeds. Fig 3(c) shows the Wordle tool.

4)TagCloud

Tag cloud is similar to wordle but has extra features like for non commercial use and can import files. Fig 3(d) shows the TagCloud.

5)Chartle.net

It is also a free web based software. Charts are created based on the word frequencies. The style of charts are like pie chart, line or bar chart, venn diagram, radar chart and scatter plots. Fig 3(e) shows the Charle.net tool.

6)Edinburg GeoParser

File is uploaded and the places in text are located. Places are marked on the google map and links can be jumped from location to location using the coordinates. Fig 3(f) shows the Edinburg GeoParser.

7)Gephi

Gephi is a node link structure. The keywords are the nodes and the link between the node shows the relationship between the words. Network of words, clusters and associations with text are formed for analysis. Fig 3(g) shows the Gephi tool.

B.Techniques

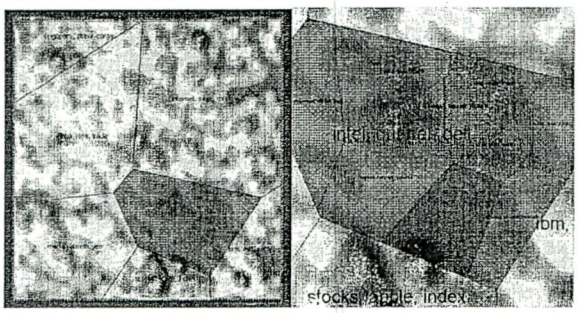
Text mining has a wide range of applications and text visualization has gained interest in recent years. Visualization allows easy and fast way of capturing the content. A brief content of the recent categories in text visualization is given in this subsection.

1) Static Text Visualization

The static text visualization is based on two features : Topic based text visualization and feature based text visualization.

In Topic based method, the documents in text corpora is getting hold of more significance for exploratory study of the data. *Tag clouds* is the depiction of keywords or named entities. They are popular and uses features like colour, size and layout based on usability and importance . Fig 4 shows tag cloud.

Figure 4: A visualization showing a search result set as a combination of tag clouds. Each polygonal area corresponds to a category of the documents in the search result set. Displayed named entities are enhanced with symbols indicating their type (person, location, data).[12]



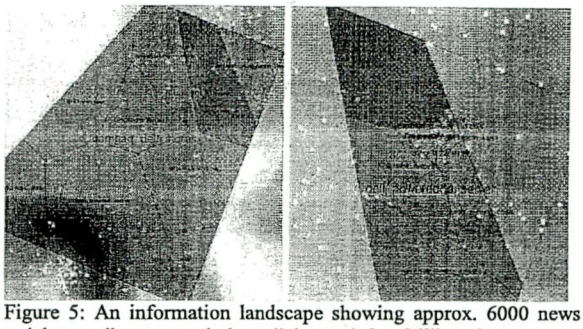


Figure 5: An information landscape showing approx. 6000 news articles on "computer industry" is used for drilling down to documents of interest: beginning with an overview (left) the user narrows down using topical cluster labels (right).[12]

Information landscape provides a geographical view of large set of documents for analysis. Navigation aids users who are unfamiliar with the data empowering to gain insight deep into the topic structure, understand the importance of various topics and slant about relationships between them. Fig 5 shows the information landscape. TextFlow[13] integrates topic mining techniques into interactive visualizations to visually analyze the evolution of topics in due course.

Whisper is a system for visual analysis of information diffusion. It uses a visual metaphor to design a hierarchical social spatial layout for visualizing the propagation of a typical event over time on Twitter. More recently, Xu *et al.*[14] studied the antagonism among topics through information diffusion on social media as well as the impact of opinion leaders on competition.

In Feature Based method, *Word clouds* that are commonly applied and provides an intuitive visual summary of documents by displaying the keywords in a compact layout. Facet Atlas integrates a node-link diagram into a density map to visually analyze the multifaceted relations of document[15]. DA viewer was designed to help linguistics researchers to study the discourse of language through tree layout. Oelke *et al* described an interesting visual analysis application for answering "How to make your writings easier to read"[16]. A visual analyltic system called VisRA was developed by them.

2) Dynamic Text Visualization

Time is the most challenging feature in text visualization. The varying content of the temporal documents has lead to dynamic text visualization. Recent study of text visualization helps in story creation, relating documents and leaving irrelevant patterns. Spark lines are combined with word cloud to form spark clouds and these projects word frequencies[17]. Theme River shows keywords with different colours with a bulge at high word frequency. Text flow allows to merge and split the text streams.

Event River groups the news corpus as a clusters. Each cluster is about a particular news and the density dependence on the period of that news[17].

A. Datasets

Datasets are the collections of records. Benchmark datasets provides the data collected from different sources. They are available to train and test the models developed by research scholars. They are even used to analyze certain aspects of products and their impact.

Text Datasets are available in repositories, blogs and resource software. The most popular benchmark repositories are Reuters, 20 Newsgroup, Datasets for Data Mining, Analytics and Knowledge Discovery by KD-nuggets, UC Irvine Machine Learning Repository by UCI KDD Archives, Frequent Itemset Mining Dataset Repository by FIMI workshops (2003/04), EDM datasets by PSLC DataShop, The KEEL dataset repository by KEEL Spanish Research Project[18,19].

Some of the the datasets available in blogs are MobBlog has datasets for research like Trust, Reputation, Recommendations and Mobility and The DataWrangling Blog has more than 400 dataset bookmarks. Datasets available in resources and softwares are Orange , KNIME (Konstanz Information Miner, RapidMiner, KEEL , Weka, Frequent Itemset Mining Implementations Repository for the Implementation of several algorithms for Frequent Itemset Mining[20].

IV. APPLICATIONS

Text mining has a broad range of applications. Uncovering the hidden pattern from deep inside the textual sources is a challenging task.Now a days there is a huge scope in mining the text data. Therefore the application of text mining is a never ending process. This chapter gives some of the text mining applications.

A.Bioinformatics

Research peoples in biomedical find it hard to analyze the abundant documents available in digitalized forms[18]. Document stored in various data sources are to be analyzed for their literature. Text mining techniques and the algorithms are used to pull the research content from the web sources. Doctor's relates the symptoms and diseases with the drugs available where Protein sequence and the study of required nutrients to a particular disease is extremely important [19].

B.Business Intelligence

Business intelligence mainly deals with decision making. Knowing the competitors, the market value, customers, products and their own position in the global trend is much more important[20]. Taking the right decision and at the right time makes the business successful. Most of the data in business are available in text like memos, reports, planning documents, regulation and the customer requirement[21]. These documents are effectively processed through Business Intelligence techniques. Competitive Intelligence is one such technique used to take decisions efficiently[20]. Cloud computing is the next technique that integrates with Business Intelligence where heterogeneous databases will be analyzed for better decision making.

C. National Security

The application of text mining is emerging and important in national security [21]. Protection of their own country from terrorism is a very critical task. Hence, government is tracking all types of resources through which information are passed. Electronic formats like Email, fax and message are the communication media which are tracked for identifying threatening and offensive materials. Text analysis method offers the tracking of materials in an effective manner. Although, not much research is conducted in this field like bioinformatics, the importance of national security is realized[21].

V. CONCLUSION

In this paper, an overview of the commonly applied text mining and visualization techniques are presented. Algorithms that could be applied for various purposes are given. Classification, clustering and association algorithms that are applied in text mining will provide and insight to further research. Current visualization tools and their application in documents make end users more interesting for clear understanding. The static and dynamic text visualization techniques with word frequencies provides clear understanding of the content. Application and their impact in the field of text mining is gaining more importance due to the huge growth of the electronically generated textual formats. Both the unstructured and semi structured formats are still a challenging task. As the growing amount of information will continue, the need of text mining and text visualization has more scope and wider research openings in all areas in the forthcoming decade.

REFERENCES

[1] Mr. Rahul Patel , Mr. Gaurav Sharma "A survey on text mining techniques" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 5 May, 2014 Page No. 5621-5625.

[2] Vishal Gupta, Gurpreet S. Lehal "A survey of text mining techniques and applications" Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[3] Mr. Rahul Patel, Mr. Gaurav Sharma "A survey on text mining techniques" International Journal Of Engineering And Computer Science ISSN:2319-7242. Volume 3 Issue 5 May, 2014 Page No. 5621-5625.

[4] Steinberger, Josef, and Karel Jezek. "Using latent semantic analysis in text summarization and

summary evaluation." In *Proc. ISIM'04*, pp. 93-100. 2004.

[5] Patil Monali, Kankal Sandip "A concise survey on text data mining" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 9, September 2014. ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940.

[6] Lokesh Kumar and Parul Kalra Bhatia, "Text mining: concept, process, applications," Journal of Global Research in Computer Science Volume 4, No. 3, March 2013.

[7] Mrs. Sayantani Ghosh1, Mr. Sudipta Roy2, and Prof. Samir K. Bandyopadhyay3 "A tutorial review on text mining algorithms" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012.

[8] Vandana Korde, C Namrata Mahender, "Text classification and classifiers: A survey text classification and classifiers: A survey" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[9] Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques." *Journal of Intelligent Information Systems* 17.2-3 (2001): 107-145.

[10] Vishwadeepak Sing Baghela, Dr.S.P.Tripathi " Text mining approaches to extract interesting association rules from text documents" International Journal of Computer Science Issues,Vol. 9,Issue 3, No 3, May 2012 ISSN(Online):1694-0814.

[11] John D.Holt and Soon M.Chung "Efficient mining of association rules in text databases" Department of computer Science and Engineering, Wright State University, Dayton, Ohio USA.

[12] Thomas J J, Cook K A. Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Ctr, 2005.

[13] Cao N, Lin Y R, Sun X et al. Whisper: Tracing the spatiotemporal process of information di®usion in real time. IEEE, Transactions on Visualization and Computer Graphics, 2012,18(12): 2649-2658.

[14] Xu P, Wu Y, Wei E, Peng T Q, Liu S, Zhu J H, Qu H. Visual analysis of topic competition on social media. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): to be appeared.

[15] Cao N, Sun J, Lin Y R et al. Facetatlas: Multifaceted visualization for rich text corpora. IEEE Transactions on Visualization and Computer Graphics, 2010, 16(6): 1172-1181.

[16] Guo-Dao Sun, Ying-Cai Wu,Rong-Hua Liang and Shi-Xia Liu "A survey of visual analytics techniques and applications: State-of-the-Art research and future challenges" Journal Of Computer Science And Technology 28(5): 852{867 Sept. 2013. DOI 10.1007/s11390-013-13838.

[17] Shixia Liu · Weiwei Cui · YingcaiWu · Mengchen Liu "A survey on information visualization: recent advances and challenges" Published online: 10 January 2014 © Springer-Verlag Berlin Heidelberg 2014.

[18] Rafael Geraldeli Rossi,Ricardo Marcondes Marcacini, Solange Oliveira Rezende "Benchmarking text collections for classication and clustering tasks" Institute of Mathematics and Computer Sciences, ISSN - 0103-2569, Sao Carlos, SP, Brazil November/2013.

[19] Wen Zhang, Taketoshi Yoshida, Xijin Tang "Text classification based on multi-word with support vector machine" Knowledge Based systems 21(2008) 879-886, Elsiver B.V. All rights reserved.

[20] Deshmukh J.J. And Tated R.R. "Weka - open source technology, its implementation and benefits", World Research Journal Of Computer Architecture Issn: 2278-8514 & E-Issn: 2278-8522, Volume 1, Issue 1, 2012, Pp.-01-05.

[21] Shaidah Jusoh and Hejab M. Alfawareh "Techniques, applications and challenging issue in text mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814.

[22] Ren´e Witte and Christopher J.O. Baker "Combining biological databases and text mining to support new bioinformatics applications" A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 310 321, 2005.c_Springer-Verlag Berlin Heidelberg 2005.

[23] Ranveer Kaur, Shruti Aggarwal "Techniques for mining text documents" International Journal of Computer Applications (0975 – 8887) Volume 66– No.18, March 2013.

[24] Shaidah Jusoh 1 and Hejab M. Alfawareh "Techniques, applications and challenging issue in text mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814.