Reviewing the Pathway of Text-Mining Approaches to Gauge the Applicability in Data Analysis

Jalaja G.
Associate Professor
Dept. of Computer Science & Engg,
BNM Institute of Technology
VTU, Belgavi, India

Sajitha N.
Asst. Professor
Dept. of Computer Science & Engg,
BNM Institute of Technology
VTU, Belgavi, India

K.R.Udaya Kumar Reddy, PhD
Prof.: Dept. of Computer Science & Engg.
BNM Institute of Technology,
Bengalore

ABSTRACT
With the increasing usage of mobile network and advancement in telecommunication standards, there is a massive growing of social networks those shares and exchanges giant forms of data. Although such forms of information are in various forms e.g. audio, video, text, image, or some specific file formats, but majorities of the transactional data are still in the form of text. Although there is various researches works being carried out in the area of datamining as well as text mining for more than a decade, the commercial usage of such tools is still not practiced owing to various challenges that are unsolved till date. Hence, the prime motive of this paper is to discuss about the fundamentals of text-mining and various significant issues associated with it. It also discusses about some of the review studies being discussed till date on same topic and updates the existing system by presenting more recent information of studies carried out towards text-mining most recently. Finally, the paper discusses exclusively the limitation explored in the existing system and then discusses about the research gap.

Keywords
Data mining, Document Clustering, Text-Mining, Machine Learning, Pattern Recognition

1. INTRODUCTION
With the rise of social networking application, mobile networks as well as with pervasive computing, the area of storage network has gained enough significance. Although there are exchange of various kinds of information in such applications but 95% of transacted information are in the form of text. Till date storage has never been a problem for repositing such massive size of textual data, but it is really a complex as well as challenging task to perform analysis on such types of text-data [1]. From more than past decade, there has been a continuous effort of enriching the field of text-mining. As majority of the existing applications runs over mobile networks, hence generalized form of data is usually found in the form of text and bears multi-dimensional challenges too [2]. Therefore in the recent years, text mining is one of the most prominent topics of research in data analytics. Basically, text mining is a type of datamining technique where the data are in the representation of text and it involves a sophisticated method of extracting a rich-quality of data from the given text. Various forms of statistical techniques are implemented in order to visualize the data patterns as well as trends of the text [3]. With the adoption of various sorts of parsing mechanism as well as inclusion of various linguistic charecteristics, the operation of text mining is carried out [4]. It also involves a significant removal of data redundancy and inclusion of significant knowledge from the highly structured data. The outcome of the text mining approach is studied using performance factor that is very closely associated with novelty, relevance, and highly informative data as knowledge. The application of text mining can be seen in the area of summarization of documents [5][6], sentiment analysis[7][8], modelling of entities [9][10], etc. The conventional process of text mining includes acquisition of textual data followed by making the data free from noise. This process is also called as pre-processing as well as transformation process. Once the textual elements were preprocessed, it is subjected to modelling, where the operation of data discovery, extraction of knowledge, and structuring of knowledge is carried out. Finally, the system conventionally uses feedback loop and perform evaluation / validation along with its usage in applications. The entire process is pictorially shown in Fig.1.

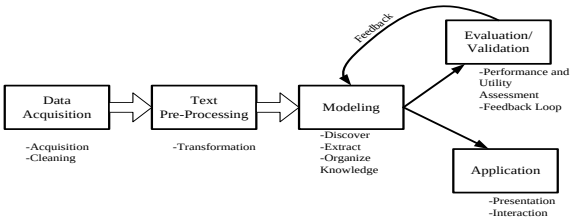

Fig 1: Text Mining Process

One of the typical applications of text-mining process is to read the entire text data that is composed in natural language and then it is subjected to classification process [11]. Usually such classification process is of predictive nature. However, there are various underlying challenges that motivate various research communities to continue investigation. Although, the applications of text mining is frequently found in research papers or various books, but at present there is such commercial application where such frequent usage of text-mining is known to users. The paper has surveyed and presented the limitations of various research techniques which are associated with the design of robustness and computational efficiency of Text Mining. Section 2 discusses about the significant issues of text mining followed by discuss of need of study in Section 3. The discussion of recent work being carried out is done in Section 4 followed by discussion of existing system in Section 4. Section 5 discusses about the frequently used techniques of text mining found in research papers while Section 6 discusses about the research gap. Finally, the summarization of the paper is done in conclusion in Section 7.

2. ISSUES IN TEXT MINING
The area of text-mining has got various advantages as well as it is also shrouded with various significant issues. The prominent challenges encountered at the time of data aggregation process in text mining are as follows:

2.1 Issues in Personalization

Majority of the existing applications of the text-mining is still dependent on the user's skills of identifying the knowledge. Hence, there is a strong dependency of usage of such tools based on the skills and expertism of a user. However, for making future usage generalized, there is still a need of intensive research work towards processing the queries using natural languages and involuntary processing of textual data (i.e. analysis). Hence, the tool should be highly customizable based on the need of the user, which at present is quite few to found.

2.2 Issues in Recognizing the Domain

It is extremely important for the text mining tool to be designed based on the domain knowledge. For an example, the term positive bears different context in medical domain as well as in mathematical domain. However, at present majority of the existing tools of text mining lacks any prior configuration of domain knowledge. It is of higher significance in text refining stage as well as in processing stage. Effective implementation of the text mining tool to recognize the domain could highly assist in performing parsing operation and can increase the potential of data filtering process.

2.3 Issues in Multiple Linguistic Supportability

Interestingly, the concept of datamining is completely free from any language, however the opposite of this is found true in text mining. Developing a function or an algorithm to perform refinement of multiple refinement techniques and generating methods free from language is highly challenging task in text-mining operation. At present, bulk of research work has been focused on English, Chinese, Japanese, Arabic, and various local languages of India. However, none of such works are standardized or found to be certified to be used in commercial markets.

2.4 Accurate Knowledge

It is quite a challenging issue to discover the accurate knowledge in text documents. At the same time, it is very much difficult to provide the user what they want from a text documents. Finding accurate knowledge or features is quite a challenging job in text mining. The previous researches on information retrieval provided many term based on protocols and the computational performance. In the term-based protocol, efficiency is very much high but the only problem arises in the term-based protocol is polysemy and synonymy. Where polysemy means, a word that has multiple meanings and synonymy means multiple words have the same meaning.

2.5 Text Mining in Online Social Media

Another major issue in the field of social media is Text mining in online social media. Here they have faced problem of gaining access to the social media data like twitter, face book API. Challenges here are further compounded by further idioms and Meta languages.

2.6 Knowledge-Base Challenges

There is a higher dimensionality of rift between the actual text as well as anticipated knowledge from it. The dependency of natural language processor as well as various protocols poses a greater deal of obstruction while designing the method of knowledge extraction. Hence, such critical challenges in text-mining technique will be required to overcome in order to visualize an effective tool of data analytics.

3. NEED OF THE STUDY

Ontology is very much essential for the formalization of common information such as product services relationship of a business. Although various data mining techniques have been introduced but they were not as much robust and computationally efficient so for the improvement of these methods in the field of data mining. In the area of text mining, a frequent occurrence of the text data that bears similar meaning. Such facts creates a huge problem in a particular document. Vector dimensions of suitable keywords are considered in the time of computation. Matrix computation is used for the feature extraction. So the research has been introduced to focus on finding the robustness and the computation efficiency of a clustering algorithm. When a user is searching for a particular topic from the documents, so the algorithm should give the relevant results with the less computation time.

4. RECENT STUDIES

This section discusses the recent survey studies which have been done to find out the technologies that are associated with the designing of robust and computationally efficient algorithm for text mining. The study based on the quality based web information retrieving with the use of natural language processing and text mining have been introduced as for discovering any information, now days people always prefer to use Google. The bar graphs they have presented to show the analysis they have done only for first ten web pages with two strategies. Where one is performed for Google relevant search, another is performed for quality based web information extraction using natural language processing and text mining [12]. Following are some of the significant review papers in text-mining published till date,

1. **Gupta and Lehal [13]**: The authors have presented a review of various techniques on text-mining. Following are the respective inferences:
    a. Pros: Discusses good theoretical details about the operations and the applications of Text Data Mining
    b. Cons: There was no discussion on performance of techniques.
2. **Clifton et al. [14]**: The authors have performs review of various security techniques on datamining.
    a. Pros: Performance measurements as well as various Methods for the domain study have been done in depth.
    b. Cons: Extremely narrowed discussion about the prior techniques.
3. **Clifton and Zhai. [15]**: The authors have investigated about the various clustering mechanism involved in text-mining approach
    a. Pros: Detailed discussion about the domain study.
    b. Cons: Very less discussion about the performance measurement.
4. **Sagayam et al. [16]**: The authors have performed an comprehensive survey on fundamentals of retrieval system of text, extraction process, as well as various indexing mechanism in text-mining.

a. Pros: Good Theory on domain.

b. Cons: Narrowed discussion of the prior techniques.

5. **Korde [17]**: The author has focused on reviewing the work done on classification techniques and various forms of available classifier design used in text-mining.

   a. Pros: Very good discussion about the domain.

   b. Cons: Comparative Observation is very few.

6. **Jensi and Jiji [18]**: The authors have carried out the review on various forms optimization techniques applicable on text-mining. Particularly, the work has reviewed various clustering techniques on documents.

   a. Pros: Good theory about the prior techniques.

   b. Cons: No comparisons between the prior techniques.

7. **Agarwal and Batra [19]**: The authors have discussed about various techniques involved in enriching the performance of text-mining.

   a. Pros: Detailed study about the domain

   b. Cons: Extremely narrow discussion about approaches towards text mining.

8. **Irfan et al. [20]**: This is the recent work done on reviewing the techniques involved in text-mining with special focus on data gather from Social Networks

   a. Pros: Good Theory about the text mining.

   b. Cons: Similar and repetitive discussion.

9. **Saranya and Munisweri [21]**: The authors have presented all the updated techniques that are used for enriching the quality of clustering method in text mining.

   a. Pros: Good study of clustering approaches.

   b. Cons: No discussion towards performance parameters..

Various researchers have also implemented various novel techniques in order to enrich the performance of the text-mining operations. Table 1 discusses about the existing techniques, performance parameters used and the limitations being observed in the implementation papers for text-mining.

Table 1 Techniques discussed in various studies

Authors	Techniques	Performance Parameters	Limitations
Zhong and Li [22]	Effective pattern discovery technique	Precision Recall	Improvement of accuracy is the one major point from pattern extraction.
Krisna and Bhabani [23]	Efficient techniques for text clustering which is based on frequent item sets, Apriori Algorithm.	Similarity measure	Apriori algorithm faces some issues related to limited memory capacity and for large item sets.
Mehta et al. [24]	An algorithm to identify the contradictory documents by clustering technique as well as optimization features.	Measurement of the Similarity of words in different sentences.	Effective feature selection
Ampofo et al. [25]	Focuses on some issues such as problem of gaining access to the social media data.	Positive sentiments for party leader, Positive sentiments for Cameron.	Ethical questions raised by social enquiries Access challenges Accurate Knowledge
Roberts et al. [26]	Structural Topic Model	Time, Expected Topic Proportion	Subjectivity Clustering Dynamism scale , heterogeneity
Xu and Reynolds [27]	IBM SPSS text analytics	Mean, Standard Deviation, Sample size, Software, Human rater	Employment of one independent human rater Key Challenges
Feinerer [28]	V-corpus and Vector source for data import. Dictionary with a character vector.	Standard operators and functions	Use of vector corpus For multiple files which are resided on disk Outcome of the performance analysis cannot be comparable with other studies.
Tseng et al.[29]	Text segmentation Summary extraction Feature selection Term association Cluster generation Topic identification And Information Mapping	Term Covering Rates for M-based terms. Topic map based on term clustering	Subjectivity Space limitations Clustering challenges. Knowledge based
Segev and Miesch [30].	Systematic and structured techniques	Prior differences of negative and positive sentences between many countries. Negativity and positivity of various sentiments.	Vastness, Vagueness, Uncertainty

5. METHODS OF TEXT MINING

This section introduces various methods observed for carrying out text-mining operations.

5.1 Mining Plain Text

The textual contents appears in applications in various format e.g. plain format, rich text format, formats that supports in web-scripts (e.g. JSP, HTML etc). However, plain text consists of semi-structured text contents. At present, the existing techniques of mining plain texts are :- a) text-summarization, b) document-retrieval, c) Information-retrieval, d) Assessing document similarity and e) Text categorization.

5.2 Summarization of Text

This mechanism performs automatic summarization of a given inputs in terms of specific paragraph or in the form of specific page. Output of text- compression algorithms are positively not human-readable, and it is also not actionable, it only supports decompression that is automatic reconstruction of the inventive text. Summarization varies from several new forms of text mining in that there are people, namely proficient abstractors, who are accomplished in the art of creating summaries and carry out the task as part of their professional life.

5.3 Document Retrieval

Document retrieval is the task of recognizes and returning for the most part significant documents. Traditional libraries provide catalogues that allow users to recognize documents stand on resources that consist of metadata. Metadata which is very efficient and well-structured document for summary, where various successful methodologies have been introduced for manual extraction of metadata and to identify the appropriate documents based on it. Availability of the libraries also supports various techniques of document retrieval system. However, document retrieval system, still have challenges lies in abstraction of significant portion of the document e.g. key words, authors, language, pertinent subjects etc.

5.4 Information Retrieval

For a given set of a text document, it is quite important to extract specific set of information. However, the process is highly dependent on the context of the text rather than just extracting text. As the term context is associated with it, hence the outcome of any processes of information retrieval system is not single. It is a set of multiple outcomes with highest similarity matches.

5.5 Assessing Document Similarity

Many text mining problems involve assessing the similarity between different documents; for example, assigning documents to pre-defined categories and grouping documents into natural clusters. These are the basic problems in data mining too and have been a focus for research in text mining. Perhaps because the success of different techniques can be evaluated and compared exploiting standard, objective, measures of accomplishment.

5.6 Text Categorization

Text categorizes is another branch of text mining techniques that permits to tag various significant portions of the text. However, it still encounters challenges with multiple languages and usage of various scientific symbols and notations that it is not able to recognize. Automatic text categorization has many practical applications excluding indexing for a document reclamation. Automatically extracting metadata, word sense disambiguation by detecting the topics a document covers and organizing and maintaining large catalogs of Web resources. Adoption of classifier is another frequent trend in text categorization.

5.7 Wrapper Induction

Various kinds of the data that exists at present are in the format that bears standards of public domain. Such form of information deploys usage of a specific markup approach for representation of the specific data over internet. Unfortunately, in that case, it poses a greater deal of problem while using conventional HTML tags as data couldn't be extracted in that way involuntarily. XML language is used to overcome this problem. The parsing operation is carried out using online wrapper classes for investigating the structure of the page and draws the significant information of the text. Unfortunately, the process is not easy as structure of a normal page may vary to highest degree.

5.8 Document Clustering

Clustering technique has always been the sole part of text mining technique. It is a technique of analysis of cluster for the text-based documents. Normally for an effective operation descriptors are extracted and operation is carried out either online or offline. There are dual forms of algorithms of document clustering e.g. hierarchical and k-means clustering.

5.9 Authoring of Web Documents

Majority of the text data are on web and hence it is important that such web-based text document should be effectively authored. By this technique, the various context of the text on the web-based interface can be easily controlled and it also offers various forms of personalization-based services of data-analytics. At present personalization poses a bigger challenge owing to multiple language usage and performing ranking of web-pages is further more challenging owing to dependency of domain idea.

6. RESEARCH GAP

Following discussion are the research gap based on the review the proposed survey study:

6.1 Narrowed Survey

The paper has briefed all the existing review work that has been published during the year 2009 till date. It can be seen that majority of the review work done till date discusses about theory of text mining or else clustering techniques and many computationally efficient techniques for text data mining. Prevention measures of repeated nature in almost all the existing review papers. The existing review papers are significantly missing better classification of the studies and discussion of the outcomes with respect to the performance parameters. It is felt that comparative analyses discussion with an aid of performance parameters of the existing studies are very much critical as it assist the reader to understand the most efficient technique till date. Majority of the contents of the existing prior studies are very much repetitive in nature with other survey papers, and no significant outcomes could be derived after gathering or reading the survey papers.

6.2 Less benchmarked Studies

Majority of the existing studies [20] [21] [24] [25] [27] [28] [29] etc. are not benchmarked at all. Hence, it becomes a challenging situation to understand the reliable nature of the discussed research papers. Benchmarking assist the readers as well as the authors to prove that under what circumstances their outcomes are superior to someone's work. If such critical information is missing from the existing implementation, adoptions of such work are less likely to happen for future

researcher.

6.3 Repetitive Nature of Implementation

The reason why this paper has choosen to segregate the papers with respect to publishers based on their impact factor, as because usually standard publishers like IEEE, Springer, Elsevier, etc. are found to have better scrutinizing policies for the manuscript, which are not that much emphasized by other non-standard publishers.

7. CONCLUSION

With the increasing size of the data worldwide from various enterprise applications, it has become quite critical to perform data analysis over the massively growing data. Hence, text-mining offers greater deal of flexibility for an organization to extract latent information from teh massively growing textual data and thereby evolve up with much better informative values. Although text mining has recently gained more attention from last 5 years, but still there are many issues in this research domain. This paper has reviewed some of the most significant work being carried out in the area of text-mining and studied some of teh most frequently used methods in text mining. Finally, the paper has explored the research gap as narrowed survey, less benchmarked studies, repetitive nature of implementation. The most frequently used data mining technique over massive and complex data are quite independent from any language, but this is not the case in text mining which is highly language dependent. The future work will be focused on developing a novel intelligent technique that can perform collaborative analysis of the multiple complicated texts. The future work will be also focused on designing a predictive tool that can extract the amount of sentiments from the textual data. Such applications can be designed by focusing on the construction of the sentiment analyzer and using semantics. The outcome could be checked for any form of the textual data with respect to false positives, accuracy and precision. The future idea is also perform comparative performance analysis of the outcome of the upcoming framework with existing one.

8. REFERENCES

[1] Miner, G.2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Books, Academic Press, Mathematics, pp.1053

[2] Weiss,S.M., Indurkhya, N., Zhang, T., Damerau, F.2010.Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science & Business Media, Computers, pp.237

[3] Han, J., Kamber, M., Pei, J.2011.Data Mining: Concepts and Techniques: Concepts and Techniques. Elsevier, Computers, pp.744

[4] Mehler, A., Kühnberger,K-Uwe., Lobin, H., Lüngen,H., Storrer,A., Witt,A.2011.Modeling, Learning, and Processing of Text-Technological Data Structures. Springer, Mathematics, pp.400

[5] Fiori, A.2014.Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding: Revolutionizing Knowledge Understanding.IGI Global, Computers, pp.363

[6] Moreno.J-M. T.2014.Automatic Text Summarization. John Wiley & Sons, Computers, pp.320, 2014

[7] Liu, B.2012.Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, Language Arts & Disciplines, pp.167

[8] Ahmad, K.2011.Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology. Springer Science & Business Media

[9] NissanE.2012.Computer Applications for Handling Legal Evidence, Police Investigation and Case Argumentation. Springer Science & Business Media, Social Science, pp.1340

[10] Karthikeyan,M.,Vyas,R.2014.PracticalChemoinformatics .Springer, 20Cheminformatics, pp. 33

[11] Berry,M.W., Kogan, J.2010.Text Mining: Applications and Theory.John Wiley & Sons, Mathematics, pp.222

[12] Ray, A. Kumar, and Kushwaha, A. (Retrieved, 2015). Quality based Web information extraction approach using NLP and Text Mining.

[13] Gupta, V., and Lehal, G.S.2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, Vol. 1, No. 1, pp. 60-76

[14] Clifton, P., Lee, V., Smith, K., and Gayler, R.2010. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv: 1009.6119

[15] Charu, A., and Zhai, C.X.2012. A survey of text clustering algorithms. In Mining Text Data, pp. 77-128

[16] Sagayam, R., Srinivasan, S., and Roshni, S.2012. A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. International Journal of Computational Engineering Research, Vol. 2, No. 5

[17] Korde, V., and Mahender, N.C.2012. Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 3, No. 2, pp. 85-99.

[18] Jensi, R., and Jiji, W.G.2014. A Survey on optimization approaches to text document clustering. arXiv preprint arXiv: 1401.2229

[19] Agrawal, R., and Batra, M.2013. A detailed study on text mining techniques. International Journal of Soft Computing and Engineering (IJSCE) ISSN 2231-2307.

[20] Rizwanairfan, C., King, G., Nielgragesi, D A., W en, D.,. Khan, S., Sajjdada.2015. A Survey on Text Mining in Social Networks. The Knowledge Engineering Review, Vol. 00:0, pp.1–24

[21] Saranya, S., and Munieswari, R. (Retrieved, 2015). A Survey on Improving the Clustering Performance in Text Mining for Efficient Information Retrieval. International Journal of Engineering Trends and Technology (IJETT)– Vol, 8.

[22] Ning, Z., Li, Y., and Wu, S.2012. Effective pattern discovery for text mining. Knowledge and Data Engineering, IEEE Transactions, Vol. 24, No. 1, pp. 30-44.

[23] Murali, K.S., and Bhavani, S.D.2010. An efficient approach for text clustering based on frequent itemsets. European Journal of Scientific Research, Vol.42, No. 3, pp.399-410.

[24] Mehta, R., Sankarasubramaniam, B., and Rajalakshmi, S.2012. An algorithm for fuzzy-based sentence-level document clustering for micro-level contradiction analysis. In Proceedings of the International Conference on Advances in Computing, Communications and

Informatics, pp. 102-105

[25] Ampofo, L., Collister, S., B.O'Loughlin, and Chadwick, A. (Retrieved, 20th July, 2015). Text Mining and Social Media: When Quantitative Meets Qualitative, and Software Meets Humans. Text Mining and Social Media: When Quantitative Meets Qualitative, and Software Meets Humans

[26] Roberts, M. E., Brandon M. S., Dustin T., Christopher L., Jetson, L.L, Gadarian, S.K., Albertson, B., and David G. Rand.2014. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science

[27] Yuejin, X., and Reynolds, N.2011. Using text mining techniques to analyze students' written responses to a teacher leadership dilemma. In Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology, China, Vol. 4, pp. 93-97

[28] Feinerer, I.2014. Introduction to the tm Package Text Mining in R. nd): n. pag. Web

[29] Y-H. Tseng. Lin, C-J., and Lin, Y-I.2007. Text mining techniques for patent analysis. Information Processing & Management, Vol. 43, No. 5, pp. 1216-1247.

[30] Elad, S., and Miesch, R.2011. A systematic procedure for detecting news biases: The case of Israel in European news sites. International Journal of Communication, Vol. 5.