ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

A Conceptual Based Approach in Text Mining: Techniques and Applications

A.Bharath

Assistant Professor, Department of Computer Science Engineering, Guru Nanak Institute of Technology,

Ibrahimpatnam, Ranga Reddy, Telangana, India

ABSTRACT: Recently, thedevelopmentand application of digital data is increasing in various field, knowledge discovery and text mining have concerned great consideration with upcomingrequirement 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 areseveralmethods and tools to mine the text and determineappreciated information for futureexpectation and decision making process. The assortment of accurate and correct text mining methodassists 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 techniques and their applications in different fields of life.

I INTRODUCTION

Text mining is defined as the non-trivial withdrawal of concealed, formerlynameless, and possibly useful information from literal data. Text Mining is a novelarea that attempts to extract eloquent information from normal language text. It can be demarcated as the process of examining text to retrieve information that is valuable for a particular purpose. Associated with the type of data stored in databases, text is formless, vague, and difficult to process. Nevertheless, in modern culture, text is the most communal way for the formal exchange of information. Text mining typicallypacts with texts whose function is the communication of definite information or feelings, and the incentives for demanding to mine information from such text instinctively is charming - even if success is only partial.Text mining is a process to extract inspiring and significantpatterns to discoverinformation from textual data bases. Text mining is a multi-disciplinary field based on information recovery, data mining, machine learning, figures, and computational semantics. Numerous text mining techniques like summarization, classification,clustering etc., can be applied to excerpt knowledge.Text mining agrees with natural language text which is depositedin semi-structured and unstructured format. Text miningtechniques are incessantly applied in industry, academia, webapplications, internet and other fields. Application areas like search tools, clientrelationship management system, clean emails, product suggestionenquiry, fraud discovery, and social media analytics use text mining for view mining, feature extraction, emotion, predictive, and trend analysis. Currently most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find stimulating patterns from large databases. The great deal of studies completed on the modeling and application of semi structured data in current database research. On the basis of these studies information retrieval methods such as text indexing methods have been established to handle unstructured documents. In customary search the user is usually look for previously known terms and has been written by somebody else. The problematic thing is in consequence as it is not relevant to users need. This is the objective of text mining to discover unidentified information which is not recognized and yet not printed down.

In this system, Text mining methodbegins with a document assortment from several resources. Text mining tool would improve a specific document and process it by examining format and eccentric sets. Text analysis is an analysis to descend high quality information from text. Many text analysis methods are accessible; subject to theaim of organization arrangements of systems could be used. Occasionally text analysis techniques are constant until

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

information is removed. The subsequent information can be located in a management information system, resilient aplentiful amount of knowledge for the consumer of that system.Extraction of cherished information from a quantity of different document is a boring and annoying task. The collection ofsuitable technique for mining text decreases the time and determinationto find the pertinent patterns for investigation and decision production. The aim of this paper is to examine different text mining systems which help to achieve text analytics commendably and professionally from large amount of data. Furthermore, the problems that arise during text mining process are recognized.

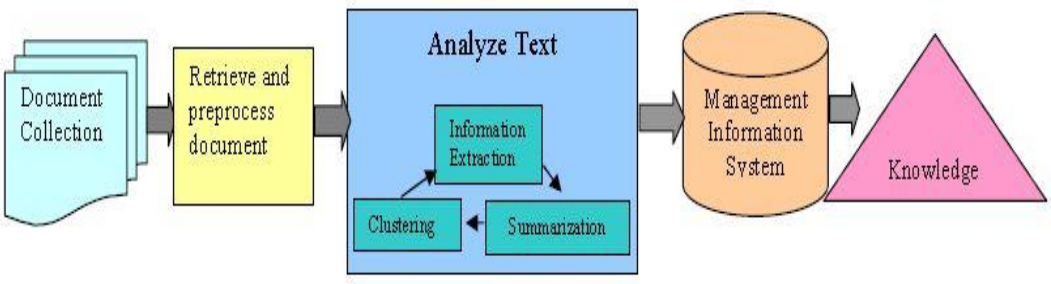

Fig.1. Text mining process

II. LITERATURE REVIEW

NaoyaOtsuka et.al(2015) has carried out theconstructing knowledge using exploratory text mining [3].Their goal is to support users who want to discover or create knowledge from a large amount of text data. Text mining is a process that extracts novel knowledge from unstructured data. A variety of applications for text mining, such as Total Environment for Text Data Mining, have been proposed. However, to the best of our knowledge, these methods are not effective at conducting text mining tasks aimed at finding novel knowledge. In this paper, we discuss characteristics of the text mining process and propose a design principle for building text mining applications based on two concepts: (i) text mining is an exploratory search task, and (ii) text mining is a process for creative knowledge work.T. Nasukawa et.al (2001) made an attempt in Text analysis and knowledge mining system [5]. Large text databases potentially contain a great wealth of knowledge. However, text represents factual information (and information about the author's communicative intentions) in a complex, rich, and opaque manner. Consequently, unlike numerical and fixed field data, it cannot be analyzed by standard statistical data mining methods. Rely on human analysis, results in either huge workloads or the analysis of only a tiny fraction of the database. We are working on text mining technology to extract knowledge from very large amounts of textual data. Unlike information retrieval technology that allows a user to select documents that meet the user's requirements and interests, or document clustering technology that organizes documents, we focus on finding valuable patterns and rules in text that indicate trends and significant features about specific topics.

Yuefeng Li et.al (2015)proposed Relevance Feature Discovery for Text Mining [6]. It is a big challenge to guarantee the quality of discovered relevance features in text documents for describing user preferences because of large scale terms and data patterns. Most existing popular text mining and classification methods have adopted term-based approaches. However, they have all suffered from the problems of polysemy and synonymy. Over the years, there has been often held the hypothesis that pattern-based methods should perform better than term-based ones in describing user preferences.It also classifies terms into categories and updates term weights based on their specificity and their distributions in patterns.Riccardo Scandariato et.al (2014) suggested Predicting Vulnerable Software Components via Text Mining [9]. This paper presents an approach based on machine learning to predict which components of a software application contain security vulnerabilities. The approach is based on text mining the source code of the components. Namely, each component is characterized as a series of terms contained in its source code, with the associated frequencies. These features are used to forecast whether each component is likely to contain vulnerabilities. In an exploratory validation with 20 Android applications, we discovered that a dependable prediction model can be

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

built. Such model could be useful to prioritize the validation activities, e.g., to identify the components needing special scrutiny.

III. TECHNIQUES USED IN TEXT MINING

Text mining is an important field that employstechniques and syndicatesapproaches from numerousareas such as

- Information Retrieval,
- Information Extraction,
- Text Categorization,
- Text Summarization And
- Text Clustering.

3.1 Information Retrieval

Information retrieval is a comparativelyancient research area. It extendedimprovedconsiderationwith the increase of the World Wide Web and the requirement for cultured search engines. The most recognized information retrieval (IR) systems are search tools such as Google which recognize those documents onthe web that are significant to a set of givenwords. The preprocessingexertion for the search engines is the informationabstraction process to generate order in a confusion of information. Google creep the web for statistics, understands it andprovisions in a detailed structure so that it can be rapidlyretrieved when users are firing search phrases. It is the task of attainingpertinent information from a collection of numerousresources. Document retrieval is restrained as apostponement of theinformation recovery where the documents that are repaidare processed to abbreviate or excerpt the specificinformation required by the user. Each operatorattempts to finddocuments that can give up information obligatory and stabs tocontent information needs for a specific user. The processof obtaining, classifying and probing the possibledocuments that may meet this information need is calledrecovery process.This system is used by manyuniversities, public libraries, government and companies toprovide access to articles, books, journals and otherdocuments.

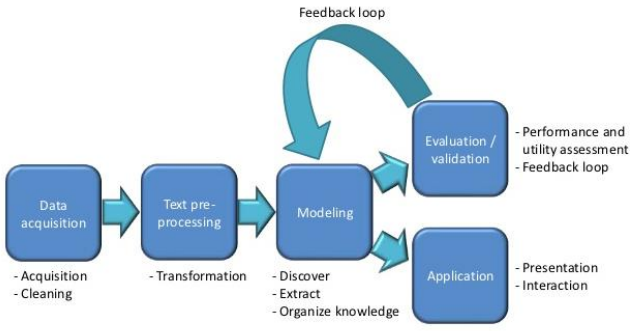

Fig.2:Information Retrieval

3.2 Information Extraction

Information extraction (IE) is the task of spontaneouslytaking outorganizeddetailed information from amorphousor semi-structured natural language. It identifies the extraction of units such as names of people, organization, location and relationship between articlesfeatures events and associations from text. The cherished information extracted is without proper sympathetic of text such as name of a person, association, position and sex. These are deposited in database like designs and are then obtainable for additional use. In most of the circumstances this actiondistresses processing human language texts by means of handling of natural text language. The information assembled is disciplined and deposited in a database automatically. IE renovates a quantity of textual information into a more organized database. Its difficulty of in use methods is contingenton the features of foundation texts. The major benefit

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

ofinformation extraction systems is the correctness of the enquiriesand the directness of the output. They can be professionallyrevised and then arrived into a file or showedvisually on screen.It is beneficial for a variation of applications chiefly given the recent propagation of internet and web document. Recent solicitations include job placement and possessions, medicalpatient records, weather information, seminars proclamations,course deference and apartment rental advertisements.

3.3 Text Summarization

The description of the instant is anoticeable one which accentuates the information that summarizing is in overall a hard task because we have to describe the source text as a whole and capture its significantcontented. The content is a substance of both information and its appearance and significance is a stock of what is vital as well as what is noticeable. Summarization is the process of dropping a text document with a computer program with the purpose of creatinganinstantaneous that conserves the most important points of the distinctive document. Technologies that can make anintelligible summary take into account variables such as length, lettering style and arrangement. Summarization systems are capable to generate both requestspertinenttext précises and general machine-generated synopsesdepending on what the user requirements. Summarization of multimedia forms, e.g. pictures or pictures is also possible. Some systems will create a summary based on a solebasis document, while others can use numeroussource forms. These arrangements are known as multi document summarization systems. Text summarization can be used to formulateinformation for use in trivial mobile devices, likeaPDA, which may require substantial reduction at ease.

3.4 Text Categorization

Categorization is the process of allocating a given text intogroups of objects whose associates are in some way relatedto each other. Appreciation of similarity across objectsand the followingcombination of like entities intogroups lead the separate to determine order in a difficultenvironment. Lacking the capability to group entities based onprofessedconnections, the individual's experience of any onething would be totally exceptional and could not be prolonged tosubsequent encounters with relatedobjects in theenvironment. This process is deliberated as a managedclassification method since a set of pre-classifieddocuments is delivered as a preparation set. The objective of this system is toallocate a grouping to a new document. By decreasing the loadon remembrance, easing the well-organized storage and recovery ofinformation, classification serves as the importantreasoning mechanism that shortens the individual'sknowledge of the environment.It can play asignificant role in aextensivevariation of areas suchas information reclamation, word sense disambiguation, topicrecognition and trailing, web pages classification, as well asany application requiring document organization. Automaticindexing of articles and by way of a controlledvocabulary such as the classification scheme (ACM) are the entries of the controlled vocabulary.

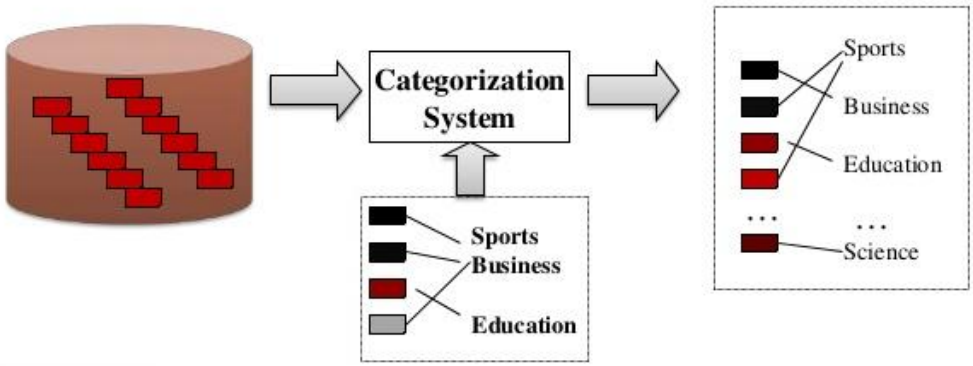

Fig.3 Text Categorization

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

3.5 Text Clustering

Clustering is a process of creating groups ofrelated objects from anassumed set of contributions. Worthy clustershave the distinctive that objects be appropriate to the samecluster are "like" to each other, while substances from twodissimilar clusters are "unlike". The indication of assemblinginitiates from information where it was smeared to numericaldata. Nevertheless, computer science and data mining inspecificstretched the conception to other types of data for instancetext or multimediaClustering is an unsubstantiated process through which objectsare categorized into groups called clusters. In the situation ofclustering, the difficult is to group the given unlabeledgroup into eloquent clusters without any previousinformation. Any labels allied with objects are acquiredmerely from the data. A benefit of clustering is thatdocuments can develop in numerous subtopics, thus confirmingthat a valuable document will not be inattentive from searchresults.Clustering is used in a wide range of systematic fields,applications and data analysis fields comprising data mining,document recovery, image separation and patternorganization.It is also used in instinctivedocument organization, topic extraction and fast informationretrieval or filtering.

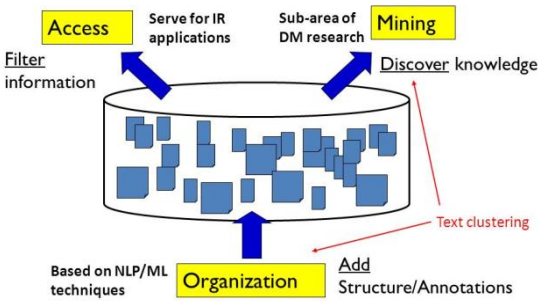

Fig.4 Text Clustering

IV. APPLICATION OF TEXT MINING

Some of the applications of text mining are given as follows as

❖ **Educational and Research Field**

In academic field, several text mining tools and techniques are used to evaluate the scholastic trends in particular region,pupil's interest in definite field and occupational ratio.Practice of text mining in research field helps to find and classify research papers and relevant material of different fields at oneplace.

❖ **Social Media**

In social media applications, Text mining software packages are available for investigatingand toobserve online plain text from internet news, blogs, email etc. Text mining tools support to recognize and evaluate number of posts, likes and followers on the social media network. This kind of explorationshows the general public reaction on different posts, news and how it spread around.

❖ **Digital Libraries**

Several text mining techniques and tools are in practice to determine the patterns and developments from journals and proceedingsfrom massivequantity of sources.Librariesare a great source of information for the academics and digitallibraries are attempting to the consequence of their collection.In this process,different operationare accomplished like documents selection, enhancement, extractinginformation and attempting entities among the documents andproducing instinctive co-referencing and summarization.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 1, January 2017

❖ **Life Science**

Life science and health care productions are creating largeamount of documented and numerical data concerning patient record,sicknesses, medications, indications and treatments of ailments. Text mining tools in biomedical field offers a prospect toextract appreciated information, their suggestion and deducingrelationship among various diseases, species, and genes.

❖ **Commercial Intelligence**

Text mining plays a substantial role in commercial intelligencethat assistsadministrations and enterprises to evaluate their consumersand contestants to take better resolutions.It alsohelps in telecommunication industry, business and commercesolicitations and client chain management system

V. CONCLUSION

In this proposed system, the awareness of text mining techniques have been familiarized and presented. Owing to its uniqueness, there are many prospective research areas in the field of Text Mining, which comprisesthe discovery of better intermediary forms for demonstrating the outputs of information extraction or reclamation. The concentration has been specified on important methods for directing text mining.Text miningtechniques are used to examine the stimulating and relatedinformation excellently and proficiently from large quantity offormless data. This paper offers a transitorysummary oftext mining techniques that assist to recover the text miningprocess.This paper enhances the process and uses of text mining, which can be applied in multitude areas such as web mining, medical, resume filtration, etc. It also clarify the unknown potential that deceits in the field of text mining and encouragedit.

REFERENCES

[1] Shilpa Dang, Peerzada Hamid Ahmad, "Text Mining: Techniques and its Application", International Journal of Engineering & Technology Innovations, ISSN (Online): 2348-0866, Volume 1, Issue 4, pp. 22-25, 2014.
[2] Shiqun Yin Yuhui Qiu1,ChengwenZhong, 2007. Web Information Extraction and Classification Method .IEEE
[3] NaoyaOtsuka, Mitsunori Matsushita (2014) "Constructing knowledge using exploratory text mining", IEEE International Symposium on Advanced Intelligent Systems (ISIS), 2015, page no: 1392 - 1397.
[4]Shah Neha K, "Introduction of Text mines and an Analysis of Text mining Techniques", PARIPEX, ISSN: 2250-1991, Volume 2, Issue-2, 2013.
[5] T. Nasukawa; T. Nagano, "Text analysis and knowledge mining system", IBM Systems Journal, 2001, Volume: 40, Issue: 4, Pages: 967 – 984.
[6]Yuefeng Li; AbdulmohsenAlgarni; Mubarak Albathan; Yan Shen; MochArifBijaksana, "Relevance Feature Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, 2015, Volume: 27, Issue: 6, Pages: 1656 – 1669.
[7] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.
[8] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
[9] Riccardo Scandariato; James Walden; Aram Hovsepyan; WouterJoosen, "Predicting Vulnerable Software Components via Text Mining", IEEE Transactions on Software Engineering, 2014, Volume: 40, Issue: 10, Pages: 993 – 1006.
[10]Amit Kumar Mondal and Dipak Kumar Maji, "Improved Algorithms for Keyword Extraction and Headline Generation from Unstructured Text", First Journal publication from SIMPLE groups, CLEAR Journal, 2013.
[11]Peerzada Hamid Ahmad, Shilpa Dang, "A Comparative Study on Text Mining Techniques", International Journal of Science and Research, ISSN: 2319-7064, Volume 3, Issue 12, pp. 2222-2226, 2014
[12]RashmiAgrawal and MridulaBatra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Volume 2, Issue-6, 2013
Varsha C. Pande and A.S. Khandelwal "A Survey of Different Text Mining Techniques", IBMRD's Journal of Management & Research, ISSN: 2348-5922, Volume 3, No. 1, pp. 125-133, 2014
[13]Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, pp. 141-160, 2003
[14] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006
[15] S. Niharika, V. SnehaLatha 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.
[16] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.
[17]Naveeta Mehta and Shilpa Dang, "A Review Of Clustering Techniques In Various Applications For Effective Data Mining", International Journal of Research in IT & Management, ISSN 2231-4334, Volume 1, Issue 2, pp. 50-66, 2011.
[18] N. Zhong, Y. Li, and S.T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.
[19] H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," International Journal of Computer Applications, vol. 75, no. 16, 2013
[20]K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," International Journal of Computer Applications, vol. 80, no. 4, 2013.