ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

A SURVEY ON TEXT AND KNOWLEDGE MINING FOR TEXT PROCESSING AND ITS DIFFERENT TECHNIQUES

[1] Priyadharshini. S.P, [2] Dr.M.Hemalatha,
[1] Ph.D Research Scholar, [2] Associate Professor
[1] Bharathiar University, [2] Sri Ramakrishna College of Arts & Science,
[1, 2] Coimbatore, India.

ABSTRACT- This survey is based on different text mining and preprocessing methods. Text mining is developed based on data mining tries to discover information hidden in scientific literature, which is not accessible by simple statistical techniques. Text mining techniques area significant subset of data mining that aims to extract knowledge from unstructured or semi-structured textual data and has widespread applications in analyzing and processing textual documents. Hence, combining textual mining techniques and bibliometric analysis can be exploited to help us discover more unseen patterns in research fields than simple bibliometric analysis. Pre-processing of documents that are from different sources is an important task during text mining process before applying any text mining technique.

Keywords: [Text mining, Text preprocessing, Text analysis, Knowledge Mining]

1. INTRODUCTION

Text mining techniques and tools are in use to ascertain the patterns and trends from journals and proceedings from immense amount of repositories. These sources of information help in the field of research and development. Libraries are a great source of information for the researchers and digital libraries are endeavoring to the significance of their collection. It provides a novel method of organizing information in such a way that make it possible to available trillions of documents online. It provides a novel way to organize information and make it possible to access millions of documents online.

2. LITERATURE SURVEY

Text mining is developed based on text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources . Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics. Green-stone international digital library that support multiple languages and multilingual interfaces provide a springy method for extracting documents that handle multiple formats, i.e., Microsoft word, pdf, postscript, HTML, scripting languages and e-mail messages. It also supports the document extraction in the form of audio visual and image format along with text documents. In text mining process various operation are performed like documents selection, enrichment, extracting information and tackling entities among the documents and generating instinctive co-referencing and summarization GATE, Net Owl and Aylien are frequently used tools for text mining in digital libraries. Text mining software packages are available for analyzing social media applications to monitor and analyze the online plain text from internet news, blogs, email etc.

Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and KhaledShaalan (2015) proposed text mining is developed based on the development in the fields of web, digital libraries, technical documentation, medical data has made it easier to access a larger amount of a textual documents, which come together to develop useful data resources. Therefore, it makes text mining (TM) or the knowledge discovery from textual databases a challenging task owing to meet the standards of the depth of natural language which is employed by most of the available documents. The available textual information in the form of databases and online source raises a question about who is responsible for keeping a check on the data and analyzing it. Keeping in view the pertaining condition, it is not possible to analyze and effectively extract the useful information manually. There is a need to employ software solutions which may employ automatic tools for analyzing a considerable amount of textual material, extract relevant data, analyze relevant data, and organize relevant information. Owing to the increasing demands to obtain knowledge from a large number of textual documents accessible on the web, text mining is gaining a significant importance in research.

BinlingNie andShouqian Sun (2015) proposed a Great design that has been the crucial point to make more aesthetically pleasing and more practical products and provide more adaptable services. Therefore, building a systematic view of design research becomes increasingly essential. Until now, analyzing the design research area from the perspective of scientometrics is still a largely unexplored area. To the best of our knowledge, there has been only one paper applying a quantitative approach to investigating the evolution and future trends of design research by analyzing citations of papers in the journal Design Studies. Additionally, few existing studies presented a qualitative method to analyze the design research area. Yet, there has been no in-depth study on keeping track of the current advances in the design research area.

M. Uma Maheswari , Dr. J. G. R. Sathiaseelan (2017) proposed text mining is developed based on text Mining is the disclosure of new obscure data, by consequently separating data from various documents. A key component is the connecting together of the separated data together to shape new actualities or new theories to be investigated by more ordinary methods for experimentation. Text mining is not quite the same as what are comfortable within web search. In pursuit, the client is regularly searching for something that is as of now known and has been composed by another person. The issue is pushing aside all the material that right now is not important to the requirements keeping in mind of the end goal to locate the applicable data. In text mining, the objective is to find obscure data, which something that nobody yet knows thus couldn't have however recorded. Text mining plays a significant role in business intelligence that helps organizations and enterprises to analyze their customers and competitors to take better decisions. It provides a deeper insight about business and gives information how to improve the customer satisfaction and gain competitive advantagesText mining algorithms will give us useful and structured data which can reduces time and cost. Hidden information in social network sites, bioinformatics and internet security etc. are identified using text mining is a major challenge in these fields.

Carlos A.S.J. Gulo and Thiago R.P.M. Rubio (2015) proposed text mining is developed based on text Mining is a common process of extracting relevant information using a set of documents. Text Mining provides basic preprocessing methods, such as identification, extraction of representative characteristics, and advanced operations as identifying complex patterns. Some achievements on text classification from various pieces of the literature. In general, text classification is a problem divided into nine steps. Those steps include data collection, text processing, data division, feature extraction, feature selection, data representation, classifier training, applying a classification model, and performance evaluationThe testing data will be used to validate the performance of the resulting classification model. There is no ideal ratio of training data to testing data. The text classification experiments presented have been used 25% for training and 75% for testing.

N. VenkataSailaja L. Padmasree, PhDN. Mangathayaru, PhD (2016) proposed text mining is developed based ontext Mining is the process of extracting fascinating information or knowledge or patterns from the unorganized text that are from diverse sources. As the text is in unorganized form, it is quite difficult to deal with it. Finding interesting information from the natural language text is the prime purpose of text mining. The text mining process is followed the three steps are as follows: Step-I: Pre-processing Text: Mining from a pre-processed document is easy as compare to natural languages documents. So, pre-processing of documents that are from different sources is an important task during text mining process before applying any text mining technique. As Text documents can be represented as - a bag of words on which different text mining methods. To reduce the dimensionally of the texts words, appropriate methods such as filtering and stemming are used. Filtering techniques remove those words from the set of all words that do not give relevant information; stop word filtering is a conventional filtering method. After this step is applied, every word is represented by its root word. Step II- Mining Technique: It is an important stage because in this step the selected algorithm is applied to text in order to process the document. The algorithm such as clustering, dimensionality reduction, knowledge representation and discovery, categorization, summarization, information extractions or visualizations could be used in general. Step III - Analysis of Text: For knowledge discovery purpose, outputs which are coming from initial stage are analyzed here. For this purpose, various tools such as link discovery tool can be used.

Yogapreethi .N, Maheswari .S(2016) proposed text mining is developed based on Text mining is to handle textual data. Textual data is unstructured, unclear and manipulation is difficult. Text mining is best method for information exchange. A non-traditional information retrieval strategy is used in text mining. For obtaining information from large set of textual documents which was done by the text mining. The figure1 is elaborated with the process of text mining. A Text mining and classification method has been used term-based approaches. The problems of polysemy and synonymy are one of the major issues. There was a hypothesis that pattern-based methods should outperform best compare to the term-based ones in describing user preferences. A large scale pattern remains a hard problem in text mining. The stateof-the-art term-based methods and the pattern based methods in proposed model which performs efficiently. In this work fclustering algorithm is used. Relevance feature discovery based on both positive and negative feedback for text mining models. The author focused towards the problem by classifying text documents on axiomatically, for the most part in English.. The techniques are clustering, classification, and information extraction and information visualization was overviewed. The process of text mining and the algorithms are also reviewed.

3. KNOWLEDGE MINING

Huda Umar Banuqitah , FathyEassa,Kamal Jambi, MaysoonAbulkhair, (2016) proposed knowledge mining is developed based on large spectrum data is being collected and generated on an unprecedented scale; this paradigm is called ―Big Data. Usage of biomedical computing systems present an explosive growth. Information Extraction (IE) techniques are the efficient exploitations of these resources that transform unstructured data into the structured form. An example of these techniques is Relation extraction (RE) which is an automatical mining of relations between the biomedical entities in text. The extraction of the relations between the biomedical entities is the procedure of determining the semantic link between those entities and characterizing the nature of this relationship. Recently RE techniques has found growing interest amongst IE community and many studies concentrate on it because it helps to find new relations and interaction between biomedical entities from raw text and minimize usage of a human resource. RE includes multiple techniques such as

Natural Language Processing (NLP), rule– based approach, and Machine Learning (ML) methods. There are three types of RE approaches which are: Supervised that uses a corpus of labeled data, Unsupervised method which needs no labeling, and Self-supervised (distant-supervised) that uses a small set of labeled examples. The Unsupervised technique extracts strings of words that exist between the entities in huge amounts of text, and then simplifies and clusters these word strings to produce relation. Unsupervised methods can use massive quantities of data and extract very large numbers of relationships, but the resulting relations may not be simple to map to relations needed for a particular knowledge base.

Smitha.T, Dr.V.Sundaram (2012) proposed knowledge mining is developed based on data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to predict trend analysis. Data mining can discover unexpected patterns that were not under consideration when the mining process started Prediction is a task of learning a pattern from examples and using the developed model to predict future values of the target variable. Many application can benefit the by the use of information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from a large database, information which is implicitly presented in the data, previously unknown and potentially useful for the users. One of the effective ways to create and use a data mining model is to get the user to actually understand what is going on so that an immediate action can take directly. There are many tools for analyzing the data. The clustering model allowed us to understand different group behavior for history of disease hit and accordingly take action. The knowledge extracted from the clustering model helped to identify the significant characteristics of insolvent inhabitants which formed a particular cluster. The supervised classification model was built on a data set. This model allowed predicting the insolvency of inhabitants well in advance so that the action measures can be taken against the insolvent inhabitants.

TipawanSilwattananusarn and Dr. KulthidaTuamsuk(2012) proposedknowledge mining is developed based on knowledge is becoming a crucial organizational resource that provides competitive advantage and giving rise to knowledge management (KM) initiatives. Many organizations have collected and stored vast amount of data. However, they are unable to discover valuable information hidden in the data by transforming these data into valuable and useful knowledge. Managing knowledge resources can be a challenge. Many organizations are employing information technology in knowledge management to aid creation, sharing, integration, and distribution of knowledge. “Knowledge management (KM) is an effort to increase useful knowledge within the organization. Ways to do this include encouraging communication, offering opportunities to learn, and promoting the sharing of appropriate knowledge artifacts”. Within the context of articles reviewed, applications of data mining have been widely used in various enterprises ranging from public health-care, construction industry, food company, retailing to finance. Each field can be supported by different data mining techniques which generally include classification, clustering, and dependency modeling. Knowledge is an important resource. Management of knowledge resources has become a strong demand for development. Discovering the useful knowledge has also significant approach for management and decision making.

Ying Cheng, Ken Chen, Hemeng Sun, Yongping Zhang, Fei Tao(2017) proposedknowledge mining is developed based on data mining appeared on the base of both the emergence of ultra-large-scale databases and the development of advanced ICTs. The DMTs is the techniques used in the data mining processes to search for the hidden information in a large amount of data. The main progresses of DMTs can be summarized as shown in Fig. 1. The development of data mining is mainly influenced by the applications of statistics, artificial intelligence, database technology, and machine learning. Most of theories and methods of DMTs are developed and extended based on the statistical theory. Artificial intelligence is used to generate the process of human thinking, which enables computer the ability of learning without precise programming and facilitates new techniques used in the data mining processes. Database technology provides the basis of data storage, organization and other functions for data mining data mining tools can be used to automatically discover interesting and valuable knowledge and patterns for the manufacturing processes. Then these knowledge and models can be used to promote the entire manufacturing processes in areas such as defect prevention and detection, quality improvement, flow time reduction, and so on.

R. Munilatha, K.Venkataramana (2014) proposed knowledge mining is developed based on Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web based applications. Web usage mining can also referred as automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or re-sources that are frequently accessed by groups of users with common needs or interests. Web structure mining helps the users to retrieve the relevant documents by analyzing the link structure of the Web. The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet.

HarunBayera , Mustafa Aksogana , EnesCelikb, AdilKondilogluc(2017) proposed knowledge mining is developed based on Big Data consists of data which is exponentially increasing in complex and non-structural forms in a never ending speed and these accumulating data has a profile that is far away from resolution option by conventional methods and techniques "Big Data" is a structure that analysing the conventional database management with analytic systems and defining the hard or impossible to resolve, rapidly growing, constantly flowing data sets data is regarded as a treasure which is acquired. It is very important that the data be increased interest in him. Agencies and organizations' need to take advantage of the data in problem solutions has become inevitable. As a result of the increase of interest in the data management, the questions of how to evaluate the data, how it can be functional, and where it will be stored are aroused. The most important feature in the structure of the data warehouse; for all systems within the company, the only source of information for reports is created. When we export a report through the system previously, we were connecting to multiple different systems and check each system's reporting accuracy. Companies will gain strength in this competitive environment after the correct analyses of social media communications, income statements, the relation between sales and advertisement budgets, office documents, stock market data, investment rates, and even the liquid instant data which is streaming in websites. The data streaming from numerous sources will worth golds if it is processed.

Hamid Mousav, Maurizio Atzor, Shi Gao, Carlo Zaniolo (2017) proposed knowledge mining is developed based on Mining Knowledge. To improve coverage of our integrated KB, we developed the IBminer and the OntoMiner systems, which generate InfoBoxes and ontologies from free text. IBminer employs an NLP-based text mining engine called SemScape to identify the morphological information in text and generate graph-based structures called TextGraphs. Then, IBminer extracts semantic links from TextGraphs using predefined patterns and converts semantic links to final InfoBoxes. In a fashion akin to

IBminer, OntoMiner uses graph pattern rules to mine iteratively ontological information from text. Integrating Knowledge. To integrate the knowledge from different sources into IKBStore, we use the existing interlink information from DBpedia. Thus, IBE was first used to address inconsistencies and other issues that surfaced during the integration of knowledge from different sources previously described, and it is currently used to manage our KB and upgrade it with information generated by crowdsourcing. Ease of access is achieved if the system can take users' queries expressed in natural language and translate them into SPARQL queries. As discussed in , good results can be obtained for simple short queries; in fact, voice input can also be used in very simple ambiguityfree requests. These approaches however cannot handle complex queries, and even for simple ones the risk of misinterpretation is high.

4. TEXT PREPROCESSING

Dr.S.Kannan, VairaprakashGurusamy,(2014) proposed text preprocessing is developed based on Text pre-processing is an essential part of any NLP system, since the characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages, from analysis and tagging components, such as morphological analyzers and part-ofspeech taggers, through applications, such as information retrieval and machine translation systems. It is a Collection of activities in which Text Documents are pre-processed. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help Text mining such as prepositions, articles, and pro-nouns can be eliminated Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens .The aim of the tokenization is the exploration of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining. These pre-processing techniques eliminates noisy from text data, later identifies the root word for actual words and reduces the size of the text data. This improves performance of the IR system Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis. Textual data is only a block of characters at the beginning. All processes in information retrieval require the words of the data set. Hence, the requirement for a parser is a tokenization of documents. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens, etc require processing as well. Furthermore, tokenizer can cater for consistency in the documents. The main use of tokenization is identifying the meaningful keywords. The inconsistency can be different number and time formats. Another problem are abbreviations and acronyms which have to be transformed into a standard form.

Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya,(2014) proposed text preprocessing is developed based on Text mining is the process of seeking or extracting the useful information from the textual data. It is an exciting research area as it tries to discover knowledge from unstructured texts. It is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT). KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining process is same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-structured data sets such as emails HTML files and full text documents etc. Text Mining is used for finding the new, previously unidentified information from different written resources. Preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process. Text mining is the process of seeking or extracting the useful information from the textual data. It tries to find interesting patterns from large databases. It uses different pre-processing techniques likes stop words elimination and stemming. This paper has given complete information about the text mining preprocessing techniques, i.e. stop words elimination and stemming algorithms. We hope this paper will help the text mining researcher‟s community and they get good knowledge about various preprocessing techniques.

B. K. Poornima, D. Deenadayalan and A. Kangaiammal(2017) proposed text preprocessing is developed based on Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and Information Retrieval(IR). In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data. The preprocessing techniques such as Tokenization, Stop word removal, lowercase conversion and Stemming are used for the text documents. Stop words removal technique is to remove the stopwords such as prepositions, articles, pronouns, etc. (that does not give the meaning of the documents i.ethe, in, a, an, with). Removing stopwords reduces the dimensionality of term space. Tokenization is to identify the meaningful keywords. The inconsistency can be different number and time formats. Challenges in tokenization depend on the type of language. Languages such as English and French are referred to as space delimited as most of the words are separated from each other by white spaces. Tokenization is also affected by writing system and the typographical structure of the words. Another problem is abbreviations and acronyms which have to be transformed into a standard form. Lowercase conversion is for a word that appears exactly the same every time it appears. The purpose of this method used for all text content is that all text is in the form of lowercase and so easy to analyze the content. Stemming is used to identify the root/stem of a word. The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space. Methods to overcome the issues indicated from the literature are the need of the hour. However, the proposed method is mainly focusing to extract the superimposed text in the form of audio in a video and apply some preprocessing to reduce the error rate and to make it useful for content analysis. Though the entire text should be extracted from the video for the content analysis, the superimposed audio from video alone is converted to text using any one of the many existing tools such as Google Docs, Braina, Apple Dictation, Dragon Naturally Speaking, etc. This work uses Google Docs app. From Google Docs app, Google Voice Search, a Google product that allows users to use Google Search by Speaking on a mobile phone or computer is used. In Google Docs, a popular service for managing documents online. Google Docs supports typing, editing and formatting via voice commands. It also helps to capture ideas, compose a letter or even write the novel without touching the keyboard.

Pritam C. Gaigole, L. H. Patil, P.M Chaudhari(2013) proposed text preprocessing is developed based on the preprocessing phase of the study converts the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified. It is the process of incorporating a new document into an information retrieval system. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall). This phase is the most critical and complex process that leads to the representation of each document by a select set of index terms. The main objective of preprocessing is to obtain the key features or key terms from online news text documents and to enhance the relevancy between word and document and the relevancy between word and category The goal behind preprocessing is to represent each document as a feature vector, that is, to separate the text into individual words. In the proposed classifiers, the text documents are modeled as transactions. Choosing the keyword that is the feature selection process, is the main preprocessing step necessary for the indexing of documents. This step is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning, and discard the words that do not contribute to distinguishing between the documents. The goal of preprocessing is to reduce the number of features which was successfully met by the selected techniques. From the results it is clear that the removal of stop-words can expand words and enhance the discrimination degree between documents and can improve the system performance. TF/IDF, the most frequently used indexing technique is used to create the index file from the resulting terms.

S. SagarImambi, T.Sudha (2011) proposed text preprocessing is developed based on the raw data in the form of text files is collected from online repository pub med. The data is converted in to xml files and stored in Database. The architecture of Preprocessing stage is shown in the fig 3. The preprocessing usually includes converting xml documents into text document, removing stop word, performing word stemming. Stop words are very frequently used common words like 'and' are' 'this' e.t.c. They are not useful in classification of documents. So they must be removed. Word stemming removes suffixes and generate the stemmed words ex. Retrieval becomes retrie. We used Porter Stemmer algorithms for word stemming. The generated features are assigned weights using various weighting techniques. The feature selection algorithm conducts a search for best subset using valuation algorithm. The valuation algorithm is run on the dataset usually portioned into internal training and test set with different set of features removed from the data. Extracting relevant feature from the text files is called feature generation. The main goal of feature generation is to transform a document in to a list of relevant features or keywords. Feature generation methods are classified into two main classes. Filter methods and wrapper methods. Filter methods use an evolution function that depends on data and is independent of inductive algorithm.(Sima C et al 2006) Wrapper methods use inductive algorithms to estimate the value of given subset. The inductive algorithm induces a classifier which is useful in classifying future set. The classifier is mapping from the space of feature values to the set of class values.

Ravi Lourdusamy , Stanislaus Abraham (2018) proposed text preprocessing is developed based on text is a leading medium for exchange of information, be it with humans, from time immemorial and of late, with machines too. The aim of text-mining applications is to trace, retrieve and operate on data of relevant information efficiently from the volume of text which continues to increase exponentially and expeditiously. TM as knowledge discovery process, first made known by Fledman [1] is an extraction of previously unknown facts by mining information from textual sources. It uses techniques and procedures from various specialized areas such as statistics, machine learning (ML), data mining (DM), natural language processing (NLP), information retrieval (IR), and information extraction (IE). The following are some major pre-processing techniques: Tokenization, Stop Word Filtering, Parts-ofSpeech (POS) Tagging, Word Sense Disambiguation (WSD), Grammatical Parsing and Chunking, Lemmatization, Stemming, Text Summarization, and Term Frequency and Inverse Document Frequency (TF-IDF). pre-processing techniques and tools available for TM were taken for an in-depth study. Feature analysis with respect to the tools studied was carried out highlighting the various features of the tools text pre-processing techniques and the tools implementing these techniques that might be useful to the researchers in order to improve the performance of the various preprocessing techniques and to enhance the existing tools or to innovate new ones. Selection and use of right pre-processing techniques and tools according to the domain might help to make the text pre-processing easy and efficient.

Ms. Nikita P.Katariya , Prof. M. S. Chaudhari (2015) proposed text preprocessing is developed based on text mining is a new area of computer science which strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. As Text documents can be represented as bag of words on which different text mining methods are based. Let Ω be the set of documents & W= {w1, w2, wm} be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form.

CONCLUSION

This paper surveys the different text preprocessing and different text mining approaches. It also presents methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side-information or meta information, which may be used in order to improve the mining process. Pre-processing activities plays a vital role in the various applications. Therefore it is concluded that the domain specific applications are more proper for text mining.

REFERENCES

[1]. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and KhaledShaalan, "Using Text Mining Techniques for Extracting Information from Research Articles", © Springer International Publishing AG 2018 K. Shaalan et al. (eds.), Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, https://doi.org/10.1007/978-3-319-67056-0_18 373 published articles during the years 2015 through 2016

[2]. BinlingNie andShouqian Sun,"Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research", Received: 23 December 2016; Accepted: 11 April 2017; Published: 15 April 2017.

[3]. M. Uma Maheswari , Dr. J. G. R. Sathiaseelan,"Text Mining: Survey on Techniques and Applications", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391 Volume 6 Issue 6, June 2017 www.ijsr.net Licensed Under Creative Commons Attribution CC BY.

[4]. Carlos A.S.J. Gulo1,2 and Thiago R.P.M. R´ubio1,3,"Text Mining Scientific Articles using the R Language", publication at: https://www.researchgate.net/publication/272417176Conference Paper January 2015 DOI: 10.13140/RG.2.1.3676.3363.

[5]. N. VenkataSailaja,L. Padmasree, PhD and N. Mangathayaru, PhD,"Survey of Text Mining Techniques, Challenges and their Applications", International Journal of Computer Applications (0975 – 8887) Volume 146 – No.11, July 2016.

[6]. Yogapreethi.N ,Maheswari.S,"A REVIEW ON TEXT MINING IN DATA MINING, Challenges and their Applications", International Journal on Soft Computing (IJSC) Vol.7, No. 2/3, August 2016.

[7]. Huda Umar Banuqitah ,FathyEassa,Kamal Jambi, MaysoonAbulkhair**,** “Big Data Knowledge Mining, Challenges and their Applications”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 11, 2016.

[8]. Smitha.T, Dr.V.Sundaram**,** “Knowledge Discovery from Real Time Database using Data Mining Technique”, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012 ISSN 2250-3153.

[9]. TipawanSilwattananusarn and Assoc.Prof. Dr. KulthidaTuamsuk**,** “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.

[10]. Ying Cheng, Ken Chen, Hemeng Sun, Yongping Zhang, Fei Tao,“Data and knowledge mining with big data towards smart production”, publication at: https://www.researchgate.net/publication/319440302, Article September 2017 DOI: 10.1016/j.jii.2017.08.001.

[11]. R. Munilatha, K.Venkataramana,“A Study On Issues And Techniques Of Web Mining”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, May- 2014, pg. 331-341 © 2014, IJCSMC All Rights Reserved 331 Available Online at www.ijcsmc.com.

[12]. HarunBayera , Mustafa Aksogana , EnesCelikb, AdilKondilogluc,“Big data mining and business intelligence trends”, Journal of Asian Business Strategy Volume 7, Issue 1(2017): 23-33 http://aessweb.com/journal-detail.php?id=5006.

[13]. Hamid Mousav, Maurizio Atzor, Shi Gao, Carlo Zaniolo,“Text-Mining, Structured Queries, and Knowledge Management on Web Document Corpora”, SIGMOD Record, September 2014 (Vol. 43, No. 3).

[14]. Dr.S.Kannan, VairaprakashGurusamy,,“A Study On Issues And Techniques Of Web Mining”, publication at: https://www.researchgate.net/publication/273127322 , Conference Paper · October 2014

[15]. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya**,** “Preprocessing Techniques for Text Mining - An Overview”, International Journal of Computer Science & Communication Networks,Vol 5(1),7-16.

[16]. B. K. Poornima,D. Deenadayalan and A. Kangaiammal “Text Preprocessing on Extracted Text from Audio/Video using R”, International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017.

[17]. Pritam C. Gaigole , L. H. Patil , P.M Chaudhari“Preprocessing Techniques in Text Categorization”, National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013) Proceedings published by International Journal of Computer Applications® (IJCA).

[18]. S.SagarImambi, T.Sudha“pre-processing of medical documents and reducing Dimensionality Advanced”, Computing: An International Journal (ACIJ), Vol.2, No.5, September 2011 DOI : 10.5121/acij.2011.2502 15.

[19]. Ravi Lourdusamy , Stanislaus Abraham, “A Survey on Text Pre-processing Techniques and Tools”, International Journal of Computer Sciences and Engineering Open Access Survey Paper Volume-6, Special Issue-3, April 2018 E-ISSN: 2347-2693.

[20]. Ms. Nikita P.Katariya , Prof. M. S. Chaudhari, “Text Preprocessing For Text Mining Using Side Information”,International Journal of Computer Science and Mobile Applications, Vol.3 Issue. 1, January- 2015, pg. 01-05 ISSN: 2321-8363