A Survey on Text Mining Techniques

Conference Paper · March 2019

DOI: 10.1109/ICACCS.2019.8728547

CITATIONS
36

READS
4,290

3 authors, including:

Sayali Tandel
Pace University
3 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Abhishek Jamadar
Smt. Indira Gandhi College of Engineering
2 PUBLICATIONS **67** CITATIONS
SEE PROFILE

2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS 2019)

A Survey on Text Mining Techniques

Sayali Sunil Tandel
Department of Computer Engineering
Smt.Indira Gandhi College of Engineering
Navi Mumbai, India
sayalitandel@outlook.com

Abhishek Jamadar
Department of Computer Engineering
Smt.Indira Gandhi College of Engineering
Navi Mumbai, India
abhishekjamadar38@gmail.com

Siddharth Dudugu
Department of Computer Engineering
Smt.Indira Gandhi College of Engineering
Navi Mumbai, India
siddharthdudugu@gmail.com

***Abstract*—As there is fast growth in digital data collection techniques it has made way for large amount of data. Greater than 85% of present day data is comprised of unsaturated and unstructured data. Determining the definite patterns and trends to examine a textual data is biggest issue in text mining The various domains associated together in data mining are text mining, web mining, graph mining , and sequencing mining. The selection of proper and correct technique of text mining enhances the hustle and by lowering the period and struggle done to mine important information. Here, we talk about text data mining, various techniques of text data mining and also application of text data mining. Text data mining is used for obtaining stimulating and fascinating designs from the unsaturated texts which are derived from various sources. It changes words, phrases and sentences of an unstructured information into mathematical value linking with the saturated information in the database and analyses it with traditional data mining techniques. Information extraction, information retrieval, summarization, categorization and clustering are the different techniques of text mining.**

Keywords—Text mining, Techniques, Clustering, Pattern, Summarization.

I. INTRODUCTION

Day by day the size of data is multiplying and expanding at an aggressive and rampant rate. Since the data is considerably huge all the corporate firms, institutions and organisations store the data in the system, electronically. Now this large amount of data is stored and exchanged through web in the form of digital libraries and textual information like blogs and other social media platforms. Hence it is difficult to extract information by using orthodox data mining techniques since they are not able to handle textual data effectively. When there is enormous amount of data and all the important and necessary information is required, Text mining is used. Text mining is also called as data text mining. Text mining examines text which is in natural language in detail and then lexical patterns are detected to extract important information. Pattern extraction from text documents and arrangements of text documents are the key goals of text mining technique development. Here both the unstructured data and semi structured data can be used for text mining. Fig. 1 represents the steps for Text Mining which is stated as follows:

- Convert unstructured data into structured data by collecting data from sources like plain text, web pages and data files.
- Pre-processing and cleansing operations are implemented to detect and remove anomalies.
- Cleansing process helps reveal the true essence of text which is available and is implemented to eliminate word stemming which is process of recognizing a certain word root and data indexing recognize the patterns from the structured data.
- Examines and inspects the designs and patterns using text mining techniques.
- In Processing technique cleaning and formatting of data is done and after the Text mining techniques like clustering and algorithms are applied to arrange to text documents.
- Extracts and roots out the useful information from the text.

Digital libraries, Web mining, drug discovery, Clustering, Social media, detection of links between lifestyle and states of health, business intelligence are some of the applications of text mining. This paper has different segments. Segment II consists of previous contribution by notable authors. Segment III defines the techniques in text mining and segments IV explains the applications of text mining. Segment V surveys the concerns of text mining and segment VI gives the conclusion of this paper.

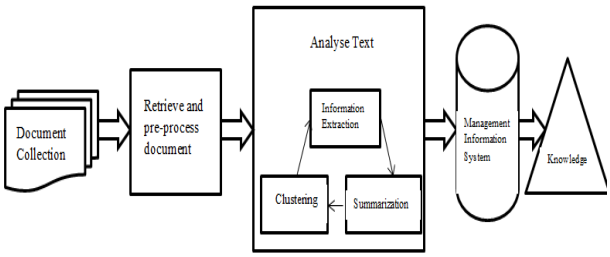

Fig.1. Example of a figure caption.

II. REVIEW OF LITERATURE

S.-H. Liao, P.-H. Chu et al (2012) surveyed the various types of techniques used for text mining, i.e., decision tree categorization, clustering, categorization which are used frequently, and their application in different fields. So they explained the part of text mining process. N. Zhong et al (2012) addressed text mining issues along with its

techniques. Further, they discussed that text mining tools and techniques get tough when traditional technique is used in unformed text. They concluded that using natural language processing and operation identification techniques can help in reducing the problems faced in the procedure of text mining. There are certain issues in text mining that still need to be addressed seriously.

A.Henriksson et al (2014) integrated a framework into MEDLINE biomedical database. The unnecessary details are eliminated and valuable information is extracted using this new framework. Laxman et al (2013) used text mining patterns for analyzing texts. Synonyms and polysemy are not analyzed properly by term based approaches. A prototype model was designed by them. This prototype model was used for arranging the patterns in various terms and conveying weights according to their distribution. Due to this approach the productivity of text mining was enhanced.

C. P. Chen et al (2014) compiled the relation discovery algorithm and presented a crime detection system. Rajendra along with Saransh (2013) presented a divisive and agglomerative method designed for text mining process which based on internet. The algorithm which they included is used in finding similarity within documents regarding specific subjects. K. Sumathy and M. Chidambaram (2013) summarized the applications, tools and issues that arise in text mining. A framework generally for concept based mining pictured as text enhancement and information extraction stages is discussed in this paper. Extracting useful information is a tedious task which may be in different forms was discussed. For any intermediate form of entity representation in text mining specific domains are responsible.

Miss. Poonam B. Kadam, Mrs.Latika R. Desai (2014), CAPTCHA images is a hybrid approach to detect and recognize texts in CAPTCHA images. The strength of CAPTCHA can be checked through this approach. It successfully detects and recognizes the text with a low false positive. Navathe et al (2000), the author proposed that the field of data mining has text mining as one of its variation. Data mining is a field wherein large databases are analyzed to find out various remarkable patterns. It is a method of mining remarkable and insignificant data and knowledge from amorphous text. V.Gupta, G.S Lehal discussed, that clustering is considered to be better if the documents are more comparable. Alike documents are grouped while using clustering technique but it is different from categorization.

III. TEXT MINING TECHNIQUES

A. Information Extraction

The primary stage for computers is to recognize amorphous typescript by recognizing important phrases and dealings within text. It aims at extracting meaningful information from huge chunk of text. The information mined is preserved in the form of a record for further access or recovery. This method identifies the extracted objects, attributes, and their relations from unamorphous texts or semi- amorphous.

In information extraction, the amorphous text files are firstly changed to structured database in which implementation of data mining technique is done so that the knowledge and interesting patterns can be extracted. The data in in relational database is stored in the form of tables within rows and columns. Required information is retrieved from structured query if the names of columns and tables are identified. But, unstructured data, the extraction is not easy like specific patterns from the text are very difficult since no link is available to locate the data like the one in structured data. Data with no structure contains information in small parts that can be created on information context and the purpose of its analysis. The practice may produce the different results subject to the purpose of the process and elements of textual data. The scope of information is defined by the elements of the textual data. These elements are tokens terms and separators can be termed as tokens. Characters like blank space or a punctuation mark are considered as separators. A token with specific semantic purpose can be a defined as a term. Various methods of extracting information can be token extraction, entity extraction, term parsing and complex fact extraction.

B. Information Retrieval

The field of Information retrieval has been under construction with database systems for more duration. The aim of Information Retrieval is to gain the document with precise information retrieved by the user. Thus recovery of document is followed by stage named as text mining which mainly focuses on request posted by the operator or it is followed by information extraction stage which uses information extraction techniques. Information retrieval, also known as IR, mainly focuses on searching and retrieving a document. Every search query has a response, in this response, out of collection of documents are retrieved. Here, the unsaturated textual data is present in large amount in a single document.

The document which is retrieved must be relevant information which is used by the user who performs queries. Database system focuses mainly on transaction and query processing of data which is structured, whereas information retrieval deals with processing of texts. Database system and Information retrieval handle the data of different types, the database problems are not there in information retrieval system like recovery, transaction management and concurrency control. And similarly, like some of problems are not easily found in database systems like the notion of relevance, and approximate search based on keywords. Information retrieval has found many applications, due to abundant of text information, the most recognized is search engine like Google, Yahoo, etc. that identifies the textual files found on the WWW that remains related to problems of the users [1]. Different methods in order are brought into use to facilitate this process using recognized search methods (Google).

Keywords based on IR technique plays an important role for enhancing the websites of various operations having ability to find files of interest. An IR system helps to contract the load of documents that are related to a specific issue. As text mining is used to apply most tough algorithms to the huge articles collections, the analysis speeds up by decreasing the count of documents for analysis using information retrieval.

C. Summarization

Summarization is collecting and producing concise

representation of documents with original text, this process is called as text summarization [2]. In summarization, first the raw text is taken and preprocessing and processing operations are performed on it. In pre-processing, three methods are applied i.e., tokenization stemming and word removal methods are applied. At handling stage of text summarization, generation of lexicon lists take place. The performance of automatic text summarization was influenced by rate of appearance of words or phrases in the last few years. Later to increase correctness of results some more methods were brought into practice with the standards procedure of text mining. Multiple documents can apply text summarization techniques at the same time. The subject of the documents depends on the quality and type of classifiers. Precise text is generated from number of documents in Summarization. It is not often possible to encapsulate huge textual file. [3]. Also, in centers used in for examining all the documents cannot be read. They basically summarize documents and make up the summary of document from important points.

Extractive and abstractive are two simple methods of summarization. In extractive method, a short version of text is formed. This is done by the appropriate selection of important sentences, and paragraph etc. from a textual document. In abstractive method, the entire document is understood. After understanding the document, natural language is used to express the concept of the entire document. Linguistic methods are used to describe the entire document. Also, it helps in delivering prominence of original document by creating a new text. Text summarization makes it easy to sort the document according to user and also checks if it should be considered for further information.

D. Clustering

The process of sectioning a group of objects or data into collection of relevant and understandable subclasses is termed as clustering. Clustering is mainly used to make a set of similar documents and files. The advantage of clustering is that the document or text files will be in multiple sub topics, which makes it safer for important documents from getting erased from search. Clustering technique separates records in a dataset into groups in such a way that themes in a cluster are same while themes between the clusters are different. Acquiring the group that has some value with regard to the difficulty being addressed is the main aim of cluster analysis. The result is not achieved always.

Therefore, the number of competing clustering algorithm is many. The analyst derives the cluster quality and compares it which is the main work of the analyst. He then chooses the cluster method which is the most relevant group. The clustering process classifies documents into non overlapping group. Each document comes into more than one topic area after classification. There are many methods of clustering and every method has different way of group data set. Clustering is mainly classified into two types:

i. Hierarchical

ii. Non hierarchical

Hierarchical clustering develops a cluster hierarchy, simply called a Dendogram, a tree like form of clusters. Each parent node cluster contains child cluster, the common parents cover the point partitioned by the relation cluster. The hierarchical clustering approach helps discovering data at complete different levels of granularity. The non-categorized method splits data set of x numbers of objects into y numbers of groups with overlapping. The methods are further separated in subdividing methods which contain categories that are equally special and contains less common stamping methods.

E. Categorization

In categorization, the important themes of a document are recognized. This is done by assigning the documents into a set of topics which are predefined. The document which is categorized it can be treated as ‘bag of words’. As information extraction does attempt to process the actual information whereas categorization doesn’t attempt to process the actual information. In this the words of the document are counted by categorization process Then using the counts they recognize the important subjects of the documents.

Categorization typically depends upon the thesaurus for which the topics are predefined and relationships are recognized by searching for comprehensive terms, narrow terms m, synonymous and related terms. The tool of categorization basically has a technique through which the ranking of the document is done by considering the documents which have most content on particular topic. Additionally, categorization is used with topic tracing and gives significance to the user looking for information on a topic. The data can be access by customers or end users readily if categorization schemes are used for classifying documents by topic. Each class belongs to more than one field. By using known examples, directed learning algorithm can be used. Also, classification on unknown examples is done automatically. Categorization associated with many application domains. The main aim of categorization is to categorize a collection of text into a fixed quantity of predefined groups.

IV. Applications in Text Mining

A. Digital Libraries

There are many techniques and tools of text mining which are used to set up patterns and trends from journals and its proceedings from massive amount of sources. For doing research, we can get great source of information for the research and the one making an effort to the significance are digital libraries. It offers interestingly new method of organizing information in such a way that it is available in trillions of documents online. It provides an innovative way to organize information and access it to millions of documents which are available online. Greenstone international digital libraries provides a springy method for extract documents of multiple formats such as MS Word, PDF, postscripts, HTML, script languages and email messages and it also supports multiple languages and multilingual interface. Various operations such as document selection, enrichment, information extraction and tackling articles in the middle of the documents are performed. The most frequently used tool for mining in digital libraries is Gate and Net Owl.

B. Web Mining

There a real many techniques and tools of text mining which are used to set up patterns and trends from journals and its proceedings from massive amount of sources. For doing research, we can get great sources of information for the research and one making an effort to the significance are digital libraries. It offers interesting new method of information organization in a way that it is available in millions of online web document. Greenstone International digital libraries provides springy methods of extracting documents of multiple format such as MS word, PDF, postscript, HTML, script languages and email messages and it also supports multiple languages and multiple lingual interface.

C. Clustering

Clustering is a process which is unsupervised to classify the text document in cluster by using various different clustering algorithms. Clustering is carried out in tip down and Bolton up approach behavior. In NLP, many mining tools and techniques are used just for the determination of unstructured data text. Clustering has various methods such as density, distribution, hierarchical, centroid and k mean.

D. Business Intelligence

Organizations and enterprises use text mining to analyse their customer and competition to take superior decision. It has a greater significance as it provides vision about the business and provides customer satisfaction and get more competition advantages. Text mining tools like such as IBM, text analysis, rapid mine, Gate supports to make conclusion about the institute that warns about the positive and negative execution.

E. Data Mining

Data Mining can be termed as an observation for the arrangements present in the data. It can be completely considered as the removal of secret, which is unidentified before, and beneficial material from the available data. Tools used in data mining can be used to calculate the behavior and forthcoming developments, and allows manufactures to create beneficial, awareness based decisions. Even Business questions can be answered using data mining tools but that might be time consuming to resolve. Complete aim of data mining method is to transform the extracted material from set of data into a logical arrangement for additional usage.

F. Social Media

Packages in Text mining software are able to access any social media applications and observe or learn the script from social media, blogs, internet, news or email etc. Text mining tools are very helpful in recognizing or analyzing total number of followers, post and likes on the social media. Due to this type of scrutiny it becomes very easy to find out peoples response on various news, post and how it expands socially. It may even show the behavior of different people who fit in to precise age crowd or people having similar opinions or variation about the similar post.

G. Resume Filtering

Large companies get thousands and lakhs of resumes every day through job seekers. Information is gained from resumes with very extraordinary accurateness and going through is very big task. Resumes can be written in a multitudinal formats Instead of creating a restricted domain, (e.g. graphs or simple text), in different languages (e.g. French and German) and in different file types (e.g. Word, EXCEL etc.). In addition, style of writing can be changed in diverse way. The recruiter looks for faults, qualification, history of the employee, fluctuations in various jobs, and individual information in the primary physical scan of the resume. Exactly gathering this data will be the first move in ignoring resumes. Thus, selection of resumes is an important task in the process of selection.

V. Issues in Text Mining

A.Henriksson et al (2016) mentioned numerous problems that take place during the process of text mining and how they affect the productivity and value of making any decision. At the early stage of text mining Complication can be surfaced. Various rules and guidelines are well-defined in preprocessing phase to control the text that makes The mining of text more efficient. Firstly unstructured data is needed to be converted into intermediate form before applying pattern analysis, but then it has its own drawbacks this stage in text mining process. Many times due to the change in the sequence in the text real subject of data may mislay its worth.

H. Solanki (2013) explored one more massive problem that it is bilingual text enhancement and its necessities. Multiple languages are supported by only few tools which are available.

A. Kumaran et al explained, to support multilingual text numerous techniques and algorithms are used self-reliantly. Many gears hesitate to support them because several important documents continue on the exterior part in the process of text mining. Such issues generate a lot of problem in discovering information and in the process of making decisions. In fact, actual advantage is quiet difficult to accomplish by using the text mining technique which is already in existence because it hardly support any multilingual documents.

Specific operations are performed on specific corpus and attain derived outcome by integrating domain knowledge, therefore it is an important area. In situations like this, domain knowledge from which amount of text to be mined needs to be combined with the calculating abilities through which information can be accomplished. Experts need to work together according to the requests of the field, and needed to work together from various domains that can mine more exact, effective and precise results [4], [5]. The texts can create confusion and complexity in the tools of text mining as they consider all these in similar contexts. It becomes hectic to segregate the textual documents as piles of documents are processed which belong to various categories but have the same domain.

Specific domain will be required to develop the plug-ins in deepness and suitable knowledge. [6], [7]. Natural language creates issues in text refinement methods and identifies entity relationship. Words have same spelling but they contribute different meanings, For example, tear and tear. Text mining tools considers both the words as same but they are distinguished as one is verb and other is noun. The field of text mining consists of grammatical rules according to your nature and perspective as it is still an undeveloped issue as [7].

A. Kaklauskas et al (2014) explained one of the biggest problems is abbreviations which gives changed meaning. Different notions of granularity alter the meaning of document which depends on condition and domain of knowledge. Defining norms can be used as a base in the area and is introduced in the tools of text mining as an attachment and is the necessity.

VI. Conclusion

Text mining techniques help in deriving different traits from amorphous textual data. Several methods and techniques lead to well-organized and accurate text mining. This paper is based on how mining should be performed on textual data. The process of text mining, its applications, Information retrieval, Summarization and various such methods have been discussed. A very convincing approach is discovered due to observations, because of which methods are examined and upgrading of method is suggested. In the long run, the technique which is proposed is implemented using machinery JAVA and very similar results are obtained. The different techniques which help in efficiently carrying out text mining are expressed as a part of this paper. Finally, it can be concluded that Knowledge-Discovery wherein the text is mined to get knowledge and information from text without a structure refers to text mining. More than 75% of data is text which means text mining has a huge market value. Information can be absorbed from different platforms (source) of material. But still unstructured data remains a source of maximum knowledge.

References

[1] R. Sagayam, S.Srinivasan, S. Roshni. 2012. ”A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques”, *International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, Issn 2250-3005(online)*.

[2] B. A. Mukhedkar, D. Sakhare, and R. Kumar. 2016. “Pragmatic analysis based document summarization” *International Journal of Computer Science and Information Security*, Vol. 14, no. 4, p. 145,.

[3] Falguni N. Patel, Neha R. Soni. 2012.” Text mining: A Brief survey”, *International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN(online):2277-7970)* ,Volume-2 Number-4 Issue-6.

[4] B. L. Narayana and S. P. Kumar. 2015 “A new clustering technique on text in sentence for text mining,” *IJSEAT*, Vol. 3, no. 3, pp. 69–71.

[5] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostrom. 2016. “Ensembles of ¨randomized trees using diverse distributed representations of clinical events,” *BMC Medical Informatics and Decision Making*, vol. 16, no. 2,p. 69.

[6] A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif. 2007. “Automatic extraction of synonymy information:-extended abstract,” *OTT06*, vol. 1, p. 55.

[7] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri. 2013. “Immune based feature selection for opinion mining,” *Proceedings of the World Congress on Engineering*, vol. 3, pp. 3–5.

[8] N. Zhong, Y. Li, and S.-T. Wu. 2012. “Effective pattern discovery for text mining,” *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44.

[9] E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernandez. 2013. “Searching research papers using clustering and text mining,” *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on. IEEE*, pp. 78–81.

[10] F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha. 2010. “Impact and usage of internet in education in pakistan,” *European Journal of Scientific Research*, vol. 47, no. 2, pp. 256–264.

[11] C. P. Chen and C.-Y. Zhang. 2014. “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” *Information Sciences*, vol. 275, pp. 314–347.

[12] V. Gupta and G. S. Lehal. 2009 .“A survey of text mining techniques and applications,” *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76.

[13] R. Steinberger. 2012. “A survey of methods to ease the development of highly multilingual text mining applications,” *Language Resources and Evaluation*, vol. 46, no. 2, pp. 155–176.

[14] R. Agrawal and M. Batra. 2013. “A detailed study on text mining techniques,” *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp.2231–2307.

[15] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan. 2004. “Text mining in a digital library,” *International Journal on Digital Libraries*, vol. 4, no. 1, pp. 56–59.

[16] R. Al-Hashemi. 2010. “Text summarization extraction system (tses) using extracted keywords.” *Int. Arab J. e-Technol.*, vol. 1, no. 4, pp. 164– 168.

[17] H. Solanki. 2013. “Comparative study of data mining tools and analysis with unified data mining theory,” *International Journal of Computer Applications*, vol. 75, no. 16.

[18] N. Padhy, D. Mishra, R. Panigrahi et al. 2012. “The survey of data mining applications and feature scope,” *arXiv preprint arXiv:1211.5723*.