TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

Web Text Mining through Machine Learning for Information Classification and Pattern Analysis: A Review

M. Venugopal, Dr. V. K. Sharma, Dr. Kalpana Sharma
Research Scholar,Dept.of CSE Professor,Dept.of EEEAssistant Professor, Dept.of CSE
Bhagwant University, Ajmer, India
venugopaltrue@gmail.comviren_krec@yahoo.comkalpanasharma56@gmail.com

Article Info
Volume 82
Page Number: 8063 - 8074
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 05 February 2020*

Abstract
In current years, there is tremendous development in information sharing and distribution on the web. In such the text data contributes the largest repositories. The main source of text data prominently comes from various information publications in multiple domains and social networks. Mining of such high volume of distributed information needs appropriate classification of the information. Due to which the mechanism of data mining is rapidly being utilizedto mine this various information to make it available for the application needs through web text mining (WTM) mechanism. However, WTMis facing many challenges to classify this information accurately due to the diversity of the information due to its context and semantic meaning. This paper aims to reviews the WTM andMachine Learning techniques to enhance the information classification through knowledge the various pattern analysis and its challenges. It intends to briefly review the importance of information classification in the field of Web text applications, especially for enterprise and social applications, and also to review existing techniques and methods for addressing the issue of extracting information from Web sources.

Keywords: *Web Text Mining, Machine Learning, Information Classification, Pattern Analysis.*

INTRODUCTION

The growth of electronic information specifically in the form of textual data on the Internet has grown tremendously over recent years. It contains very important and useful information resources, as well as a large amount of textual information in structured and unstructured form. Structured data is usually supervised by storage systems, but text data is usually supervised by search engines because of the deficient in structure form [1], [2].Search engines enable users to easily find useful information from collections through keyword queries, and various studies to suggest on how to advance search engine efficiency and effectiveness is an important research topic in the Information Retrieval (IR) [3], [4], [5].It also examines many related search topics forth clustering of text, information categorization, and recommendation methods [6].

The need to process text data automatically to retrieve beneficial information from a huge volume of data repositories [7], [8] to help human analysis is obvious. However, majority of distributed information is in the unstructured form and it is challenging to make its structure to extract information. Therefore, finding meaningful information from massive data is a huge challenge. In the past the processing and mining of text are performed by machine learning (ML)

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

techniques and data mining approaches through the knowledge of statistics and linguistic computing[9], [10], [11].The aim of Text Mining (TM) is to acquire most valuable information from text, and process to transforming unstructured text into structured data objects using a set of algorithms and a quantitative method for analyzing those data objects.

However, research on IR has typically encouraged the inclusion of more information than analyzing model information, which is the primary goal of TM.The purpose of access to information is to connect the precise information with the correct users at the precise time, with minimum focus on processing or converting text information.

The TMcan be viewed beyond the information access field, which helps users to analyze, digest, and make decisions in many TM applications. The main purpose to analyze and find appealing patterns in text data through the trends and avoid the dependency of query or relevant information for information mining. The basic purpose of TM is to help users extract data from text-based resources and supervise the operations for supervised or unsupervised retrieval and classification. It budding up integrates the technique of data mining with NLP and linguistics computing operations to mine the most web information over web [12], [13].

Most data mining analysts believe that "mining" information is already in the structure of a relational database.
Unfortunately, the digital information available in several applications appears in unstructured natural language documents instead of structured storage forms[14], [15], [16].For this reason, KDD is becoming increasingly important in the field of TM, and which is useful for discovering knowledge in structured or unstructured text data.

In the past, for IR many terminology-based methods were provided in WTM [17], [18] technology.The most term-based approach is proposed in a probabilistic, rough set, and SVM-based methods [19]. However, every one of these terminological processesundergoes from ambiguity problem, that is, each word hasseveral meanings and synonyms having an equivalent or related meaning. Thus it is difficult to answer the user'saspirationwith the semantic knowledge of the various terms of the applications.Technically, mining techniques are based on the foremost methods of these classes, algorithms, and applications able to discover from various types of text data.

The following papers are organized in different sections to introduce the insights of this review. In the current Section-2 WTM and its Process, Section-3 discusses ML in text categorization, and Section-4 introduces information classification and its challenges. In Section-5, it introduces the pattern analysis in WTM, Section-6 discusses several related works, and Section-7 introduces the future scope of WTM research. Finally, in Section 8, the conclusions of the review are presented.

I. WEB TEXT MINING

The Internet is a tremendous resource of information; it is so difficult to choose meaningful information for humans without assistance.It is the main source for information publication and accessing informationplatform.

A common template with content can automatically populate web pages on many websites to improvise the publication and retrieval efficiency [20]. For humans, even without explicitly declaring templates, templates make it easy for readers to access content that is guided by a consistent structure. However, when it comes to systems, the unfamiliar templates are pretendedto

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

be not useful because they minimizecorrectness and performance due to not having important terms in the template.So, template recognition and extraction procedures have recently pulled muchconsideration for improving the need for web-based application performance by means of classification and integration of documents [21], [22], [23], [24].For example, many organizations publish biogen data on the Internet in variousstructures,many researchers are trying to integrate this information to construct integrated storage, even various business data integration from multiple domains for sake of analysis or market prediction is also a need of WTM. So, to improve the performance of such application various web template data extractions methods are required.

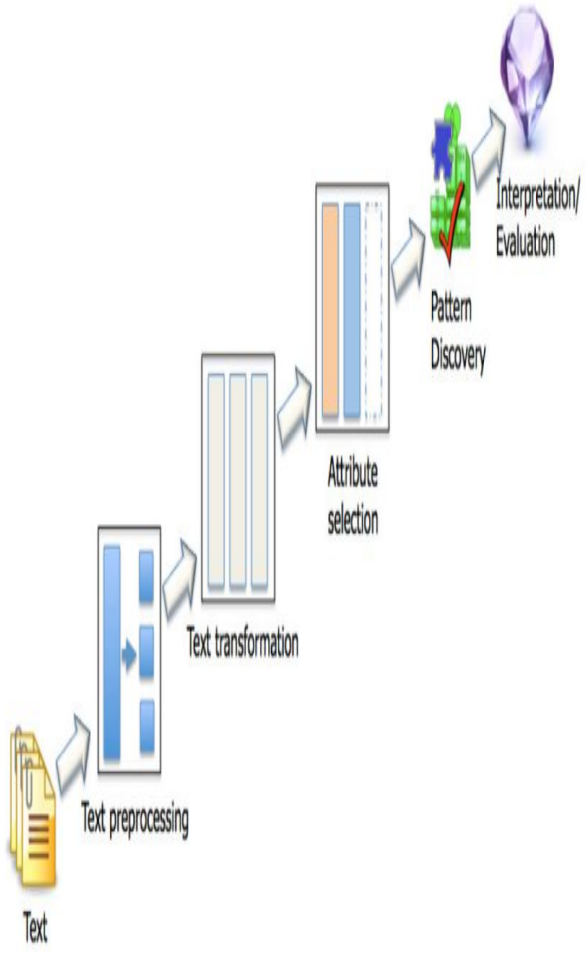

Fig. 1: WTM Methodology [Source: Internet]

The TM mechanism facilitates identifying significant knowledge from the text documents. It solves the problem of discovering novel, hidden and unidentified information automatically from collection of unstructured distributed text resources in the form of documents utilizing different Automated information discovery using sophisticated algorithms and techniques.

The TMis a branch of data mining that focuses on extracting information from text documents containing natural language text, while data mining extracts information from databases [25]. Although web mining also retrieves information from web documents that may also contain natural language text, there are significant differences between the two technologies of network mining and TM.The methods or procedures of TM extracts knowledge from structured and unstructuredweb documents [20], [26]. In addition, TM is dissimilar from IR or information access because IR does not retrieve any accurateand novel information like TM.

The TM process includes a set of sub-processes such as text preprocessing, text conversion, attribute selection, pattern discovery/data mining, and final evaluation or interpretation, as shown in Fig.1. A large amount of text data is pre-processed to clean text and to organize data in structured. After the text data is pre-processed, it is usually expressed as a word format in the form of a vector space representation or a word bag. This step is called text conversion. Then, the features of the various text data are selected in the attribute selection step. Unrelated attributes will be ignored and only the features of interest will be selected. Pattern discovery is a traditional data mining method because text data is available in a structured format.

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

II. MACHINE LEARNING IN TEXT CLASSIFICATION

The classification of a document automatically through various means is a significant topic for research in today's WTM. The different means of methodology to classify text document correctly utilizes the traditional ML, IR, and NLP techniques [27].

The abundance repositories of online text documents available give us a wealth of information. This wealth of information existing ought to be methodically prepared for its appropriate utilization. The organized and association of information facilitates the storage, search, and retrieval of relevant text content for utilization by various applications in need [28]. The practice of text categorization is essential for systematic categorization of text documents into the associated classes[29], [30]. The methodology of text categorization is significant in managing information documents precisely with low cost, error and processing time [31].

In many areas, automatic classification of text is an important issue due to the inaccuracy of the document text processing. The utilization of text processing is needed in many applications associated with the publishing, identification, classification, and indexing of the information in news, scientific, medical articles and also in other information in different domains.

ML is associated with the design and progress of mechanisms and procedures that enable systems to "learn" to enhance the anticipated predictions in the future[32].The text classification applications in IR utilize the algorithms which are weakly understood the association process due to the deficient in learning and generalization process [33]. The advantage of ML techniques provides a superior learning perspective toensuring better classification and IR by means of parameter setting for future development direction.

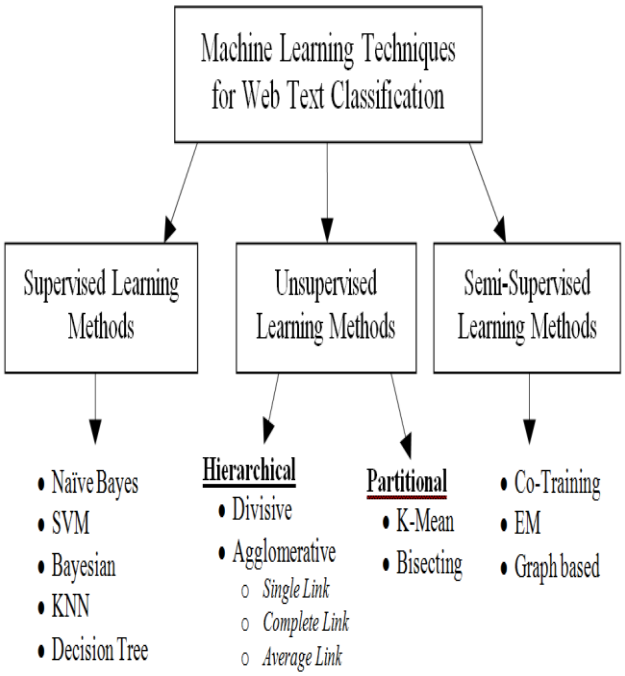

Fig.1 Classification distribution of ML Algorithms for WTM.

The main methods of ML are divided into “supervised learning”, “unsupervised learning” and “semi-supervised learning” [34]. More and more learning algorithms are including in day by day. The most frequent are "regression model", "nearest neighbor classification", "Bayesian probability method", "decision tree", "neural network", "online learning", "support" vector machine (SVM), "cooperative training", "maximum expectation", "graph-based approach", etc. [19].

The process of learning utilizes a set of training data which initially marked with a class or category, to support the classifiers to classify the new text documents.For all the categories with labels are believed as relevant and others as irrelevant. Based on the learned knowledge the text classifiers mostly compute the documents "weight vector" to predict the best classifying class it associated [35]. In Table-1 we present the most ML methods being used.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

Table-1: ML methods advantage and disadvantages

Methods	Description	Advantage	Disadvantages
K-Nearest Neighbor	This is a well-known pattern recognition algorithm. The algorithm is based on the assumption that the characteristics of members of the same category should be similar. Therefore, observations that are close together in the covariate space belong to the same class.	This method is effective, simple, parameter less and easy to implement.	The main disadvantage of this method is that it becomes slower as the training set size increases. The presence of irrelevant features severely reduces their accuracy.
Naïve Bayes	The Bayesian method that makes the independence assumption is called a naive Bayes classifier. It makes predictions by reading the examples in a set of attribute value representations and then using Bayes' theorem to estimate the posterior probability of all qualifications. The feature independence assumption makes the feature order uncorrelated, and the existence of one feature does not affect other features in the classification task.	This method requires a small amount of training data to estimate the parameters required for classification. Classifiers based on this algorithm show high accuracy and speed when applied to large databases.	This method works well only if the assumed functions are independent. When dependencies occur, performance decreases.
Decision Trees	Decision trees categorize educational documents by constructing well-defined true / false queries in the form of a tree structure. These leaves represent the corresponding categories of the text document, and the branches represent the combination of features that link to these categories.	This method works for all types of data. Even if a large amount of attributes is present, it is the fastest.	The main risk of implementing a decision tree is to align the training data with the occurrence of the replacement tree.
Rules-Based Classification	This method uses rule-based inference to classify documents into annotated categories. This classifier is useful for analyzing nonstandard data. Configure a set of rules that describe the profiles in each category. A rule is in the form of "Conditions following conclusions" in which the conditional condition is filled with the features of the category and the conditional part is represented by the name of the category or another rule to test.	This method can perform semantic analysis.	The main disadvantage of this method is the involvement of human experts to construct or update the ruleset.
Support Vector Machines	Statistics-based learning algorithm. This algorithm solves the general problem of learning to distinguish positive and negative members of n-dimensional vectors of a given class. It is based on the structural risk minimization principles of computer learning theory. SVM requires both positive and negative training sets that are not common in other classification methods. The performance of SVM classification does not change even if a document that does not belong to a support vector is removed from the set of training data. This is one of the main advantages.	Among the existing supervised learning algorithms for TC SVM, it was recognized as one of the most effective text classification methods because it can manage the functions of large space and high generalization ability.	However, this makes the SVM algorithm relatively complex, requiring high time and memory consumption during the training and classification phases.

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

In the text document processing the identification of correct knowledge to understand the user requirement is an unsolved problem. Initially, IR offers several term-based approaches to answer this problem, for example "Rocchio and Probabilistic Models" [36], "Rough Set Models", "BM25 and SVM based filtering models"[37], [38].The benefits of "term-based approaches "consist of proficient calculation enhancement and the well-established terminology weighting theory that has emerged from the IR and MLsocietyin the past few years.

However, the term-based approach has polysemy and synonym problems, where "polysemy" signifies that a word has several senses, while "synonyms" are several words with the identical sense [39]. The semantic sense of several of the terms found cannot determine the user's answer.

III. INFORMATION CLASSIFICATION

Information is vital to us in every way possible. We rely on information sources every day to accomplish a variety of tasks. However, the growth rate of information sources is alarming. In the past few decades, the concept of information and the consequent exchange of information have changed dramatically. However, as awareness grows and information grows accordingly, it is clear that we need to organize information in a way that makes it easier for everyone to access information [40].

With the fast growth of available digital data in recent times, "knowledge discovery" and "data mining" have drawn much interest to the requirement to turn such data into positive information and knowledge. A lot of applications, such as "market investigation" and "business administration", be able toassist from using information and knowledge pull out from huge amounts of data [41]. The discovery of knowledge can be considered as the process of implicitly extracting information from large databases, that is, information that is implicitly contained in the data and previously unknown and potentially useful to the user.

In the past decades, there has been a tremendous increase in information with the ability to connect between different parties. Thanks to the internet, everyone can now access almost endless sources of information via the web. As a result, the task of organizing this rich information becomes increasingly difficult every day. Had the other parties agreed to the structured web from scratch, it would have been much easier to classify the information correctly. But in fact, the information on the web is not organized or structured incorrectly. These facts have led to many attempts to classify information from the web and other sources, thereby establishing easier and systematic access to the information [42].

4.1 Challenges in Information Classification

The development of the World Wide Web has attracted many researchers to try to design ways to organize such a huge source of information. Scalability issues and the quality of automated organization and classification are all at work. Documents on the web have a variety of themes, their structures are different, and most of the structures are unreasonable. From very simple personal homepages to huge corporate websites, the nature of websites on the web varies, all of which contribute to building a vast repository of information.

Google, Yahoo! Search engines have been introduced to help tolocate relevant information on the web such as. However, the search engine does not automatically organize the documents, it just retrieves the relevant documents by the specific query issued by the user. Search engines are well known in the information search community, but they do not solve the problem of automatically organizing search documents.

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

The problem of classifying a lot of information into groups of similar topics is still unresolved. The real motivation of the research in this paper is to go one step towards a satisfactory solution and help solve this problem. Information is about creating a system that can effectively classify web documents based on a more informative document data representation and achieve a high level of classification quality. The learner is provided with numerous labeled training instances, and the main issue is to classify some test examples that were not previously seen.

IV. PATTERN ANALYSIS IN WTM

The various data mining techniques have been proposed that are practical models in text documents.However, finding and updating the identified patterns effectively remains an unresolved research topic, particularly in the TM field.The majority of the TM methods utilize a terminology-based approach, but all of them have an issue of ambiguity and synonymous setbacks.For many years, pattern-based or phrase-based methods have often been considered improved than term-based methods, however, several types of research do not sustain this supposition [43], [44].

Over the years, it has often been hypothesized that "phrase-based methods" may achievesuperior to"term-based methods" because phrases mighthaveadditionalsemantics information in compare to a term.This hypothesis is not well represented in IR history.Although sentences are less ambiguous and different from individual terms, they are less statistical, less frequent and contain a large number of repetitive and noisy phrases[45].

In light of these obstructions, the "sequential pattern" utilized in the data mining society has proven to be a promising substitute to this formulation[44], [46], because the "sequential pattern" has high-quality numerical characteristics related to terminology. In order to solve this issue the shortcomings of the "phrase-based method", a "pattern mining based method" or a pattern classification model [47] has been recommended. It adopts the concepts of "closed"or"non-closed"sequential pattern mode. These methods are based on "pattern mining" have revealed a degree of effectiveness improvement. Paradoxically, however, it is believed that a pattern-based approach may be an important alternative, but its effectiveness has not improved much compared to terminology-based approaches. Two basic questions about the validity of the pattern-based approach are due to "low frequency" and "misinterpretation."

With the past few years, IR hasbuilt up with a number of proven technologies that prove the term is an important feature in text documents. In many cases, the related document association is being classified based on the traditional "term frequency and inverse document frequency (TF-IDF) weighting method.

In the last decade, a variety of data mining techniques have been proposed for different knowledge work.These techniques consist of "association rule mining", "frequent itemset mining", "sequential pattern mining", and "closed pattern mining".Most of these methods propose the development of proficient mining algorithms to determine specific patterns within a realistic and satisfactory period frame.The problem of how to efficiently use and update these patterns still remains an unresolved research issue for designing a variety of data mining methods.

V. RELATED WORKS

In the past, various kinds of TM demonstrations have been proposed.A popular "word bag" is utilizedfor the keywords as factors in feature space vectors.In [48], the "TF-IDFweighting method" is discussed for TM illustration for the"Rocchio classifier". The"TF-IDF" is

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

considered as a global scheme proposed in [49], it illustrated a 30% improvised performance in the field of TM. Apart from this there are various schemes based on "word bag" notation are presented in [5], [40], [43].The difficulty with the "word bag approach" is how to choose an inadequate number of utility in a huge collection of words or terminology to enhance system competence and prevent overuse [41]. To reduce the number of features, several methods have been developed to decrease the dimensionality using "feature selection techniques", such as "information gain", "mutual information", "chi-square", " Odds ratio", and others [32], [44].

In [50], the analysis of the text is discussed utilizing the TM techniques through mining the co-occurrence terms from a collection of documents as descriptive phrases. But, the efficiency of TM systems based on phrases as textual representations has not improved significantly. As described in [37], the probable cause might be the"phrase-based method"have a lower uniformity of assignments and a minimum frequency of term in the documents.

The "term-based ontology" mining method also provides a few ideas for text representation. For instance, "hierarchical clustering" [51] is utilized to resolve synonyms and subordinate relationships among key terms. In addition, pattern development techniques were commenced in [52] to advance the efficiency of "term-based ontology mining".

In the field of TM, the studies of pattern mining are broadlybeing explored in the past years. In such few most popular algorithms are "Apriori-like algorithms" [53], [54], "PrefixSpan" [55], "FP-tree" [56], "SPADE" [57], "SLPMiner" [58] , and "GST" [59].These studies effort to focused on increasing the effectiveness of the mining algorithms to generate patterns in huge volume of data. However, generating the needed patterns and rules is still an open challenge in TM to construct representations using new data features[27], [34], [35], [40].

For challenging problems, closed-order patterns are utilized for TM in [43], [44], [47], it suggests that the closed-patternperception in TM is valuable and has the prospective to improve the efficiency of TM. Various pattern classification models are been discussed in [50] and [52] to progressthe efficiency bymeans of"closed patterns" in TM. Additionally, a "two-stage model" using together a "team-based approach" and a "pattern-based approach" was presented in [55], [56] to considerablyprogress the efficiency of information mining.

VI. FUTURE RESEARCH SCOPE

In response to the above problems, we aim to explore and propose solutions within the scope of WTM's future research by using ML techniques for information classification and pattern analysis:

7.1 Heterogeneous Information Learning and Pattern Analysis

In the trend of information analysis and development of many compelling search engines like Google, boost the TM researches and also the utilization of information in several applications of online society for analyzing the diverse data in multiple domains for detection, prediction and searching.The conventional data mining algorithms typically make an effort to discover patterns in data sets that containindependent and evenly distributed samples.

Many interesting areas of today's needs are illustrated by an interrelated heterogeneous network of objects [14], [16], [25]. This study will propose a link mining method that can aimat the combination of "link mining algorithms" for various informationfindingjobs. In addition, the analysis of various applications is dynamic and the development of incremental link mining

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

algorithms is important, and also to knowing "links", "objects" and "networks", it can build ontology and structured information networks based on structured information.

7.2 Information Mining in Heterogeneous and Unstructured Text Data

The web is common for users from different domains to post data, share surveillances, practices, and interchangethoughts. There is a big volume of information storage over the web. For example, "Wikipedia" has a huge amount of information repositories available on a variety of topics and domains [19], [26].As a result, web information has become the best platform for access information for the multiple domains, it provides a huge collection ofwebpage link access that contains"text" and"multimedia" data, but it also provides support for the query-enabled "databases" for deep web access.

Today TM is being applied in the web for IR over both structured and unstructured to simplifies the IR in many information-sharing systems related to "digital libraries", "biological information", "research articles", etc. The techniques utilize to extract and process information is through "conceptual linkage", "topic tracking", "categorization" and "clustering". To understand potential of information association among the IR it generates patterns to relate the domain concept and knowledge constructed in relevant to "Web content mining", "Web structure mining", and "Web usage mining" [14], [16].

There are many research issues in this area that require multidisciplinary efforts, together with"IR", "databases", "data mining", "NLP", and "ML".For various applications that present information, structured and semi-structured data are found, with specific factors for text and multimedia data.Therefore, it can mine and constructcomparatively structured web storage. The focus of this research is to present promising research methods for heterogeneous information integration, information extraction, and in-depth Web semantic analysis to support accurate information classification.

7.3 Heterogeneous Information Classification through Pattern Analysis

Most applications today typically handle high-dimensional and heterogeneous massive amounts of data. The objective of pattern mining is to discover the set of items that appear in the dataset. The frequency of the subsequence or substructure is not below the user-preciselimit. Pattern investigation is a importantmeans for discovering the "correlations", "clustering", "classification", "sequence and structural patterns", and "outliers".

For more than a decade, "frequent pattern mining"has been a key subject in data mining research [30]. A large part of the studies is devoted to this research and has made great progress, from proficient and extending algorithms for mining frequent itemset in operational repositories in the field of "sequential pattern mining", "structural pattern mining", "correlation mining", "association", and"classification".

Recently, research has been carried out to scalable methods to mine huge patterns [20], [31], where the size of the pattern may be quite large, so that it cannot be extended graduallyutilizing"Apriori-like methods", and these methods are used for pattern firmness, and for the mining of superior-top-k patterns [32]. But further research is still is required to considerablydecrease the length of thegenerating patterns, and effectively mining such patterns and improving the superiority of the retained patterns. In addition, the "classification", "correlation analysis" and "pattern understanding"for mining patterns remainsan interesting contribution to TM research in future.

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

VII. CONCLUSION

This paper aims to explorer thepaststudies and examines the TM and ML techniques and their respective strengths and weaknesses in WTM.In the majority of the studies, it is evident that the performance of the ML algorithm on text classification isaffected due the structure and quality data form.The characteristic illustration technique reduces the accuracy and performance of the classifier due to its unrelated and repetitive attributesof data.It suggests the need and challenge of using ML technology in WTM. It also discusses the challenges of information classification and pattern analysis to accurately extract information. It suggests that the future development of WTM involves heterogeneous information sets, which will become a necessary condition for making a more general text classification system that will effectively utilize a large amount of unlabeled data for classification of various systems.

VIII. REFERENCES

[1]. Z. Tan, C. He, Y. Fang, B. Ge, W. Xiao, "Title-Based Extraction of News Contents for Text Mining", IEEE Access, Vol. 6, 2018.

[2]. M. I.Varlamov and D.Y. Turdakov, "A survey of methods for the extraction of information from Web resources", Program. Comput. Softw., vol. 42, no. 5, pp. 279-291, 2016.

[3]. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. "Open information extraction from the Web". In Proc. of the 20th Int. Joint Conf. on Artificial Intelligence, pp. 2670-2676, 2007.

[4]. C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of Web information extraction systems", IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1411-1428, 2006.

[5]. D. Freitag, "Information extraction from HTML: Application of a general machine learning approach", In Proc. 15th Nat. Conf. Artif. Intell. 10th Innov. Appl. Artif. Intell. Conf. (AAAI), pp. 517-523, 1998.

[6]. Z. Gao, Y. Fan, C. Wu, W. Tan, and J. Zhang, "Service recommendation from the evolution of composition patterns", in Proc. IEEE Int. Conf. Services Comput. (SCC), pp. 108-115, 2017.

[7]. S. Wu, J. Liu, and J. Fan, "Automatic Web content extraction by combination of learning and grouping", in Proc. 24th Int. Conf. World Wide Web (WWW), pp. 1264-1274, 2015.

[8]. M. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender, "Automatic web news extraction using tree edit distance", In WWW, 2004.

[9]. M. Tang, Y. Xia, B. Tang, Y. Zhou, B. Cao, R. Hu, "Mining Collaboration Patterns Between APIs for Mashup Creation in Web of Things", IEEE Access, Vol. 7, 2019.

[10]. B. Liu, C. Wang, Y. Wang, K. Zhang, C. Wang, "Microblog Topic Mining Based on FR-DATM Chinese Journal of Electronics, Vol. 27(2), 2018.

[11]. Y. Zuo, J. Wu, H. Zhang, D. Wang, K. Xu, "Complementary Aspect-Based Opinion Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 30(2), 2018.

[12]. O. Shapira, H. Ronen, M. Adler, Y. Amsterdamer, J. Bar-Ilan, and I. Dagan, "Interactive abstractive summarization for event news tweets", in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), pp. 109-114, 2017.

[13]. D. Gibson, K. Punera, and A. Tomkins, "The volume and evolution of Web page templates", In Proc. 14th Int. Conf. World Wide Web (WWW), pp. 830-839, doi: 10.1145/1062745.1062763, 2005.

[14]. Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in large uncertain databases", IEEE Trans. Knowl. Data Eng., vol. 26, no. 5, pp. 1171-1184, 2014.

[15]. J. Wang and F. H. Lochovsky, "Data extraction and label assignment for Web databases", in Proc. 12th Int. World Wide Web Conf. (WWW), pp. 187-196, doi: 10.1145/775152.775179, 2003.

[16]. Y. Li, J. Bailey, L. Kulik, and J. Pei, "Mining probabilistic frequent spatiotemporal sequential patterns with gap constraints from uncertain databases", In Proc. IEEE ICDM'13, pp. 448-457, 2013.

[17]. T. -Y. Chan, Y.-S. Chang, "Enhancing Classification Effectiveness of Chinese News Based on Term Frequency", IEEE 7th International Symposium on Cloud and Service Computing (SC2), Pages: 124 - 131,2017.

[18]. P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic is a knowledge", In Proc. of the 22nd ACM International Conference on

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

Information and Know. Manag., pp. 1401-1410, 2013.

[19]. A. Khan, B. Baharudin, Lan Hong Lee, "A Review of Machine Learning Algorithms for Text- Documents Classification", Journal Of Advances in Information Technology, Vol. 1 , No. 1, 2010.

[20]. Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Relevance Feature Discovery for Text Mining", In IEEE Trans. Knowl. Data Eng., vol. 26, no. 6, pp., Jan. 2015.

[21]. A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages" In SIGMOD, 2003.

[22]. Z. Bar-Yossef and S. Rajagopalan, "Template detection via data mining and its applications" In WWW, 2002.

[23]. M. N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim, "Xtract: A system for extracting document type descriptors from xml documents", In SIGMOD, 2000.

[24]. D. Gibson, K. Punera, and A. Tomkins, "The volume and evolution of web page templates", In WWW, 2005.

[25]. T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefie, "Probabilistic frequent itemset mining In uncertain databases", In Proc. ACM SIGKDD'09, 119-128, 2009.

[26]. S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of Wikipedia entities in web text", In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457-466, 2009.

[27]. H. Ye, B. Cao, Z. Peng, T. Chen, Y. Wen, J. Liu, "Web Services Classification Based on Wide & Bi-LSTM Model", IEEE Access, Vol. 7, 2019.

[28]. Z. Zhang, Q. Li, and D. Zeng, "Mining evolutionary topic patterns in community question answering systems", IEEE Trans. Syst., Man, Cybern., vol. 41, no. 5, pp. 828-833, 2011.

[29]. G. Xu, Y. Meng, X. Qiu, Z. Yu, X. Wu, "Sentiment Analysis of Comment Texts Based on BiLSTM", IEEE Access, Vol. 7, 2019.

[30]. Y. Long, Y. Liu, Y. Zhang, X. Ba, J. Qin, "Coverless Information Hiding Method Based on Web Text", IEEE Access, Vol. 7, 2019.

[31]. E. Ferrara, P. D. Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey", Knowl.-Based Syst., vol. 70, pp. 301-323, 2014.

[32]. F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47, 2002.

[33]. A. Dhar, N. S. Dash, and K. Roy, "Application of TF-IDF feature for categorizing documents of online bangla Web text corpus", Intell. Eng. Inform., vol. 1, pp. 51-59, 2018.

[34]. M. K. Dalal and M. A. Zaveri, "Semi supervised learning based opinion summarization and classification for online product reviews", Appl. Computer Intelligence Soft Computing, 2013.

[35]. D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec", Inf. Sci., vol. 477, pp. 15-29, 2019.

[36]. W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding", In Proceedings of the 2012 ACM SIGMOD Int. Conf. on Management of Data, ser. SIGMOD, pp. 481-492, 2012.

[37]. S. Shehata, F. Karray, and M. Kamel, "A concept based model for enhancing text categorization", In Proc. ACM SIGKDD Knowl. Discovery Data Mining, pp. 629-637, 2007.

[38]. M. S. Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.

[39]. N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms", ACM Comput. Surv., vol. 43, no. 1, pp. 3:1-3:41, 2010.

[40]. M. Bruno, G. Canfora, M. Di Penta, and R. Scognamiglio, "An approach to support Web service classification and annotation", in Proc. IEEE Int. Conf. E-Technol., E-Commerce E-Service, pp. 138-143, 2005.

[41]. Y. He, Cheng Wang, Changjun Jiang, "Discovering Canonical Correlations between Topical and Topological Information in Document Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 30(3), 2018.

[42]. W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE Transactions on Knowledge and Data Engineering, Vol. 29(3), 2017.

[43]. M. Bouazizi, T. Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter", IEEE Access, Vol. 5, 2017.

[44]. S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text

January - February 2020
ISSN: 0193 - 4120 Page No. 8063 - 8074

mining", In Proc. IEEE Conf. Data Mining, pp. 1157-1161, 2006.

[45]. K. Sun et al., "Web content extraction based on maximum continuous sum of text density", In Proc. Int. Conf. Asian Lang. Process. (IALP), pp. 288-292, doi: 10.1109/IALP.2016.7875988, 2016.

[46]. N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents", Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[47]. S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining", Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[48]. X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data", Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

[49]. S. T. Dumais, "Improving the Retrieval of Information from External Sources", Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

[50]. J. Zhu, K. Wang, Y. Wu, Z. Hu, H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", IEEE Transactions on Knowledge and Data Engineering, Vol. 28(7), 2016.

[51]. X. Wang, Ji-R. Wen, Z. Dou, T. Sakai, R. Zhang, "Search Result Diversity Evaluation Based on Intent Hierarchies", IEEE Transactions on Knowledge and Data Engineering, Vol. 30(1), 2018.

[52]. Z. Hu, H. Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, "Discovery of rare sequential topic patterns in document stream", In Proc. SIAM SDM'14, 2014, pp. 533-541, 2014.

[53]. J. Shen, E. Zheng, Z. Cheng, C. Deng, "Assisting Attraction Classification by Harvesting Web Data", IEEE Access Vol. 5 Pages: 1600 - 1608, 2017.

[54]. J. Ruohonen, "Classifying Web Exploits with Topic Modeling", 28th International Workshop on Database and Expert Systems Applications (DEXA) Pages: 93 - 97, 2017.

[55]. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining sequential patterns by prefix projected growth", In Proc. IEEE ICDE'01, pp. 215-224, 2001.

[56]. J. W. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 1-12, 2000.

[57]. M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences", In Machine Learning, Vol. 42, no. 1-2, pp. 31-60, 2001.

[58]. M. Seno and G. Karypis, "SLPMiner: An algorithm for finding frequent sequential patterns using length-decreasing support constraint", In Proc. IEEE ICDM'02, pp. 418-425, 2002.

[59]. Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques", Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.

[60]. Y. Li, W. Yang, and Y. Xu, "Multi-Tier Granule Mining for Representations of Multidimensional Association Rules", Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 953-958, 2006.