Text Mining Methods and Techniques for Information Extraction in Web Data - A Review

Sridhar Mourya, Dr. P.V.S. Srinivas, Dr. M. Seetha

Abstract— The amount of text generated each day is increasing rapidly. Web text mining is the procedure of mining significance information, knowledge, or patterns from unstructured text from other sources. Discovering patterns in text and document association in documents is a well-recognized difficulty in data mining. The amount of information stored in the world nowadays is increasing day by day. Because it is typically an unstructured format, it cannot be exploited to extract constructive information, so it able to utilize several techniques, such as classification, clustering, and information extraction. Various techniques of text classification have recently been developed to find efficient and effective classification techniques for text classification. Several of them are directed and various of them are not supervised in sorting documents. This review paper focuses on the concept of web text mining methods and techniques utilized in information extraction and describes the actual application of text mining. It also provides a brief description of the benefits and limitations of text mining.

Index Terms— Information Extraction, Text mining methods, Web Text Data.

1 Introduction

THE growth of information is increasing constantly with the sharing of web data over the internet. A huge count of organizations, institutions, and professionals work efficiently and in an appropriate mode using this data. This data is distributed as structured (STRUCT) and unstructured (UN-STRUCT) data, and about 90% is collected in a non-regular format. Therefore, there is a need to retrieve meaningful data from this UN-STRUCT pool. Text mining (TM) helps it overcome this problem and easily accomplish these goals. This technology provides the ability to automatically retrieve text from other written sources. The former is different from Data Mining (DM) because it handles UN-STRUCT data and the latter deals with STRUCT data. The foremost objective of TM is to identify previously unknown information. Application areas such as "search engines, customer relationship management systems, filter emails, product suggestion analysis, fraud detection, and social media analysis" utilize TM for "opinion mining, feature extraction, emotion, forecasting, and trending analysis" [1]. The TM techniques are widely utilized in a variety of key areas such as "Web applications, the Internet, business intelligence, and content enhancement" [2], [3].

TM is in the field of DM, that attempts to discover appealing patterns in large databases [4]. TM, as well recognized as "intelligent text analysis", "text data mining", or "knowledge discovery of text" typically related to the method of retrieving appealing and insignificant information and knowledge from UN-STRUCT text. TM is an immature interdisciplinary area that combines "information retrieval, data mining, machine learning, statistics, and computational linguistics". Because most of the details (above 80%) are accumulated as text, TM is considered to have huge industrial perspective. While known to be found in several sources, the UN-STRUCT Text is the most eagerly presented basis of the knowledge [4], [5]..

TM able to work with UN-STRUCT or semi-STRUCT data groups for instance "email", "full-text documents", and "HTML files", except for DM tools [6], which are considered to manage STRUCT data in databases is related to DM. Consequently, TM is an improved answer for dealing with document retrieval. However, so far major of R & D attempts have focused on DM methods by means of STRUCT data. The difficulty posed through TM is noticeable. Natural languages (NL) have been extended that allow humans to correspond and confirm information, but computers are taking an extensive approach in recognizing the NL. Beginning with a document set, the TM tool searches for a specific document and identifies and preprocesses the format and character set. It then goes through the text analysis stage and occasionally repeats the technique until the information is retrieved.

According to [7], the boundary between DM and TM is blurred. The dissimilarity among the regular DM and TM is that in TM the patterns are mined from NL text, not a factual STRUCT database. The goal of DM is to determine formerly unidentified trends and patterns in the database. There is an extensive variety of differences between humans and computer languages, but technological advances have begun to narrow the gap. The NL processing discipline has taught computer NL and has created a technique to evaluate, recognize, and even to produce text information.

Several of the technologies developed and utilized in the TM method are "information extraction, subject tracking, summarization, classification, clustering, concept linkage, information visualization, and question answering" [1], [8]. The DM has a lot of practices, for instance, the "classification, clustering, neural networks, and decision trees". The following sections describe these techniques and their role in TM.

2 Web Text Mining

Web Text Mining focuses on finding interesting patterns in huge databases relatively textual information [9]. Information

- *Sridhar Mourya, Department of CSE, JNTUH - Hyderabad, India*
- *Dr. P.V.S. Srinivas, Department of CSE, SNIST - Hyderabad, India.*
- *Dr. M. Seetha, Department of CSE, GNITS - Hyderabad, India.*

recovery methods such as "text indexing techniques" have been building up to manage UN-STRUCT documents. Web TM is similar to DM but applies to free, UN-STRUCT and semi-STRUCT text. It able to be utilized to categorize documents into taxonomies and to process machine-supported text analysis [10]. It utilizes "information retrieval (IR), information extraction (IE) and Natural Language Processing (NLP)" techniques and associated with the database's knowledge search algorithms and methods of DM and machine learning.

Existing studies assume that users primarily search for known terms previously utilized or drafted by others. The foremost difficulty is that the search outcomes are not related to the necessities. A clarification to utilize TM to locate appropriate information that has not been unambiguously stated or has been written so far. The TM procedure begins by collecting documents through multiple resources. In certain documents able to be improved during TM tools and by examining the types and character sets. This instrument is pre-processed. The document undergoes a text analysis step. Text analysis consists of "semantic analysis" to acquire enhanced eminence information via text. It able to utilize other text analysis methods. It able to utilize different methods depending on the organization's goals. In several scenarios, the text analysis method repeats until information is retrieved. The results able to be accumulated in a "management information system" that presents a significant quantity of valuable information to the users of that system.

TM automatically extracts information from a diversity of "text-based sources" and detects unrecognized information. The STRUCT data able to be processed through DM tools, but UN-STRUCT or semi-STRUCT data categories for instance "full-text documents", "e-mails", and "HTML files" are able to be processed through TM. In general, information is kept in the natural form of text. TM is not like web mining, in general, the user searches on the web is something that was previously known and created through a different one. For instance, in e-commerce, the main problem with web mining is the acquisition of all the data that is not related to the user's search. It does not display unidentified information and the main purpose of TM is to retrieve anonymous information, somewhat that is not acceptable to anyone [11].

Data is an essential type of information that must be ordered and mined for knowledge creation. Discovering patterns and learning from the vast quantity of data is an important challenge. The foremost objective of DM is to uncover unknown trends and patterns in the database correctly. The "Data preprocessing" is a necessary method before applying any other method. Many methods, for instance, "clustering", "classification", and "decision trees", are involved with DM. All text-based information is accumulated electronically on the client's personal computer or web server, because of the increase in hardware storage devices.

It able to utilize TM techniques to retrieve "relevant information", "knowledge", or "patterns" from diverse sources in UN-STRUCT forms. The general structure of TM includes two repeated actions of "text refinement" and "knowledge distillation". The TM solutions are open-form text documents and transformed into an intermediate form, while knowledge characterization derives patterns or knowledge from in-between forms. Intermediate forms able to be STRUCT as semi-STRUCT or relational data diagrams as shown in the theoretical graphs. An IF can be document relies on, where all entities symbolize the document, and all units able to be conceptual relies on the entity or perception of attention in a particular area.

Web TM is made of numerous tasks, such as "clustering, classification, association, prediction, and normalization". This pattern mining approach depends on the classification concept. In general, classification is a type of "supervised learning model" [12]. This allows it to retrieve models that describe significant data classes or predict future trends. This model is utilized to predict goal values or attributes. It knows about background knowledge through processing the data that prepares the classification which able to eliminate data transformations, such as reducing the noise or processing the missing values to analyze relevant data to generalize the data with irrelevant or redundant attributes and high-level concepts, or to normalize the data. Classification is an effective way to distinguish between groups or object classes. This approach focuses on the external structure of relationships [14].

3 Text Mining And Information Extraction

3.1 Text Mining

The expansion of the "Web, digital libraries, technical documentation, and medical data" areas has equipped it simpler to retrieve bigger amounts of text documents to extend helpful resources. Thus, knowledge discovery of TM or text databases is a challenging task because it meets the NL standards utilized in most available documents. Database and online sources available in the form of text information [5], [6], [8], [10], raises questions concerning who is responsible for checking and analyzing data. It is not achievable to manually analyze and retrieve beneficial information in view of the relevant conditions. It needs to utilize a software clarification that able to analyze a large quantity of textual data, retrieve relevant data, analyze relevant data, and utilize automated tools to organize relevant information. TM has gained important importance in research as the demand for knowledge gained from the big count of text documents retrievable on the web increases [11], [13].

In general, TM and DM are believed to be similar, and there is recognition that the similar techniques can be exploited in mutual consideration to mining text. However, TM is different in that it involves STRUCT data, whereas text is a specific function, is relatively UN-STRUCT, and usually requires preprocessing. TM is also a field that is correlated with NLP. It is one of the upcoming topics of interest in interpreting and interpreting the language utilized by humans, as well as the correlation between an enormous quantity of UN-STRUCT Text [15], [16], [17].

3.2 Information Extraction

The starting point for computer evaluation of UN-STRUCT manuscripts is using IE. The software identifies the main expression and links contained in the manuscript. This is done by discovering a predefined arrangement in the text; This technique is term as "pattern matching model". The information in standard language text documents contains unable to make use of for mining. IE agree with the documentation, selects the suitable articles and associations between them to construct them additional useful for adding directions [20], [26]. Contrary to information retrieval, such as IR processes identification of related documents in a set of documents, IE generates structured post-processing information, which is necessary for different applications of web mining and search tools. [24] IE handles the discovery and extraction of important information from NL texts [25]. This involves separating the corresponding parts of the text, retrieving the data provided in those elements, and converting data into purposeful forms. The mining of the rating can now be performed by domain-specific texts; although the full IEs of random texts are still a continuous research objective [2], [3].

3.3 Retrieving Knowledge from Text

In the majority of circumstances, basically explicit data from the IR from UN-STRUCT Texts, rather than abstract knowledge, is obtained. In such a scenario, it is necessary to utilize a textual mining assignment, together with supplementary techniques to transfer knowledge from data [18], [19]. The "DiscoTEX (Discovery from Text EXtraction)" is among such approaches utilized for TM. This engages by means of IR primary to collect STRUCT data from UN-STRUCT text, subsequently using conventional KDDs for discovering knowledge from these data. This textual mining framework was described in [21]. In this scheme, the learned "IR system" is utilized to translate UN-STRUCT text into more STRUCT data. These data are then mined to build up consequential associations. In the case where the information retrieved from the corpus of documents is in the form of conceptual knowledge, rather than specific data, IR is likely to provide as a "knowledge discovery" of text. Detection of knowledge through retrieving information, such as key phrases or keyword extraction from the text, able to be utilized for other TM assignments, i.e. for the "classification", "grouping", "tabulation", and "topic disclosure" [22], [23], [24].

Patterns extraction evolution is exploited to update word-like models in documents and upgrade the d-pattern. This technique helps to reduce the effects of noise patterns due to low-frequency problems. This method is called an internal model evolution because it supports the pattern of the model [19]. This algorithm delivers a good result and provides efficient updates of detected models separated from text documents.

4 Text Mining Methods And Techniques

TM is typically utilized to get rapid results, and it is also the subject to research under various app areas[6], [12], [18]. Relies on the respective areas of the application, TM able to be classified as "text classification, text clustering, association rule extraction, and text visualization". These are presented in the following sub-sections.

4.1 Text Clustering

Text clustering relies on a cluster hypothesis suggesting that related documents are further similar to each other than unrelated documents [25]. Clustering technology is a reliable technique commonly utilized to analyze big quantity of data like DM. Text clustering has been proven to be the most popular effective tools utilized for text theme analysis [26]. In addition, topic analysis method [14] that performs clustering processing that groups frequently appearing named entities and applies "hypergraph" based methods to place frequent items. Each group of named units is represented through a cluster relating to the ongoing issues in the corpus. The content tracking process in dynamic text data attracts interest from researchers engaged in text clustering research in the digital field. Different methods and algorithms are included in the document clustering process. In the clustering process, the number of grouped sets, properties, and associations is not known first. A set of documents is organized by categorizing them into specific categories such as "medical, financial, law", etc. [13], [25].

4.2 Association Rule Extraction

The works of [28], [29] argues that the method of related to "association rule mining (ARM)" is utilized to recognize associations within big variables in the dataset. ARM recognizes the variable grouping that is likely to happen frequently. The method of the ARM in DM is also known as KDD in the database. This is related to the association analysis that discovers the relationship between the two variables. Wong et al. [31] present association rules for TM are primarily concerned with exploring relationships between various topics or de facto concepts utilized to characterize the corpus. They are aiming to determine main association rules related to the corpus in that manner that the happening of a particular topic in the article corresponds to the happening of a further topic.

4.3 *K*-Means Algorithms

The *k*-means approach segregates the dataset into "*k*-clusters", where each cluster is symbolized through an average of points. It is called a "centroid". A two-stage iterative method is adopted for the application of the algorithm. (1) "Allocate each point to the nearest centroid", and (2) "Evaluate the center of gravity of the recently developed group". When the cluster center of gravity reaches a certain value, the process ends. The "*k*-means algorithm" has extensive applications due to its straight parallelization. In addition, the sequence of each data does not influence the "*k*-mean algorithm" attributing numeric characteristics to that data. It is necessary to mention the maximum value of *k* at the establishment of the process. The illustration of the cluster is done by the "*k*-medoid algorithm" which selects the entity adjacent to the center of the cluster. However, the assortment of *k* items is prepared arbitrarily in the algorithm. The particular object is useful for determining the distance. Clusters are formed which are relies on the adjacent entity to *k*, but other items recursively attain the situation of *k* until the necessary eminence of the cluster is accomplished [13], [25].

5 Text Classification

Text classification is a "supervised learning" assignment for conveying a text document to a document of a predefined class. It is utilized to locate important information from the vast number of text documents existing on the knowledge database, the "World Wide Web" and the company-wide intranet. Classification of text is an assignment of repeatedly categorized a series of documents from predefined sets into categories. Categorization often depends on the domain on which the topic is predefined and identifies the relationship by searching for "broad, narrow, synonyms and related terms". The classification tools typically have a way of ranking documents in the sequence that documents have the major substance for a foremost topic.

Text data sets are generated by "natural speech processing", "typed text", and "handwritten text". A data set is an UN-STRUCT set of documents that is pre-processed by means of the three rules as, "Tokenize the file into entity tokens with space as the delimiter", "Removing the stop word which does not express any meaning", and "using porter stemmer algorithm to stem the words with common root word".

There are several ways to mine text, which able to be distinguished from a different view depending on the inputs made in the TM system and the DM works to be performed. Important methods rely on the type of input data are as follows.

5.1 Keyword -Based Association Analysis

This method collects a set of keywords or words that happen frequently as input and discovers that there is a relevance or relationship between the groups. Analysis of the association before processing the message data by separating, blocking, delete the stop words, then run the mining association algorithm. A set of closely-connected or nearby keywords is a word or phrase. The association's mining process able to help to determine compound relationships, terms or phrases that depend on a domain or non-compound relationship. Mining relies on these relationships is called mining-level association. The difficulty of association mining in a document database is mapped to association DM in a transactional database, which has developed many interesting methods. This analysis able to reveal superficial relationships, such as the discovery of mixed nouns or co-occurring variants, but it might not bring a deep understanding of the message.

5.2 Document Classification Analysis

Document classification is utilized for auto-tagging, creating topic directories, identifying documents, and categorizing the purpose of hyperlinks associated with a document group. The general procedures first organized documents as groups of training packages. Training packages will be analyzed to obtain the classification model. The classification is able to be utilized for other online document classification. This tagging method might rely on tags received by manual tagging, which is an expense and cannot be made. In the case of bulk documents or several mechanized categorization algorithms, which might process minute tag sets and necessitate prior classification.

5.3 Document Clustering Analysis

This is majorly significant techniques for classifying documents in an unmanaged manner. Due to the dimensional curse, the design of the document is a smaller area, the structure of the document space is comprehensible. In low-dimensionality areas, traditional clustering algorithms able to be utilized. For this reason, segmentation of mixed-grouping spectra, grouping using latent indexing and clustering using localized storage indexes is the best-known technique. This analysis brings meaningful information such as information facts or article discovered with data extraction. This approach is further moving forward and might direct to the finding of deep knowledge, however, necessitate the analysis of the meaning of the text, with the understanding of NL and the way of machine learning methods.

6 Text Classification Methodology

The process of automatically retrieving STRUCT data, such as inter-agency relationships, and attributes that describe entities from UN-STRUCT Text, is called information extraction. Often, it involves the generalization of interest data as templates. The prototype is utilized to implement the extraction process through "NLP, text processing and Feature selection and reduction".

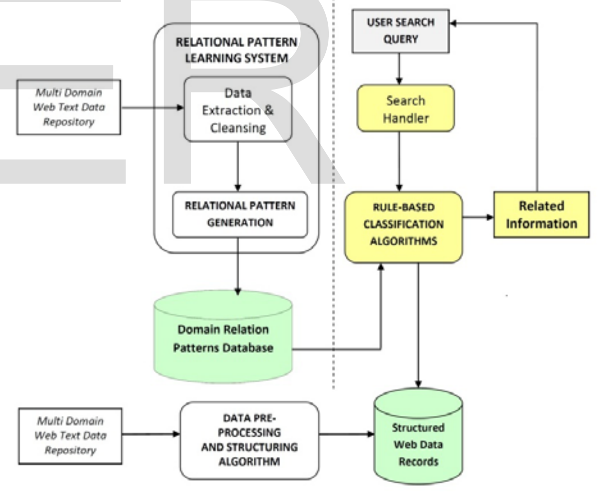
RELATIONAL PATTERN LEARNING SYSTEM
Multi Domain Web Text Data Repository
Data Extraction & Cleansing
RELATIONAL PATTERN GENERATION
Domain Relation Patterns Database
USER SEARCH QUERY
Search Handler
RULE-BASED CLASSIFICATION ALGORITHMS
Related Information
Multi Domain Web Text Data Repository
DATA PRE-PROCESSING AND STRUCTURING ALGORITHM
Structured Web Data Records

Fig. 1. A Context Framework for Relational Pattern Learning for Information Classification.

Fig. 1 shows the framework context for adaptive learning in relational modeling for data classification in web DM. The text describes the interactions of components for retrieving web data using data. UN-STRUCT and non-grammatical forms. Data sources are collected from various domain sites. The received data page is processed by parsing the data and retrieving it to obtain a new theme. Patterns are utilized for mapping user searches, which are scaling for UN-STRUCT and UN-STRUCT data.

6.1 Natural Language Processing (NLP)

The NLP is a theoretical motivational field for computational techniques to analyze and display naturally occurring text more than one levels of "linguistic analysis" for the reason of accomplishing "human-like linguistic processing". The function of NLP in TM is to offer the language information needed to complete the mission to the system of the data extraction process. This is prepared by describing the document with information such as sentence limitations, partial word tags, and parsing of search results. Search results able to be read by the import tool. Apply NL processing methods to the extraction process and reduce the attributes of the message classification process.

6.2 Text Preprocessing

The primary action in retrieval systems is to recognize the keywords for displaying this document through tokenization. Text documents are grouped into words by eliminating all punctuation and replacing non-tab and non-text characters. It utilizes the value that is specified to be tokenized as the only space. A group of words comes from a set of documents. All the text in a set is called a dictionary. It able to utilize the filtering method to decrease the dimension of the dictionary and create a narrative of the document within the collection.

6.3 Feature Selection and Reduction

The purpose of the feature selection approach is to decrease the dimension of the data set by eliminating features that are not related to classification. This conversion process has many benefits, including small data sets, low computing requirements for text classification algorithms, and reduced search space. The goal is to reduce the curse of dimensions so that they are more accurate in classifying and reducing crosstalk and widening. The process of picking a feature will include the following steps: (1) "Create a candidate set, including a subset of original features through a specific research strategy", (2) "Evaluate candidate sets and evaluate the ability to utilize features in the listing applicant". Depending on the assessment, some features of the candidate package may be discarded or added to the selected feature set depending on its relevance and (3) "whether the selected feature set is good enough today using a specific stop criterion". If yes, the feature selection algorithm will return the selected feature set, otherwise, it will repeat until the threshold is reached.

Reducing the appearance involves a combination of three common approaches, namely, "stopping words", "inhibiting" and "filtering statistical data". The proposal of stopping word filtering is to eliminate words that contain modest or no content, such as "articles, conjunctions, and prepositions" as a stop or pause a word list.

7 Related Work

A lot of research has contributed to IE using a variety of technologies. The main focus of this study is to determine how various TM cases able to be utilized with STRUCT data sets in text documents. This section commences by defining research topics, estimating previous research, and then using key technologies using data and DM. Each research area defines the subject and develops the connection between the layers and the evolution between these topics. In [24] it utilizes the TM method for topics, which is provided through visualization tools. These tools also display links between these topics and provide interactivity so that users able to easily find topics in their domain and identify trends in cross-domain research.

Y. Li et al. [6] discuss the problem of message mining and the existing message classification techniques. Either way is acceptable by the term. They analyze how previous techniques have suffered from multiple-dimensional problems and synonyms. It also demonstrates the need for effective tools to efficiently utilize large format tools. It has introduced a "relevance feature discovery (RFD)" to locate the related character in text documents. It addresses two key concerns in mining, for instance, low-level support mining leading "WFeatures" and "FC-clustering algorithms" in RFD format. The "FC-clustering algorithm" describes the process of grouping features and discover a set of pattern algorithms. The "feature algorithm" is utilized to calculate the weight of the given specifications.

M. Yadav et al. [9] Summarize the preliminary discussion of WEB mining on web applications, critical science applications, and present some of the most significant computer science in the field of mining and areas that are promising for future research. From the outset, the potential to retrieve valuable knowledge from the web is clear. The application of DM techniques to mine knowledge from web mining to meet this capability. In web data, at least one structure or usage data is utilized in the mining process. Attention towards web mining has developed quickly in the short term in research communities and practitioners.

K. Radinsky et al. [11] describes and evaluates how to learn to predict interesting events in materials with 22 years of news articles. This case illustrates the possibility of outbreaks, deaths, and riots before these events occur. Here it will find detailed information on how to distinguish and summarize events from news archives and web resources and research. It evaluates the predictability of access to actual events in the system.

I.A. Moloshnikov et al. [14] develops algorithms for probing for documents on specific topics relies on reference collections. In addition, the "contextual graph" for the visualization theme in the search results has been extended. This algorithm relies on a combination of keyword weighting with mining and entropy developers, probability and meaning for a group of words that describe a given topic. The results demonstrate that the average accuracy is 99% and the recovery rate is 84%. Specific techniques for graphing relies on algorithms and weight able to be derived. This will give it the opportunity to display an array of bullet points in a large document in dense graphs.

S. Shehata et al. [16] presented a "concept-based model". It analyzed not only document levels but also terms in sentences. The proposed model includes a conceptual statistical analyzer, a conceptual graphical representation on and a concept retriever. In this model, every sentence in a text document is automatically labeled with the assist of "PropBank notation". In together the verbs and influence terms are supposed. The model also includes a "concept-based statistical analyzer", a "COG representation", and a "concept extractor". It is utilized to keep up sentence semantics relies on statistical analyzers. The weights calculated by the "concept-based statistical analyzer" are displayed with the COG demonstration.

S. T. Wu et al. [18] proposed a "pattern taxonomy extraction method" for retrieving descriptive "frequent sequential patterns" through subtracting meaningless ones. Instead of "keyword-based concepts", it utilized a "pattern-based model" that utilized frequent sequential patterns. It is mostly utilized to explain the difficulty of "mining sequential patterns" in text documents. Pattern mining helps to remove meaningless patterns in the tree structure and pattern classification that show relationships between patterns retrieved from a collection of text. They aim to apply pattern mining to filter user profiles to filter out inappropriate documents relies on a user's profile. When it utilize pattern mining to get a topic or pattern, the centroid, or feature vector, is utilized to process the representation of the subject area. The patterns retrieved from the training set are represented.

The discovery of information of important features in text documents appears to be an effective approach to determining document relevance, that is, relevant or irrelevant documents in the IR domain. Existing technologies rely on a "terminology-based approach". The "term-based approach" mines the conditions set in the training to describe the relevant functions but suffers from a low level of support issues. Therefore, the problem of retrieving relevant information in web TM occurs in the context of various application domains for instance "the web, social networks, and other digital collections". The core goal of this research is to find extremely applicable information relies on constructive relationship pattern learning for information classification, taking advantage of useful features in text documents that describe Web TM results. This is considered a difficult assignment in contemporary information investigation for mutually empirically and theoretically [27], [30]. This issue is the main anxiety for numerous Web applications and has attracted awareness from researchers in "Data Mining, Machine Learning, Information Retrieval, and Web Intelligence communities".

8 Conclusion

TM has recently become an important area. The immense quantity of information and other sources presented on the web has proven to be beneficial to organizations in many different areas. This information is not available in a readable format and requires pre-processing. In other words, a TM process is required for analysis. Information extraction, information retrieval, and the visual representation of information are essential elements of the TM process. These various technologies are utilized for various needs. If the simple query is input, retrieving information is the best choice, whereas if it needs complex information from UN-STRUCT or semi-STRUCT data, it needs to create a model that automatically extracts the information. Alternatively, simple techniques such as visualization are useful for simplifying the process of retrieving relevant information. TM is useful in many areas. From education and health care to business organizations, TM is utilized extensively for a diversity of intention. Despite some problems, TM certainly saves a great deal of time through automated work.

References

[1]. S. A. Salloum, M. Al-Emran, A. A. Monem, K. Shaalan, "A Survey of text mining in social media: facebook and twitter perspectives", *Advances in Science, Technology and Engineering Systems Journal*, Vol. 2(1), pp. 127 - 133, 2017.

[2]. W. Hua, Z. Wang, H. Wang, K. Zheng and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", *IEEE Transactions On Knowledge And Data Engineering*, pp. 1041-4347, 2016.

[3]. P. V.-Elizondo, R. M.-Piña, S. Vazquez-Reyes, A. Mora-Soto, J. Mejia, "Knowledge representation and information extraction for analysing architectural patterns", *Science Computer Program*. Vol. 121, pp. 176–189, 2016.

[4]. H. Hassani, X. Huang, E.S. Silva, M. Ghodsi, "A review of data mining applications in crime", *Statistical Anal. Data Mining ASA Data Sci. Journal*, Vol. 9(3), pp. 139–154, 2016.

[5]. X. Zhai, Z. Li, K. Gao, Y. Huang, L. Lin, L. Wang, "Research status and trend analysis of global biomedical text mining studies in recent 10 years", *Scientometrics*, Vol. 105(1), pp. 509–523, 2015.

[6]. Y. Li, A. Algarni, M. Albathan, Y. Shen, and M.A. Bijaksana, "Relevance Feature Discovery for Text Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27(6), Pp. 1656 - 1669, 2015.

[7]. J. Zhu, Member, K. Wang, Y. Wu, Zhongyi Hu, and H. Wang, "Mining User-Aware Rare Sequential Topic Patterns in Document Streams", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 28(7) Pp. 1790 - 1804, 2016.

[8]. N. Venkata, L. Padmasree and N. Mangathayaru, "Survey of Text Mining Techniques, Challenges and their Applications", *International Journal of Computer Applications*, vol. 146, no. 11, pp. 30-35, 2016.

[9]. M. Yadav and P. Mittal, "Web Mining: An Introduction", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3 (3), March 2013.

[10]. Y. Zhang, M. Chen, L. Liu, "A review on text mining", *Proc: 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 681–685, 2015.

[11]. K. Radinsky, E. Horvitz, "Mining the Web to Predict Future Events", *Proc. ACM WSDM'13, Rome, Italy*, 2012.

[12]. N. Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, NO. 1, January 2012.

[13]. M. Chebel, C. Latiri, E. Gaussier, "Extraction of interlingual documents clusters based on closed concepts mining", *Proc. Comput. Sci.* Vol.60, pp.537–546, 2015.

[14]. I. A. Moloshnikov, A. G. Sboev, R. B. Rybka, D. V. Gydovskikh, "An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach", *4th*

International Young Scientists Conference on Computational Science, Vol. 66, pg. 297–306, 2015.

[15]. C.-H. Chang, M. Kayed, M. Girgis and K. Shaalan, "A Survey of Web Information Extraction Systems", *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1411-1428, 2006.

[16]. 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.

[17]. 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.

[18]. S. T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining", *In Proc. IEEE Conf. Data Mining*, pp. 1157-1161, 2006.

[19]. W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding", *In Proc. of the ACM SIGMOD International Conference on Management of Data*, pp. 481-492, 2012.

[20]. M. Banko, Michael J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. "Open information extraction from the Web". *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 2670-2676, 2007.

[21]. 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 Conf. on Info. & Knowledge Management*, pp. 1401-1410, 2013.

[22]. Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai, "Semantic annotation of frequent patterns", *ACM Trans. Knowledge Disocvery from Data (TKDD)*, vol. 15, pp. 321-348, 2007.

[23]. L. W. Han and S. M. Alhashmi, "Joint Web- Feature (JFEAT): A Novel Web Page Classification Framework", *Communications of the IBIMA*, pp.8, 2010.

[24]. A. Herrouz, C. Khentout, M. Djoudi, "Overview of Web Content Mining Tools", *International Journal of Engineering And Science (IJES)*, Vol. 2(6), 2013.

[25]. 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.

[26]. G. Li, D. Deng, and J. Feng, "Faerie: Efficient filtering algorithms for approximate dictionary-based entity extraction", *In Proc. for ACM SIGMOD International Conference on Management of Data*, pp. 529-540, 2011.

[27]. 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.

[28]. C. H. Mooney and J. F. Roddick, "Sequential pattern mining approaches and algorithms", *ACM Computer Survey*, vol. 45, no. 2, pp. 19:1-19:39, 2013.

[29]. 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*, pp. 533-541, 2014.

[30]. Y. Li, A. Algarni, and N. Zhong. "Mining positive and negative patterns for relevance feature discovery", *Proc. for 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 753-762, 2010.

[31]. P. C. Wong, P. Whitney, J. Thomas, "Visualizing association rules for text mining", *IEEE Symposium on Information Visualization*, pp. 120–123, 1999.