A SURVEY ON TEXT MINING TECHNIQUES FOR RETRIEVING INFORMATION FROM WEB PAGES

[1]Ramkumar V , [2]Dr.N.Krishnaraj
[1]Research Scholar, Shri Venkateshwara University, Uttar Pradesh
[2]Professor, Department of Computer Science and Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, Andra Pradesh, India

Abstract

Information Retrieval (IR) is a rapidly rising domain in Natural Language Processing (NLP), wherein search engines are the most popular as well as useful applications. Some engines such as Google or AltaVista are utilized by several individuals in their day to day lives. Research within the domain of Information retrieval has centred on certain important problems which retain importance in the age of commercial web search engines that work with millions of web pages. This review article addresses the issues in web page classification over the last two decades.

1. Introduction

During last two decades, most research in Information Retrieval was aimed at retrieving documents and the focus on the task given during the Text Retrieval Conference (TREC) evaluations in the 1990s reinforces the notion that Information retrieval was considered a synonym to retrieval of documents. Search engines on the web are generally the most frequently occurring kind of Information Retrieval systems. Theoretically, information storage as well as retrieval are simple tasks.

The primary aim of IR systems is the discovery of relevant information or documents which fulfil user information requirements. For achieving this aim, information retrieval systems typically execute the procedures mentioned below:

- During indexing, a document is represented in the form of summarized content
- During filtering, all stop as well as common words are discarded.

Searching is the primary procedure in information retrieval systems.

Semantic Web refers to web with a particular meaning. It delineates things in a fashion comprehensible to machines. It is an extension to the typical web and is not concerned with the relations among things and their characteristics. Traditional Web comprises human operators and utilizes machines for jobs such as discoveries, searches as well as aggregation while Semantic Web refers to that which is comprehended by machines, performs the searches, aggregation as well as combination of information with

no human operators present. It is easy to process for machines, on a huge scale. It is also the most effective method of denoting information on the World Wide Web.

1.1 Text mining

Text mining in another context is also called as text data mining, nearly may sound as text analytics. It means the processing of derivation of information with superior from the given text data. Due to the rapid development of the information there exists a great need to extract and discover valuable knowledge from the available huge information received from various sources of data like, World Wide Web etc., In general data mining, refers to the field of retrieving purposeful information and at times top level knowledge from the given set of large set of raw data [1].

Text mining is normally considered as high difficulty task compared to conventional data mining and it is related to the fact that traditional databases have constant and known structures whereas text documents are known to be unstructured and at the same time, web document can be of semi-structure. Hence text is supposed to involve a series of steps such as data pre-processing and modelling for condition the given raw data in to structured data. Text mining helps in various tasks where in other case it may need huge manual work. General issues solved through text mining includes searching through document, organizing document, comparing document, extraction of significant information and document summarization, etc., Natural Language Processing (NLP) refers to the concept including latest computational technique and the systematic process to investigate and evaluate claim on human language. NLP is the process that relates back into the history of Artificial Intelligence (AI) and the common investigation by computational process on cognitive function with the importance towards the representing of knowledge.

Text mining uncovers novel unidentified or secret information from the given text data through the process of extraction by various methods. Text mining is known to be multidisciplinary area that concerns with retrieval of information, analyzing the textual data, extracting of information from data, categorizing, clustering, visualizing, data mining, and machine learning. Various important techniques that differ from each other are used by text mining. The information retrieval techniques used unstructured text is different from that of which is used for the structured databases. The Summarization techniques are applied for summarizing the documents to reduce size but to keep remains the same meaning. The process of categorization falls under supervised process type and it uses a predefined set documents based on the content to be categorized. Clustering deals with large dimensional data given, identifying significant pattern related with the data given. It also includes that feature that it is a set of similar data and its relationship[2].

Most of the text mining techniques depend on Vector space model (term frequency). It holds the significance of the term inside the documents, but at some point of time two terms of the same document

can have same frequency. Hence there is a chance of redundancy and irrelevancy in the result. But the meaning contribution of one term may be highly appropriate when compared to the other.

The exploration of sequential and appealing pattern altogether along with knowledge present in the unstructured text document is meant to be the normal text data mining method that also adopts certain discovering strategy and their knowledge on the concerned databases. The information received so far can be linked together for forming novel hypothesis to be included for future study through setting up of fresh experiment. MEDLINE biomedical databases by the integration a frameworks to named entities recognitions, classifying of texts, hypothesis generating and testing, relationships and synonym extracting, extraction of abbreviation. This novel framework assists in elimination of non-significant details and to help to extract only precious information [3]. To analyze the textual data with text mining pattern and shows term dependent approach is not able to analyze synonym and polysemy appropriately and also a prototype system was designed to specify the pattern in term of weight assigning as per its distributions[4].

To present novel and effective pattern discovery technique. It uses the pattern developing and discovering technique for enhancing the effectiveness of identifying related and exact information. It performs BM25 and vector support machine depended filtering on router corpus volume 1 and text retrieving conference text data for estimating the efficiency of the suggested techniques [5]. Performing several experiments of classification process using multi-word feature on the textual data. The researcher proposed a hand-crafted technique for extracting multi-word feature from the given data sets. For classifying and extracting multi-word textual data they divided the text into linear and nonlinear polynomial form in support of vector machine that improves the efficiency in the extraction of data [6].

Google and Yahoo search engines most often use information retrieval system for extracting relevant document as per the phrase given on www. They use query based algorithm for tracking the trend and achieve much more important result information. They also provide user more related and exact information in order to meet the needs of the user [7]. For extracting synonyms and abbreviations text, co-reference method is often in use for NLP. Natural Language (NLP) holds more complexities since the text retrieved from various sources does not have similar word or abbreviations and moreover there is no requirement for detecting these problems and develop rules for their uniform finding [8].

2. Text Mining Techniques

2.1 Concept based mining:

Various techniques of text mining executes on statistical analyzing of terms, either words or phrases. Statistical analyzing of term frequencies captures the significance of terms within documents only but anyhow the two terms of the same document can have the same frequency with one term contributing highly

to meaning of its phrase compared to other terms. Thus, the fundamental text mining prototype indicates term that captures the semantic of textual data. Here, the mining models captures term that presents the concept of the sentences, which lead to identify the topics of the documents. A novel concept-based mining model that analyzes term on the sentences, documents, and corpus levels is given. The concept-based mining model effectively discriminates among non significant term with respect to sentences semantic and term that holds the concept that represents the sentences meaning. New mining model contains of sentence-based concept analysing, document-based concept analysing, corpus-based concept-analysing, and concept-based similarity measures. The terms that contribute to the sentence semantic is analyzed on the sentences, documents, and corpus level compared to conventional analysing of the documents. The proposed system effectively finds interesting matching concept between documents based on the semantics of its sentence. The resemblances between documents are measured with new concept-based similarity measures. The proposed similarity measuring technique is completely advantageous of using the concept analysis measures on the sentences, documents, and corpus level for measuring the resemblance between the documents. Huge set of experiments were done with the new concept-based mining model on various data set in textual data clustering. The experiment demonstrates wide comparison among the concept-based analysing and the traditional analysing. Experimental result demonstrates the substantial improvement of the clustering quality with the sentences-based, documents-based, corpuses-based, and merged approaches of concept analysing [9].

In novel concept-based mining model, the semantic structures of every term found in the sentences and documents compared to frequencies of the terms lying in the document. Every sentence is labelled with semantic role labellers that determine the term which contributes to the sentences semantic linked with their semantic role of sentences. Every term that have semantic roles in the sentences are known as concept. A Concept shall refer to either a word or phrase. They are completely based on the semantic structures of the sentences. While new documents are introduced to the set up, the proposed mining model detects concepts that match from the given documents to the entire processed documents previously present in the data sets through the process of scan to the new documents and extracts the matching concept. Thus this setup fills the gap present between natural language processing technique and text mining discipline. A novel concept based mining model consists of two components and it is supposed to enhance the quality of text clustering. By exploring the semantic structures of the sentence in document, the best text clustering results are obtained [10].

2.2 Conceptual clustering:

The need to conceptual document clustering technique is increasing for managing several kinds of extensive information given on WWW. The authors utilize formal concept analysing (FCA) methods for cluster document based on the formal context. Concepts hierarchy of document is constructed using the formal concept of the document in the documents corpuses. It uses tf.idf (term frequency /spl times/ inverse document frequency) terms weighting model for reducing minimum useful concept from the formal concept and the association and correlation mining technique for analyzing the link of terms in the documents corpuses [11].

General clustering technique has the disadvantages such as it does not provide intentional description of the cluster received whereas Conceptual Clustering technique provides such description but the only disadvantage is that it is slower. This approach has two steps: First being that common (non-conceptual) clustering algorithm is applied and in this case a variant of the well-known k{Means algorithm is used for decreasing the size of the problems. Next, the resulting clusters are clustered using conceptual clustering techniques. This paper related to Formal Concept Analysis, the problem of text clustering is focused. For improving the quality of the clusters, added to this the background knowledge present as thesaurus is used. With the given set of document and thesaurus, this paper provides clustering of the document with remarkable performance that is done with intentional description of the cluster [12].

Short text clustering is rapidly growing as significant methodology, but only the problem is that it has sparsity and large dimensionality of textual data. Earlier concept decomposition systems had reached concept vector through the centroid of cluster using *k*-means-type clustering algorithm on normal, complete text. This investigation proposes a novel concept decomposition technique that develops the concepts vector by locating semantic words community from weighted words co-occurrence networks retrieved from short text corpuses or subsets. The clustering membership of short text is the calculated by mapping the actual short text to the supervised semantic concept vector. New method is found not only robust towards the sparsity of short texts corpora, instead it also overcome the problem of dimensionality that refers to the huge quantity of short text input due to the concept vector that is achieved from term to term instead of document to term space [13].

2.3 Semantic based text clustering

A new system of ontology for identifying the similarity that preferably exists between the concept and the sentiment in terms of the semantic. Unlike having semantic similarities, this method focuses on the links that establish connections between the concept and its relationship of the constructive sentiment [14].

Much familiar techniques for text mining usually depend on the statistical analysing of the term or phrase and in this the frequencies of the term are considered for finding the significance of a term in the document alone, whereas the term which contributes to the sentence semantics is important that leads to the identification of the topics. Here, in the proposed system the NLP technique is effectively used for capturing the semantic of the texts can be used for enhancing text clustering and also concept based mining model is given. The terms that contribute to the sentence semantic is processed on the sentences, documents and corpus level compared to conventional analysing of documents alone. Based on the semantic of the sentences, the system identifies concept matching between the document and uses for the improvement of the cluster [15].

Text mining refers to the process of data extraction from textual database or document given. Every word present in the document gives a structure to the text and reduces the dimension. In text mining technique, the basic metric such as term frequencies of terms are calculated for working out the weights of the terms in the documents. Even if with the present statistical analysis, the actual meaning of the terms does not give the exact meaning of the terms. The proposed system depends on concept based model. Concept based approach, executes as the concept is retrieved from the document, and semantic based weights are calculated for effective index and cluster process using MeSH ontology for concept extracting and concept weight calculation with respect to the identity and synonymy relation. K-means algorithm is applied to cluster the document based on semantic similarity [16].

Hierarchical classification refers to the effectual method for categorizing large-scale of text. This work introduces a relaxed strategy along with the conventional hierarchical classification technique for improving the system's performances. A novel term weighting model based on the Least Information Theory (LIT) is used to hierarchical classification process. These processes quantify information in probability distribution change and offer a fresh document representing model in which the contribution of every term is exactly calculated. The experimental result shows that the relaxed strategical method builds a high reasonable hierarchy and also enhances classification performances and also outperforms existing classification techniques such as SVM (Support Vector Machine) in efficiency. The approach is also much more efficient

in case of large-scale text classification task. While comparing the traditional classic term weighting method TF*IDF, LIT-based method gives significant increase in the classification [17].

Text classification in the context of hierarchical taxonomy of topic is very general and realistic problem. Existing methods simply consider bag-of-word and has received better result. But, while there exist a number of labels with various topical granularities, bag of-word representation is not sufficient. Deep learning model is proved to be effective for automatic learn different level of representation for image data type. This gives interest for study the best approach for representing text. Here it is proposed, a graph-CNN based deep learning model for converting text to graph-of-word primarily, and then to use graph convolution operation for convolving the word graphs. A graph-of-word representation of text has the advantage of capturing non-consecutive and long-distance semantic. CNN model has the advantages of learning various levels of semantic. For further leveraging, the hierarchy of label, the deep architecture with the dependency among label is regularized. Experimental result on both RCV1 and NYTimes dataset shows the significant improvement in large-scale hierarchical classification of text compared to existing deep model and classification. Two accepted deep learning architectures received great concentration for text data, i.e., recurrent neural network (RNN) [18]

CNN uses convolutions mask for sequential convolve process over the data. For text, an easy mechanism is the recursive convolution of the neighbouring lower-level vector in the series for composing higher-level vectors. These approaches of using CNN easily evaluate the semantic compositionality of consecutive word that relates to the n-grams seen in conventional text model. Similar to images, such convolution generally represents various levels of semantic given by the text. High level corresponds to semantic captured by larger “n”-grams

Ontology, as conceptual model, provides the actually needed framework to semantic representing of text. The principal relationship between textual data and an ontology refers to terminology that connects term to domainText is the predominant medium for exchange of information between the experts [19].

Bio- medicinal areas have seen rapid growth in the publication quantity and have the lack in standard scientific terminology. This is more challenging, involves high complexity and also more time consuming. Earlier information retrieval method uses term matching and term frequencies and they are developed for finding the parts of document that contains the definition and association of scientific concept being investigated. This work presents a novel method for extracting parts of publication which are highly relevant towards biomedical concept that are already defined through domain ontology and rule. This work uses hierarchal clustering for identifying the parts of the text that contains the highly relevant term to the concept

and uses vector space modelling for extracting the text. Next it uses cosine similarity measure to measure the text correlation to the concept, rank it and retrieves the one with high relevancy. This method is applied to knowledge bases of autism phenotype definition that are simulated with web ontology language, OWL, and its rule language, SWRL. Finally the accuracy and relevancy of three level of semantic using the ontology hierarchy, the rule definition, or both is compared. The result shows that the ontology hierarchy provides the highest accuracy (73%) in identifying the matching text defined in the domain concepts. But the combined ontology hierarchy and rule definition approach provides the highest accuracy (76%) in retrieving any text that referred to the domain concept. [20]

2.4 Self Organising Map

Self-organizing map (SOM) is considered to the highly popular neural network method for cluster analysing. Clustering method using SOM normally has two-stage procedure: Primarily original data are projected onto a set of prototype on an ordered grid by SOM. This prototype is noticed as proto-cluster that will be linked together in the later stage for obtaining final clustering result [21].

The aim of SOM is to present entire input vectors in a high-dimensional space by prototype in a low-dimensional space where it preserves the distance and topology with most possibility. SOM refers to the much familiar neural network method to cluster analysing for 2 reasons. Prime is SOM consists of the property of self-organized and topology-preserved. Closed data in the input data spaces are projected onto closed prototype vector present on the grid next to training phase. Hence 2 input vectors that are inserted onto closed prototype mostly belong to the unique cluster. Next is that it has major visualization property which provides that the ordered low-dimensional grids are considered as natural visualization surface for exhibiting cluster structures. For visualisation, U-matrix is general, yet powerful tool . As the training of the map is over, U-matrix is portrayed on the top of the grid for giving the focus into local distance structure among the closed prototype structures.

The definition of the Self-organizing map (SOM) states that its definitive goal is the process to convert incoming signals to the dimensions that are arbitrary to low discrete maps of one or two dimensions for performing by adoption itself to the topologies of the networks. Few properties of SOM assists in the implementation of useful change in the networks. For example the fundamental nature of the SOM which makes the detainment of topological link among the input data brings data interpreting task as simpler. Similarity sharing task between the input data is other way of enabling visualization of data. The cluster exploration task is also made easier by features of SOM. It is applied in the machine learning process for completing the tasks such as data clustering, classification and graph mining.

A system that includes a combination of novel growing self organizing map (GSOM) and latent semantic analysis (LSA) to enhance the quality of the result of textual data clusters instead of depending on GSOM only [22]. An unlearned and competitive kind of learning algorithm known as SOFM. The resemblance and the relationships among the input sample is grouped into map of different topology in the structure of space having high dimension and at last it is changed into relationship which are spatial among two dimensional neurons [23]. Conventional SOM algorithm trains from data with the application of fixed maps. This work proposes a new SOM learning algorithm which expands the map sideways and also in hierarchical. The adoption of the mapping structures are based on topic identified from the fundamental document clustering. This technique differs from conventional approach that is classically driven by the cluster data density. Experiment results suggested the outperforming of the proposed algorithm compared to similar existing approaches [24].

To develop a language-neutral technique for tackling the linguistic difficulty in the text mining task with the variation of automatic clustering technique that applies neural net approaches such as Self-Organizing Map (SOM). SOM generates two maps viz., the word cluster map and the document cluster map that reveals the relationship between words and document correspondingly. Search processes incorporate this map and effectively find the related document based on the keyword given in the query list. The conceptually linked web document is found not merely with the specific keyword but the related word located through the word cluster map. This work presents new SOM-based technique for text mining. The document given is first converted to a group of feature vector where every, component relates to a varying word and the value of the component presents the word- occurring frequencies in the documents. The vector is used as inputs for training the self-organizing maps. Two maps such as the word clustering map and the document clustering maps were got through the labelling of neurons with word and document correspondingly. Related documents are hence found very easy with this process [25].

A new SOM learning algorithm proposed and which expands the map sideways and also in hierarchical that could identify the relationship between documents in both perceptions. The proposed new algorithm clusters a set of training document with typical SOM algorithm and then identifies the topic of every cluster and uses them for evaluation of the constraint on expansion of the maps [26].

A topic-oriented SOM method that is apt to document cluster and organization is designed. Newly proposed SOM automatically adapts the quantity and structure of the map based on the identified topic. Contras to existing data-oriented SOM's, the method expand the mappings and generate the hierarchy both based on the topics and its characteristic of the neuron [27].

3. Genetic Algorithm for text clustering

Genetic algorithm technique according to the latent semantic model (GAL) to cluster the text was studied. The major constraint is this application of genetic algorithm (GA) to cluster the document is the existing multiples of thousands of dimensions available in feature space that is apt for text. Since many of the straightforward and popular approach signify text with vector space model (VSM) which means that every single term present in the vocabulary signifies dimensions. Latent semantic indexing (LSI) refers to a booming technology in the concept of information retrieval that tries to explore the latent semantics inferred by queries or documents through their representation in a dimension-minimized space. At the mean time, LSI considers the impacts of synonymy and polysemy that construct semantic structures in text. Genetic Algorithm belongs to search technique that efficiently evolves the optimal solutions in the minimized space. This work proposes a variable string length genetic algorithm that is explored for automatic evolve of the exact quantity of clusters and provide best optimal clustering of data. Genetic Algorithm can be applied in combination with the minimized latent semantics structure and enhance clustering effectiveness and accurateness. The supremacy of GAL technique on the traditional GA used in VSM model is given through best Reuter document clustering result [28].

Aim of automatic text summarization method is the generation of summary of the actual data that permits the end user to achieve the major piece of information existing in the text but with very much least reading time [3]. Additionally, significant data pre-processing task leading to effective classification lies with attribute selections that consist of selection of the highly related attribute for classification purpose. The aim of this work is deploying Genetic Algorithm in the concept of term weightage and feature selections in text mining processes. Genetic Algorithms are used in text processing. In this proposed concept, “concept” refers to mention certain relevant words that points to specific entities or impressions in documents. With the controlled vocabulary set, the concept group of document are extracted from textual data. Every concepts or terms may be weighted between 0 and 1. In order to bring down the issue, the weights are considered as binary digit which means that either the relevant concept or term belongs to the document or not. Chromosomes are defined as set of concepts or term weight that has real or binary digits [29].

Genetic Algorithms (GA) are applied for extracting term (feature) as needed as per term significance computed by the equation seen. The investigations revolve around features selection techniques for lowering computational complexities and for increasing analytical performances. A novel genetic algorithm is designed for extracting feature in text mining task. TF-IDF is applied for reflecting document-term relationship in feature extractions and with the iterative procedure, feature is chosen as many as the predetermined count. Clustering experiment is done on group of spam mail document for verifying and

improving feature selection's performances and it is found that the proposed FSGA algorithm shows better performance of Text Clustering as well as Classification compared to the use of other features. Documents classification experiments show better performance in classification than using FSGA but only except the recall result. This provides the result that if the large quantity of object then the classification performance is high since the classification is used for learning [30]

.

The research targets at proposal of a novel system to apply the genetic algorithm for grasping the technology evolution path. To achieve this, at first the technology and natural evolution's analogy is investigated. And next the morphology of the chosen technology is defined through extraction of the keyword information received from patent document that is gathered from USTPO (United State Trade and Patent Office) database by text mining task. Finally, genetic algorithm is used for analyzing the evolutionary patterns of technologies and derives the configuration of hopeful technologies with software, describing the cell, mutations and selections of technologies. Primarily, patent data of interests are gathered for applying the new approach towards a particular technology. As patent data is considered to be the highly efficient information that exhibits the technology nature, it is chosen for deriving the evolutionary characteristic of technologies. Secondly, the chromosomes of technologies shall be identified related to genetic algorithms by text mining. Next as thirdly, chromosome of existing patent is reproduced with selections, crossovers and mutations for producing children for the next generation. Fourthly, the probability of the chromosome is calculated by fitness function that is given as value of patent in the present work. At last, the evolutions of the gathered technologies are studied through comparison of the result of genetic algorithms and the chromosome of real patent. As patent information is used for investigating technology evolutions, patent filed to USPTO (United States Patent and Trade Office) in the duration is known as the initial generation. Next three process such as selections, crossovers and mutations are used for preceding the genetic algorithms as reproduction method [31].

4. Suffix tree for text clustering

STC refers to a linear time clustering algorithm (linear in the size of the document set). It relies on the identification phrase which is common to group of document. Phrases are ordered sequence of one or more words. STC algorithm differs from existing clustering algorithms. STC is a data structure that has all the suffix of a given string that could efficiently run various significant string operations. This algorithm works based on the concept that documents are the string of words and not the collection of words and hence works on proximity information of words. STC uses suffixes tree structures to identify set of documents that shares common phrase and term efficiently, and takes this information for creating cluster and presents their content concisely to the user. STC mainly works on four logical steps: 1. Document "cleaning"; 2.

Construction of a generalized suffix tree; 3 Identification of base cluster; 4.Combination of base cluster into cluster. In the process of clustering, the important algorithm is the Suffix tree clustering. This algorithm has linear complexity as O(n) which makes this algorithm as the best clustering algorithm since its response time is minimum. This low complexity returns fast searching in suffix tree; hence this algorithm is often used for online clustering as well as web document clustering. [32]

The investigation of the clustering issue precedes its application to the text domains. The hierarchical organization of document into coherent category is most useful for organized browsing of the set of document. Clustering technique provides a coherent summary of the group in the form of cluster-digest [83] or word-cluster [17, 18] that is used for providing summary insight into the entire content of the fundamental corpus. As the text document is taken from naturally high-dimensional domains, it is helpful to see the issue in two ways where important cluster of word can be located and used for identifying cluster of document [33].

In various text classifications application, it is interesting to consider each document as string of character than considering as bag of word. Earlier researches mainly concentrated on various variants of generative Markov chain model. Discriminative machine learning techniques such as Support Vector Machine (SVM) successfully classifies text with word features. But they are not effective or efficient in straightforwardly considering the entire substrings available in the corpus as feature. This research proposes to partition the substring into statistical equivalence group and select the significant group (in the statistical sense) as feature (named key-substring-group feature) for classification of text. In specific it proposes a suffix tree based algorithm that extracts linear time features (based on the complete characters present in the corpus) [34].

This work presents a topic discovery system that aims to disclose the implicit knowledge available in the new stream. The implicit knowledge is given as hierarchy of topics/subtopic which has the set of document related to them and the summary that is retrieved from the given document. Summary thus constructed are helpful for browsing and selecting topic of interest from the developed hierarchy. Proposed system has a new incremental hierarchical clustering algorithm that mergers both partitional and agglomerative approach with the major advantages [35].

In text mining research, familiar methods use often bag-of-words model that represents documents as vectors. This method ignores the word sequences information and hence the good clustering results are limited to certain special fields. This work proposes a novel similarity measure with respect to suffix tree model of text document. This functions, as analyzing the word sequence information and computing the

similarity between the text documents of corpus by using a suffix tree similarity that merges with TF-IDF weighting technique [36].

5. Conclusion

In this review article , over the past decade the text mining , and classification algorithms used for web page classification are analyzed and presented It is suggested that semantic mining model is incorporated in the web page search , will improve in the quality of the identification web pages. The cluster based approach will provides efficient searching in web pages when comparing with other traditional approaches.

References

1. Oikonomakou, Nora, and MichalisVazirgiannis, "A Review of Web Document Clustering Approaches." Data Mining and Knowledge Discovery Handbook, 2005.
2. Shilpa Dang, Peerzada Hamid Ahmad , Text Mining: Techniques and its Application, IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4, November 2014 ISSN (Online): 2348-0866
3. A. Henriksson, Jing Zhao, Lars Asker, Henrik Boström , "Detecting adverse drug events with multiple representations of clinical measurements " IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014, pp. 536-543.
4. A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and ˇ M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.
5. P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on. IEEE, 2015, pp. 634–638.
6. Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45
7. N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012
8. E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce, and J. T. Fernan- ˜ dez, "Searching research papers using clustering and text mining," in Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on. IEEE, 2013, pp. 78–81.
9. Shehata, S., Karray, F., Kamel, M.: An efficient concept-based min-ing model for enhancing text clustering. IEEE Trans. Knowl. Data Eng. 22(10), 1360–1371 (2010)

10. S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model,"Proc.Sixth IEEE Int"l Conf. Data Mining (ICDM), 2006.
11. N.N. Myat ; Khin Haymar Saw Hla,A combined approach of formal concept analysis and text mining for concept based document clustering, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05).
12 Andreas Hotho, Gerd Stumme, Conceptual Clustering of Text Clusters, Proceedings of FGML workshop, 2002.
13 CaiyanJia ,Matthew B.Carson, XiaoyangWang ,JianYu, Concept decompositions for short text clustering by identifying word communities, Pattern Recognition, 2018 – Elsevier.
14. Yang, H., Wang, Z., Xu, H. (2015) On-line text mining and recom-mendation based on ontology and implied sentiment inclination. In 2015 17th International Conference on Advanced Communication Technology (ICACT), pp. 613–617. IEEE.
15. Reshma R, Vinai George Biju, A Semantic Based Approach For Text Clustering Using An Advanced Concept-Based Mining Model, International Journal Of Advanced Computing And Electronics Technology, Volume-2,Issue-2,2015.
16. S.Saranya1 and S.Logeswari, A Semantic Model for Concept Based Clustering, International Journal of Innovative Research in Computer and Communication Engineering, 2017.
17. YongpingDu , JingxuanLiu ,WeimaoKe ,XuemeiGong. Hierarchy constructionand text classification based on the relaxation strategy and leastinformation model. Expert Systems with Applications , Volume 100, 15 June 2018, Pages 157-164.
18. Hao Peng , Jianxin Li , Yu He , Yaopeng Liu , Mengjiao Bao, Lihong Wang , Yangqiu Song , and Qiang Yang ,Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN 2018 IW3C2 (International World Wide Web Conference Committee) .
19. Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar, Text mining and ontologies in biomedicine: Making sense of raw text , Henry Stewart publications 1467-5463. Briefings in Bioinformatics. vol 6. no 3. 239–251. september 2005.
20. Saeed Hassanpou, Siddharth Taduri , Semantics-based Text Mining of Biomedical Concepts in Scientific Publications, CS224 Final Project Report, June 4, 2010.
21. Le Yanga , Zhongbin Ouyanga , Yong Shia.b, A Modified Clustering Method Based on Self-Organizing Maps and Its Applications, International Conference on Computational Science, ICCS 2012
22. Matharage, S., Alahakoon, D.: Enhancing GSOM text clustering with latent semantic analysis. In 2010 Fifth International Con-ference on Information and Automation for Sustainability, pp. 441–446. IEEE (2010)

23. Kohonen, T., 1990. The self-organizing map. Proceedings of IEEE, 78(9), pp. 1464-1480.
24. Hsin-Chang Yang, Chung-Hong Lee, A novel self-organizing map algorithm for text mining, 2010 International Conference on System Science and Engineering, Pages: 417 - 420
25. Yi Ding, Xian Fu, The Research of Text Mining Based on Self-Organizing Maps, 2012 International Workshop on Information and Electronics Engineering (IWIEE), Elsevier.
26. Hsin-Chang Yang, Chung-Hong Lee, A Novel Self-Organizing Map for Text Document Organization, 2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications
27. Hsin-Chang Yang, Chung-Hong Lee, and Kuo-Lung Ke ,TOSOM: A Topic-Oriented Self-Organizing Map for Text Organization , International Journal of Computer and Information Engineering Vol:4, No:5, 2010.
28. Wei Song, Soon Cheol Park, Genetic algorithm for text clustering based on latent semantic indexing, Computers and Mathematics with Applications 57 (2009) 1901–1907, Elsevier.
29. S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil, Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution, International Journal of Mathematical and Computational Sciences Vol:2, No:1, 2008.
30. Sung-Sam Hong , Wanhee Lee, and Myung-Mook Han, The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification, Int. J. Advance Soft Compu. Appl, Vol. 7, No. 1, March 2015 ISSN 2074-8523.
31. B. G. Yoon and J. S. Yang, "Applications of Genetic Algorithm and Text Mining on Technology Innovation", Applied Mechanics and Materials, Vol. 145, pp. 287-291, 2012.
32. Bharadwaj, D., Shukla, S.: Text mining technique using genetic algorithm. In: Proceedings on International Conference on Advances in Computer Application (ICACA), 2013.
33. Milos Ilic , Petar Spalevic, Mladen Veinovic, Suffix Tree Clustering - Data mining algorithm, ERK'2014, Portorož, B:15-18
34. Charu C. Aggarwal, and C.X. Zhai (eds.),A Survey of Text Clustering Algorithms, Mining Text Data, DOI 10.1007/978-1-4614-3223-4_4, Springer Science+Business Media, LLC 2012
35. Dell Zhang, Wee Sun Lee , Extracting key-substring-group features for text classification , ,Published 2006 in KDD ,Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining , Pages 474-483.
36. AuroraPons-Porrata, RafaelBerlanga-Llavori, JoséRuiz-Shulcloper, Information Processing & Management , Volume 43, Issue 3, May 2007, Pages 752-768.