ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse71832019.pdf
https://doi.org/10.30534/ijatcse/2019/71832019

Review and Comparative Analysis of Topic Identification Techniques

Deepti Sehrawat[1], Nasib Singh Gill[2]
[1,2]Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana (India)
dips.scorpio@gmail.com, nasibsgill@gmail.com

ABSTRACT

Topic identification is an area of data mining that finds common text/ themes from several documents. It is a data summarization technique that helps to summarize documents. This area is of great interest among researchers as its applications in the real world are very wide. This paper presents a review of topic identification techniques. Existing solutions include text clustering, latent semantic approach, probabilistic latent semantics approach, latent Dirichlet allocation approach, association rule-based approaches, document clustering, and soft computing approach. Soft computing techniques including fuzzy logic, neural networks, support vector machine, ant colony optimization, swarm optimization, and their hybrid approaches provide a good solution for text clustering. This paper presents a comparative study of different text mining techniques with their strengths and weaknesses. A future dimension is also proposed to develop a hybrid approach for topic identification using different techniques.

Key words : Topic identification, cluster labeling, label identification, document summarization, K-means, LSA and LDA.

1. INTRODUCTION

The growing size of unstructured text data on the World Wide Web has raised the need for understanding the ways to extract meaningful information from this large scale text. Mining topics from large documents and collection of documents including research publications, technical reports or summarizing large documents is very helpful. Topic identification, also known as “topic discovery”, “topic finding”, “label identification”, “cluster labeling”, “category labeling”, or “document summarization” . The main aim is to find common patterns from a number of documents or to summarize a single large document. The extracted and summarized text carries consistent semantic meaning. This text helps in future casting or trend forecasting.

By examining various publications for a period of time it would be beneficial for many real-world applications. We can discover more prominent areas and also predict their future trends. Authors in [1] studied several methods for stock market prediction using data mining and Genetic Algorithm. It is clear from their study that stock prediction is not an easy task and for accurate forecasting, numerous influences should be considered. Furthermore, data mining can benefit to model the research direction, helps to predict future trends of the industry, forecasting stock market trends, emerging trends of new products by examining reviews of the customers [2] [3]. The problem is twofold for all these applications, these are:

(i) First, topic mining by summarizing and generating meaningful topics from a set of documents; and
(ii) Second, forecasting of topic trend in the future.

A number of solutions exist to find meaningful topics and summarizing a set of documents. Existing solutions generally use text clustering, hierarchical clustering, association rule based data mining, latent semantic approach and a number of other techniques for topic discovery [4]. Context information is not considered in words, sentences or paragraphs; every single document corresponds to an instance. An efficient AI tool for text summarization is presented in [5]. A set of topics/patterns are first discovered from the documents and then the temporal correlation is used to predict their popularity. The above-mentioned problem is a time series forecasting issue [6], as the popularity in different years can be considered as time series data that can be predicted using time-series forecasting. Topic forecasting is not as simple as it seems to be. There is a correlation among research topics which may be considered as strongly correlated, weakly correlated, inversely correlated, positively correlated, negatively correlated, or there may be no correlation, thus complicating the process of topic forecasting.

There are two ways provided by the modern search engines to retrieve useful information: the first approach requires a cluster labeling that allows documents browsing based on human-made subjects. The second approach is the classification that retrieves documents by word query [7]. In order to assign the documents to a particular browsing, category classification is generally used. Authors in [8] proposed a text summarization algorithm for Telugu e-newspaper. This paper presents numerous existing text mining techniques along with their advantages and limitations.

Organization of the rest of the paper is as follows: Section 2 covers a brief introduction of existing techniques which can be used for topic identification. In Section 3, we have presented related work carried out over the years by different researchers along with their pros and cons. Section 4 concludes the paper.

2. TOPIC IDENTIFICATION TECHNIQUES

Various techniques are used for extracting useful information from documents; a few techniques are explained as under:

2.1 Document Clustering

Data distribution information is extracted from unstructured textual data with the help of document clustering; it requires no prior knowledge [9]. Partitioning like a distance-based approach is used to cluster documents in groups. Due to the increasing size of the data simple document clustering is not enough to extract meaningful information. As a result, there is a need for automatic topic discovery from a set of documents. A topical space navigation method that group related documents into subcategories is proposed by Sahami [10] which uses hierarchical clustering. Document clustering using Self Organizing Maps to automatically cluster unstructured web documents was proposed in [11]. Different variants of K-Means for documents clustering are also implemented and it is the most popular clustering algorithm which has improved efficiency and accuracy.

2.2 Association Rule-Based Approaches

An association rule has a set of items on its left side known as antecedents and set of items on the right side known as consequents, it has a form $X \rightarrow Y$. within a corpus and a number of association rules are discovered that implies the presence of an article term if another term is present. Authors in [12] proposed a topic discovery method based upon association rule mining. In this approach, each document is represented in the vector format that can be taken as a transaction. Association rule mining is then applied to the transaction set that extracts the patterns.

2.3 Latent Semantics Approaches

When we have to find the similarity between two documents then the Latent Semantics approach proves to be the best. It is very useful for cases even when there is no similarity (no common words) between two documents. In this proposed approach, a low dimensional space of terms is created that can be termed as "latent" because it is a combination of the terms, not a single term to which it relates. LSA discover the semantic similarity between documents and between topics [13].

2.4 Probabilistic Latent Semantics Approaches

Probabilistic Latent Semantics Analysis (pLSA) is an addition of the probabilistic model to the earlier latent semantic analysis. It is a new statistical technique that analyses co-occurrence data and two-mode data. pLSA approach can be used in numerous applications including machine learning, natural language processing, filtering, information retrieval, and related areas. The proposed pLSA approach provides a better statistical base to the earlier similar kind LSA approach. pLSA can also be used to find topics by using historical data [14].

2.5 Latent Dirichlet Allocation Approaches

Authors in [15] proposed a model that addresses the limitations of pLSA and gives a new model known as Latent Dirichlet Allocation (LDA) model. The new proposed model is similar to the earlier pLSA model with some difference that the newly proposed LDA model assumes Dirichlet prior to topic distribution [16]. LDA approach is among the family of Bayesian nonparametric approaches that use unsupervised learning style. Documents in this approach are represented in vector space to cluster the data which are then treated as topics. Authors in [17] used an LDA approach to generate a topic hierarchy of the documents.

2.6 Soft Computing Approaches

Current soft computing techniques like fuzzy logic, genetic algorithm, neural network, Self-Organizing map, Support Vector Machine, Differential Evolution and a number of hybrid approaches have recently been used for data mining problems. Soft computing techniques are mainly used to evaluate the web applications according to the new characteristic, so as to achieve intelligence of the web [18]. These approaches attempt to provide fairly accurate solutions at a low cost, thereby speeding up the process [19]. These nature inspired soft computing techniques when combined with other algorithms results in an optimized approach to produce more meaningful, accurate, efficient and useful results. Particle Swarm Optimization (PSO) is used for document clustering that utilizes swarm-based algorithms. Authors in [20] proposed a model using SQL relational databases with Naïve Bayes classifier, a machine learning algorithm. Other techniques include Self Organizing Map [11], Genetic Algorithm [21], Particle Swarm Optimization [22], lexical chain method [23] and Differential Evolution (DE) [24].

3. REVIEW OF LITERATURE

In the past two decades, topic identification has emerged as an important approach in the field of data mining. At present, a number of solutions exist to find meaningful topics and summarizing a set of documents from the web. Existing solutions frequently use text clustering, hierarchical clustering, association rule based mining, Latent Semantic Models, and various soft computing approaches. The following section presents some latest research in the field of topic identification and future forecasting.

Eder V´azquez *et al.* (2018) presented a genetic algorithm based automatic extractive text summarization method. The proposed approach creates a more relevant and less redundant summary by using measures like coverage, sentence positioning, title similarity and length of the sentence. It uses the term length for creating new sentence length weighting [25].

Kun Dong *et al.* (2018) research offers an integrated method for interdisciplinary topic identification from the scientific literature. The proposed approach integrates co-occurrence network analysis method with high-TI terms analysis and burst detection. There are four steps in the proposed method, these are Data collection, data processing (data cleaning by using Thomson Data Analyzer (TDA) tool and cleaning by using fuzzy matching, terms merging and terms clustering which is based on principal component analysis), topic identification and topic prediction. This method can be used for knowledge discovery and in the future can also be used for scientific measurement. There are some limitations of this method such as it considers only a few relationships between topics and is only limited to the late fusion stage of integration [26].

Xingxing Zhang *et al.* (2018) proposed a latent variable extractive model based on a neural extractive summarization model and sentence compression model. To conclude gold summaries this model consider the sentences as binary variables and those sentences are used which have activated latent variable (i.e., ones) [27].

Krushna Sharma *et al.* (2018) proposed an automatic classifier method by using deep learning methods (Recurrent Neural Network with Convolutional Neural Network). It improves the classification process by using a different type of data for training and propose relevant documents. There are three steps in the proposed method, these are data pre-processing, text summarization by using a graph-based method, and text classification with the help of natural language processing and machine learning techniques [28].

Qingyu Zhou *et al.* (2018) presented a framework for extractive document summarization using neural networks. The proposed approach used the hierarchical encoder for the representation of the sentences from which summary is generated. It integrates the selection strategy with the sentence scoring model in a single phase. Whenever a sentence is selected, the model scores the sentences according to the present extraction state and partial output summary [29].

Fei Liu *et al.* (2018) presents a data-driven, trainable and wide domain framework for abstractive summarization. First, the source text is given by an AMR graph (Abstract Meaning Representation) which is then converted to a summary graph. The summary is then generated from this summary graph. It makes use of structured prediction algorithm for graph to graph conversion. The semantic graphs of input are transformed into a single summary semantic graph [30].

M. Anjaneyulu *et al.* (2018) proposed an unsupervised model for extractive summarization from multiple documents. This approach uses Latent Semantic Analysis (LSA) and the redundant sentences are removed from the multiple documents by analyzing their semantic and syntactic information. The proposed approach is a two-phase process. In the first phase, those sentences are identified which are representing the topic of the document and then in the second phase summary is generated by removing redundant sentences and combine the remaining ones. For removing redundant sentences statistical, semantic, lexical and syntactic features of the sentences are studied [31].

J. Lin *et al.* (2017) proposed a supervised method for topic refined method for competitive perspective identification. The proposed method refines perception classifiers with the document-topic distributions mined from texts. Another proposed framework which is user based bootstrapping method, semi-supervised in nature minimizes human labor in data annotation. Good quality classified texts can be selected from unlabeled online corpus [32].

H. Wen Jing *et al.* (2017) proposed a topic computation model for documents to satisfy range queries. To identify documents, indexes are created then for documents subsets topic models are pre-computed. The proposed approach gives a better and improved solution to identify correct sets of documents. The presented approach can also be parallelized with ease, including the building of the indexes, the pre-computing of the topic models, and the query processing [33].

Jose L. Hurtado *et al.* (2016) proposed an ensemble forecasting model, an approach that finds from documents meaningful topics and also predicts future trends by future forecasting. In the proposed approach first sentences in documents are converted to a transaction format which is then applied to association rule mining to find a frequent pattern and in turn discover topics from the documents. To refine the frequent patterns into meaningful topics set inclusion/exclusion operations were then performed. Temporal frequency was recorded for the extracted documents, for some years that describe the temporal evolution of the topic. Correlation between topics was found by using correlation coefficient [34].

Shrikant Malviya and Uma Shanker Tiwary (2016) proposed a novel method of knowledge representation at various levels like word level, sentence level, and paragraph level. Document summarization of research papers was carried out using a Bayesian network based approach with various information retrieval techniques. In order to accomplish this task, a semantic knowledge tree was built to store and represent retrieved information. The score estimation was done in a bottom-up manner. The proposed MDS (Multi Document Summarization) is a kind of query-based extractive summarizer. It retrieves the relevant content from the knowledge tree based on its score [35]. Better summaries can be given by using spatial information based retrieval content.

Kartik Asooja *et al.* (2016) evaluated a basic approach for forecasting emerging trends. The approach used regression models for predicting the distribution of keywords. A time series data set of topics and popularity was generated. They computed Spearman rank correlation and average root means square [36]. Given approach is not diverse, data collected from only one conference, correlation is also not done, forecasting can be included to improve the performance of the proposed work.

N K Nagwani (2015) presented a text summarizer for a large text collection based on the MapReduce framework. It is a multi-document summarizer that uses a semantic similarity based clustering technique. The proposed approach has four stages: First stage performs document clustering using K-Means on the MapReduce framework and collects the text information from each document in aggregate. The second stage generates topics from all text document clusters using the Latent Dirichlet Allocation technique on collective information. The third stage computes semantic similar terms for every topic term using WordNet Java API. The last stage extracts sentences from individual documents using parsing techniques [37].

Yogita K. Desai and Prakash P. Rokade (2015) proposed a system based on cross-document structure theory (CST) that generates a relevant summary from documents by identifying CST relationship among multi-documents. It has revealed from their research that there are 24 types of CST relations that exist among the documents with the same domain. To accomplish the task in the proposed system first documents are preprocessed and then features are extracted. Then, the final summary is generated by defining the threshold value for the score of the sentence [38]. In the proposed approach topic correlation and future, forecasting is not done.

Pankaj Bhole and A. J. Agrawal (2014) proposed an application to automatically generate short news from a large article.It involves following steps: Pre-processing, Sentence clustering (using K-means), Cluster ordering (based on the number of important words it contains), Representative sentence selection (ranking for most informative sentences) and summary generation (ordered list of sentences from clusters) [39]. The proposed technique is for a single document which summarizes the text and does not consider multiple documents, also future forecasting is not done and there is no correlation.

R. Priyadarshini and Latha Tamilselvan (2014) presented a clustering approach based on K-Means. The proposed approach is based on keywords and implemented in the mahout component of Hadoop. User queries are used to retrieve the files and graphs are displayed to give clustered documents from MongoDB. Similarity measures along with concept matching were used to perform semantic document analysis. To cluster, the documents NLP tool was used [40]. A threshold for a number of cluster formation is not fixed.

Pengtao Xie and Eric P.Xing (2013) proposed a multi-grain clustering topic model. It performs document clustering and modeling simultaneously. Experiments on two datasets demonstrate the fact that these two tasks are closely related and can mutually promote each other. Their experiments on document clustering reveal that through topic modeling, clustering performance can be improved by finding more coherent topics and can differentiate topics into group-specific ones and group-independent ones [41].

Stuti Karol and Veenu Mangat (2013) proposed two algorithms KPSO and FCPSO using a hybrid approach to cluster text documents. The first approach was the KPSO algorithm by hybridizing two popular K-Means and PSO. KPSO begins with an initial module of K-Means and after obtaining its results PSO was applied on the results of the previous step that were generated by K-Means. KPSO being a hard clustering algorithm generate disjoint clusters. The second FCPSO soft clustering algorithm hybrid FCM and PSO. It first applies the Fuzzy C-Means algorithm on the data to obtain initial clusters followed by PSO to give optimum clusters. Clusters are created from data points having a degree of membership in each cluster. Their results have proven that FCPSO is as far better than KPSO [42]. The proposed approach can be improved for better convergence and Accuracy correlation and future forecasting can also be included for a better approach.

Rocio Chongtay (2013) presents a design proposal of the MIDIGESTS tools for summarization-based learning. The presented approach attempts to streamline the focus on the specific learning topics to support effective ubiquitous learning. Ubiquitous computing supports learning at anytime from anywhere. A combination of NLP and Interactive Systems was employed in order to reduce information

overload. This approach allows automatic summarization, visualizations, and rich interaction [43]. Performance can be improved further by techniques such as semantic enrichment with the help of Linked Open Data.

Ruben Sipos *et al.* (2012) proposed a sub-modular framework to summarize temporal data. A citation structure is not required, the proposed approach relied only on words to infer influence across time. The proposed approach can be applied to any time-stamped textual document collection. Monotone sub-modular functions were used to optimize the objective through an efficient greedy algorithm. It uses the K-neighbour approach and sub-modular word coverage is applied for summarizing text [44].

Kumar Shubhankar *et al.* (2011) proposed an improved iterative PageRank algorithm for topic identification and clustering the research papers into identified topics. The proposed approach works by forming closed frequent keywords-sets extracted from the research paper titles to identify the topics. It finds within a topic cluster a ranked list of research papers. The proposed approach uses hierarchical clustering and modified page rank algorithm techniques for clustering of research papers [45]. Topic correlation and future forecasting are not done which can be implemented in the proposed work to improve efficiency.

Liangjie Hong *et al.* (2011) proposed a model (collection model) to automatically analyze multiple correlated temporal text streams. The proposed approach is an enhancement over LDA that has applied Collapsed Gibbs sampling in order to find hidden patterns [46]. The proposed approach is a time-dependent model that can be used for multiple text streams. The proposed model can be further improved if Bayesian non-parametric techniques can be used to find a number of topics automatically from the dataset.

Weiwei Cui *et al.* (2011) presented a system named TextFlow to help users to visually analyze and evolve topics at different granularities. The proposed model enables flawless communication among interactive visualization and topic mining. The proposed system has few limitations like it is not so easy to select meaningful threads and cluttered visualization, when using more keywords [47].

Sandha Harabagiu and Finley Lacatusu (2010) proposed two topic representations to enhance content selection and quality of information ordering for Multiple Document Selection (MDS). An evaluation of 40 MDS methods was presented which are based on structured sets of topic themes. Evaluation of the summarization method was done using ROUGE automatic scoring packages and the manual Pyramid evaluation method [48].

David J. Newman and Sharon Block (2006) in their work in [14] presented a method that finds topics in the Pennsylvania Gazette, a newspaper in early America. The proposed model, Probabilistic Latent Semantic Analysis (pLSA) can be used for historical research. Their experimental results have proven that pLSA proves to be a good method for historical research by efficiently computing significant topics form large texts. It also allows the mixing of a number of topics within a particular topic [14]. Trend analysis is not included in the proposed approached for which it can be enhanced.

Qiaozhu Ma and Cheng Xiang Zha (2005) presented an unsupervised framework that helps to discover evolution theme patterns from the text. The proposed approach used clustering, SVM classification, and Kullback Leibler Divergence probabilistic approach. In the proposed approach authors do not apply hierarchical clustering and stemming [49].

Bing Liu *et al.* (2003) used agglomerative clustering and PERL language for topic mining and extracting definitions on the web [50]. The proposed approach does not support meta-data and parsing of web pages is very slow.

A number of text summarization techniques are available these days. Table 1 enlists major techniques used in recent works along with their advantages and limitations. These limitations can be taken as a future research area.

Table 1: Important Publications and techniques used in their approach

Pub. Year	Title of Paper	Authors	Techniques Used	Advantages
2018	"Sentence features relevance for extractive text summarization using genetic algorithms" [25]	E. V´azquez, R. Arnulfo, G. Hern´andez and Y. Ledeneva	Genetic Algorithm	Create more relevant and less redundant summary.
2018	"An integrated method for interdisciplinary topic identification and prediction: a case study on information science and library science"	K. Dong, H. Xu, R. Luo, L. Wei, and S. Fang	Integrated method; uses co-occurrence network analysis with high-TI terms analysis and bursts detection.	This method can be used for knowledge discovery and in the future can also be used for scientific measurement. For interdisciplinary topic

	[26]			identification from the scientific literature.
2018	"Neural Latent Extractive Document Summarization" [27]	X. Zhang, M. Lapata, F. Wei and M. Zhou	Neural Network	It improves over a strong extractive baseline trained on heuristically approximated labels.
2018	"Automated Document Summarization and Classification Using DeepLearning" [28]	K.Sharma, A. Gaikwad, S. Patil, P. Kumar, and D.P. Salapurkar	Deep Learning; Recurrent Neural Network with Convolutional Neural Network	It gives more accurate and abstractive results.
2018	"Neural Document Summarization by JointlyLearning to Score and Select Sentences" [29]	Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao	Neural Network	It integrates the selection strategy with the sentence scoring model in a single phase.
2018	"Toward Abstractive Summarization Using Semantic Representations" [30]	F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith	Semantic Graphs	Gives abstractive summarization of input data
2018	"Topic Oriented Multi-document Summarization Using LSA, Syntactic and Semantic Features" [31]	M. Anjaneyulu, S. S. V. N. Sarma, P. Vijaya Pal Reddy, K. Prem Chander and S. Nagaprasad	Latent Semantic Analysis	Useful for multi-documents summarization
2017	"Topic and user based refinement for competitive perspective identification" [32]	J. Lin, W. Mao, and D. Zeng	Supervised Refined Perception Classifiers with document-topic distributions mined from texts. User-based bootstrapping method, semi-supervised in nature.	Good quality classified texts are selected from unlabeled online texts.
2017	"Accelerating Topic Exploration of Multi-Dimensional Documents" [3]	H. Wen-Jing, L. You and L. Z. Qi	Indexes are created then for documents subsets topic models are pre-computed.	Gives a better and improved solution to identify correct sets of documents. Can be parallelized and pre-computing is possible.
2016	"Topic discovery and future trend forecasting for texts" [34]	J. L. Hurtado, A. Agarwal, and X. Zhu	Association rule mining and Ensemble topic forecasting	Has better performance
2016	"Knowledge Based Summarization and Document Generation using Bayesian Network" [35]	S. Malviya and U. S. Tiwary	Bayesian Network	Generates around 50% of relevance of the summary.
2016	"Forecasting Emerging Trends from Scientific Literature based on Keyword Extraction and Prediction" [36]	K. Asooja, G. Bordea, G. Vulcu, and P. Buitelaar	Regression models, Spearman Rank Correlation, Average Root Mean Square	Future Forecasting is done.
2015	"Summarizing large text collection using topic modeling and clustering based on MapReduce framework" [37]	N. K. Nagwani	K-Means clustering and Latent Dirichlet Allocation	Faster and powerful for analyzing Big Text Data.

4. CONCLUSION

So far our literature survey shows that the authors are interested in topic discovery using different techniques including classification, association rule, latent semantics, probabilistic latent semantics, latent Dirichlet allocation, and various soft computing approaches. Soft computing paradigms which are based on meta-heuristic are very less exploited in the field of text mining. The hybrid approach will enhance the effectiveness of text mining. The study of the literature shows that various techniques exist for topic identification. Some of them are not fully optimized and can be explored further. Some are computationally complex and

expensive. A combination of soft computing techniques is also used for topic identification but for topic mining and topic future forecasting they have some limitations. Only hybrid soft computing approaches and topic mining techniques can provide a promising solution.

REFERENCES

1. M. Tawarish, and K. Satyanarayana. An enabling technique analysis in Data Mining for Stock Market trend by Approaching Genetic Algorithm, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1, pp. 27-33, 2019. https://doi.org/10.30534/ijatcse/2019/06812019
2. C. Tucker, and H. M. Kim. **Predicting emerging product design trend by mining publicly available customer review data,** in *Proc. of international conference on engineering design, ICED 11,* 2011, pp. 43-52.
3. R. Schumaker, and H. Chen. **Textual analysis of stock market prediction using breaking financial news: The azfin text system,** *ACM Trans Inf Syst, vol. 27, no. 2, pp. 1-19,* 2012. https://doi.org/10.1145/1462198.1462204
4. D. M. Blei. **Introduction to probabilistic topic models**, *Commun ACM,* vol. 55, no. 4, pp.77-84, 2012.
5. N. Chatterjee, and S. Mohan. **Extraction-Based Single-Document Summarization Using Random Indexing,** in *Proc. 19th IEEE International Conference on Tools with Artificial Intelligence*, 2007, pp. 448-455. https://doi.org/10.1109/ICTAI.2007.28
6. T. C. Fu. **A review on time series data mining,** *Eng Appl Artif Intell., Elsevier*, vol. 24, no. 1, pp. 174-181, 2011. https://doi.org/10.1016/j.engappai.2010.09.007
7. W. Mettrop, and P. Nieuwenhuysen. **Internet search engines—fluctuations in document accessibility**, *J Doc*, vol. 57, no. 5, pp. 623-651, 2001. https://doi.org/10.1108/EUM0000000007096
8. R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra. **Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers**, in *Proc. Smart Computing and Informatics, Smart Innovation, Systems and Technologies, Springer, Singapore,* vol. 77, pp. 555-564, 2018. https://doi.org/10.1007/978-981-10-5544-7_54
9. A. C. Joshi, V. R. Padghan, J. R. Vyawahare, and S. P. Saner. **Enforcing document clustering for forensic analysis using weighted matrix method (wmm),** *International Journal of Advance Research in Computer Science and Management Studies*, vol. 3, no. 3, pp. 88-94, 2015.
10. M. Sahami. **Using machine learning to improve information access**. *Technical report, Stanford University*. 1998.
11. R. Freeman, H. Yin, and N. M. Allinson. **Self-Organising Maps for Tree View Based Hierarchical Document Clustering**, in *Proceedings of the IEEE IJCNN'02, Honolulu, Hawaii,* 2002, pp. 1906-1911. https://doi.org/10.1109/IJCNN.2002.1007810
12. P. C. Wong, P. Whitney, and J. Thomas. **Visualizing association rules for text mining,** in *Proc. IEEE Symposium on Information Visualization,* 1999. (Info Vis '99). https://doi.org/10.1109/INFVIS.1999.801866
13. T. K. Landauer TK, P. W. Foltz, and D. Laham. **An Introduction to latent semantic analysis,** *Discourse Processes*, vol. 25, pp. 259-284, 1998. https://doi.org/10.1080/01638539809545028
14. D. J. Newman, and S. Block. **Probabilistic topic decomposition of an eighteenth-century american newspaper,** *J Am Soc Inf Sci Technol*, vol. 57, no. 6, pp. 753-767, 2006. https://doi.org/10.1002/asi.20342
15. D. M. Blei, A. Y. Ng, and M. I. Jordan. **Latent dirichlet allocation,** *J Mach Learn Res*, pp. 993–1022, 2003.
16. A. K. Jain. **Data clustering: 50 years beyond k-means**, *Pattern Recognit Lett*, vol. 31, no. 8, pp. 651-666, 2010. Award winning papers from the 19th international conference on pattern recognition (ICPR). https://doi.org/10.1016/j.patrec.2009.09.011
17. N. Akhtar, H. Javed, and T. Ahmad. **Hierarchical Summarization of Text Documents Using Topic Modeling and Formal Concept Analysis**, in Proc. *Data Management, Analytics and Innovation*, Springer, Singapore, 2019, pp. 21-33. https://doi.org/10.1007/978-981-13-1274-8_2
18. A. R. Deshmukh, and S. R. Gupta. **Data mining based soft computing methods for web intelligence**, *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 3, no. 3, pp. 376-382, 2014.
19. Y. K. Mathur, and A. Nand. **Soft Computing Techniques and its Impact in Data Mining**, *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 8, pp. 658-662, 2014.
20. I. S. Makki, and F. Alqurashi. **An Adaptive Model for Knowledge Mining in Databases "EMO_MINE" for Tweets Emotions Classification**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 7, no. 3, pp. 52-60, 2018. https://doi.org/10.30534/ijatcse/2018/04732018
21. K. Premlatha, and A. M. Natrajan. **Discrete PSO with GA operators for Document Clustering**, *Int. J. Recent Trends Eng*, vol. 1, no. 1, pp. 20-24, 2009.
22. J. Kennedy. **Particle Swarm Optimization.** *Encyclopedia of Machine Learning*, Springer, Boston, MA, pp. 32-45, 2011. https://doi.org/10.1007/978-0-387-30164-8_630
23. H. M. Lynn, C. Choi, and P. Kim. **An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms**, *Soft Computing*, vol. 22, no. 12, pp. 4013-4023, 2018. https://doi.org/10.1007/s00500-017-2612-9
24. A. Abraham, S. Das, and A. Konar. **Document Clustering using Differential Evolution**, in *Proc.*

International Conference on Evolutionary Computation, IEEE, 2006, pp. 1784-1791. https://doi.org/10.1109/CEC.2006.1688523

25. V. Eder, A. G. H. René, and L. Yulia. **Sentence features relevance for extractive text summarization using genetic algorithms,** *Journal of Intelligent & Fuzzy Systems*, vol.35, no. 1, pp.353-365, 2018. 10.3233/JIFS-169594
26. K. Dong, H. Xu, R. Luo, L. Wei, and S. Fang. **An integrated method for interdisciplinary topic identification and prediction: a case study on information science and library science**, Scientometrics, vol. 115, no. 2, pp. 849-868, 2018. https://doi.org/10.1007/s11192-018-2694-x
27. X. Zhang, M. Lapata, F. Wei, and M. Zhou. **Neural latent extractive document summarization**, *arXiv preprint arXiv:1808.07187*, 2018. https://arxiv.org/abs/1808.07187v2
28. K. Sharma, A. Gaikwad, S. Patil, P. Kumar, and D. P. Salapurkar. **Automated Document Summarization and Classification Using Deep Learning**, *International Research Journal of Engineering and Technology*, 2018; 5 (06).
29. Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao. **Neural document summarization by jointly learning to score and select sentences**, *arXiv preprint arXiv:1807.02305*. 2018. https://arxiv.org/abs/1807.02305v1
30. F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. **Toward abstractive summarization using semantic representations**. *arXiv preprint arXiv:1805.10399*. 2018. https://arxiv.org/abs/1805.10399v1
31. M. Anjaneyulu, S. S. Sarma, P. V. Reddy, K. P. Chander, and S. Nagaprasad. **Topic Oriented Multi-document Summarization Using LSA, Syntactic and Semantic Features**, in *International Conference on Innovative Computing and Communications*, Springer, Singapore, 2019, pp. 487-502. https://doi.org/10.1007/978-981-13-2354-6_50
32. J. Lin, W. Mao and D. Zeng. **Topic and user based refinement for competitive perspective identification**, in *Proc. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing*, 2017, pp. 131-133. https://doi.org/10.1109/ISI.2017.8004888
33. H. Wen-Jing, L. You and L. Z. Qi. **Accelerating Topic Exploration of Multi-Dimensional Documents**, in *Proc. IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lake Buena Vista, FL*, 2017, pp. 1520-1527. https://doi.org/10.1109/IPDPSW.2017.113
34. J. L. Hurtado, A. Agarwal, and X. Zhu. **Topic discovery and future trend forecasting for texts**, *J Big Data. Springer Open*, 2016, pp. 3-7. https://doi.org/10.1186/s40537-016-0039-2
35. S. Malviya, and U. M. Tiwary. **Knowledge Based Summarization and Document Generation using Bayesian Network**, in *Proc. of Twelfth International Multi-Conference on Information Processing*, Elsevier. 2016, pp. 333-340. https://doi.org/10.1016/j.procs.2016.06.080
36. K. Asooja, G. Bordea, G. Vulcu, and P. Buitelaar. **Forecasting Emerging Trends from Scientific Literature based on Keyword Extraction and Prediction**, in *Proc. 10th Language Resource and Evaluation Conference (LREC), Portoroz, Slovenia*, 2016.
37. N. K. Nagwani. **Summarizing large text collection using topic modeling and clustering based on MapReduce framework**, *Journal of Big Data, Springer*, pp. 2-6, 2015. https://doi.org/10.1186/s40537-015-0020-5
38. Y. K. Desai, and P. P. Rokade. **Multi Document Summarization: Approaches and Future Scope**, *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 5, no. 3, pp. 2015.
39. P. Bhole, and A. J. Agrawal. **Single Document Text Summarization Using Clustering Approach Implementing for News Article**, *International Journal of Engineering Trends and Technology (IJETT)*, vol. 15, no. 7, pp. 364-368, 2014.
40. R. Priyadarshini, and L. Tamilselvan. **Document clustering based on keyword frequency and concept matching technique in Hadoop**, *International Journal of Scientific & Engineering Research*, vol. 5, no. 5, pp. 1367-1372, 2014.
41. P. Xie P, and E. P. Xing. **Integrating Document Clustering and Topic Modeling**, in *Proc. Twenty-Ninth conference on Uncertainty in Artificial Intelligence (UAI)*, 2013, pp. 694-703. https://arxiv.org/abs/1309.6874v1
42. S. Karol, and V. Mangat. **Evaluation of text document clustering approach based on particle swarm optimization**, *Cent. Eur. J. Comp. Sci.*, vol. 3, no. 2, pp. 69-90, 2013. https://doi.org/10.2478/s13537-013-0104-2
43. R. Chongtay, M. Last, M. Verbeke, and B. Berendt. **Summarize to learn: summarization and visualization of text for ubiquitous learning**, in *Proc. 3rd IEEE Workshop on Interactive Visual Text Analytics, At Atlanta, GA.* 2013.
44. R. Sipos, A. Swaminathan, P. Shivaswamy, T. Joachims. **Temporal Corpus Summarization Using Submodular Word Coverage**, *CIKM'12, Maui, HI, USA, ACM*, 2012, pp. 754-763. https://doi.org/10.1145/2396761.2396857
45. K. Shubhanker, A. P. Singh, and V. Pudi. **A frequent keyword-set based algorithm for topic modeling and clustering of research papers**, in *Proc. of 3rd Conference on Data Mining and Optimization (DMO)*, IEEE, 2011, pp. 96-102. https://doi.org/10.1109/DMO.2011.5976511

46. L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. **A Time-Dependent Topic Model for Multiple Text Streams**, in Proc. *17th ACM SIGKDD international conference on knowledge discovery and data mining, New York*, NY, USA, 2011, pp. 832-40. https://doi.org/10.1145/2020408.2020551
47. W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, X. Tong, and H. Qu. **TextFlow: Towards Better Understanding of Evolving Topics in Text**, *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412-2421, 2011. https://doi.org/10.1109/TVCG.2011.239
48. S. Harabagiu, and F. Lacatusu. **Using Topic Themes for Multi-Document Summarization**, *ACM Transactions on Information Systems*, vol. 28, no. 3, article no. 13, 2010. https://doi.org/10.1145/1777432.1777436
49. Q. Ma, and X. Zha. Discovering Evolutionary theme patterns from text: An exploration of temporal text mining, in *Proc. eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, 198-207*. https://doi.org/10.1145/1081870.1081895
50. B. Liu, C. W. Chin, and H. T. Ng. **Mining Topic-Specific concepts and definitions on the web**, in *Proc. 12th international conference on World Wide Web*, 2003, pp. 251-260. https://doi.org/10.1145/775152.775188