International Journal of Advanced Technology and Engineering Exploration, Vol 5(44)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2018.543018

Review Article

A survey and analysis based on topic based classification

Chandni Sikarwar[1*], Kailash Patidar[2] and Rishi Kushwah[3]
M.Tech Scholar, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[1]
Professor and HOD, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[2]
Assistant Professor, Department of Computer Science, School of Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[3]

Abstract
In this paper a survey and analysis based on topic based data classification has been presented. It includes the topic based data orientation, data categorization, document clustering, etc. This study provides the analytical way to analyze the methods previously published and provide explorative way of the approaches presented. It also provides the discussion based on the attributes and property used and explored. This study provides the discussion of different partitioning algorithm, different grouping methods and classification approaches. Based on the study, future enhancements have been suggested.

Keywords
Data categorization, Classification, Clustering, Partitioning algorithms.

1.Introduction
In current scenario proper document retrieval is a major concern. Retrieving the appropriate document is important for efficient retrieval of document. It is helpful in appropriate retrieval and categorization.

In certifiable exigency, intentionality producer's bear unendingly reasonably add to conflicted goals and a far reaching satisfies with contrastive sprinter options [1−5]. The multi-basis improving additionally give a battle sooner rather than later. Its includes traverse thorough spaces like the outline space, consolidating the characterizing factors of the competitor arrangements, and the expectation space, constituting the mapping of every hopeful answer for the various target capacities values [6]. The last is where optimality is getting going, tradeoffs are investigated, and choices are ordinarily come to. So there is the need of arrangement in light of different choice criteria which can be heuristic, it will be conceivable by client characterized imperatives and numerous specific requirements.

*Author for correspondence

It can be smarter to locate a legitimate bunched approach to arrange the records, at that point apply some characterization criteria which will be fulfilled some edge an incentive to give the compelled method for these issue. It can be accomplished through affiliation lead mining [7], we can utilize apportioning system additionally in light of the fact that it can diminish the looking time and upgrade the seeking ability [8, 9].

For characterization we can utilize affiliation run mining with some grouping procedures like k-means and fuzzy c-means, it will be a superior choice [10]. At that point we can upgrade it utilizing a few streamlining systems like ant colony optimization (ACO), particle swarm optimization, Mimetic calculation etc. [11−13]. Subset superset dividing can be utilized for parceling and better grouping [14].

There is an exponential increment in the quantity of advanced information and content records. Accordingly, it is exceptionally hard to sort out these expansive accumulation of content reports in an successful way and finding fascinating data or examples [15, 16] has turned into an essential undertaking. To accelerate the looking procedure for comparable reports for finding required archives, record grouping is a strategy generally utilized.

Grouping out an extensive arrangement of records into various comparative groups [17]. Along these lines, the records in a similar bunch are more like each other than to records in another group. Preprocessing is done to change over the words to their base frame, to expel stop words, copy words before applying vector space model to the content records. The separations between records are estimated utilizing closeness measures like Cosine, Jaccard [18] and so on. At that point the bunching calculations are connected till required number of groups is framed. There are two normal grouping calculations. Dividing calculations in which groups are registered straightforwardly.

The remaining of this paper is organized as follows. In section 2 literature review has been discussed. In section 3 previous result analyses has been presented. In section 4 problem analyses has been presented. In section 5 concluding remarks along with the future suggestions have been presented. Finally references are given.

2.Literature review

In 2011, Zuhtuogullari and Allahverdi [19] observe that an extendable and improved item set generation approach has been constructed and developed for mining the relationships of the symptoms and disorders in the medical databases. The algorithm of the developed software finds the frequent illnesses and generates association rules using Apriori algorithm. The developed software can be usable for large medical and health databases for constructing association rules for disorders frequently seen in the patient and determining the correlation of the health disorders and symptoms observed simultaneously.

In 2010, Yang et al. [20] suggest that the SOM has main disadvantage of the need to know the number and structure of neurons prior to training, which are difficult to be determined. Several schemes have been proposed to tackle such deficiency. Examples are growing/expandable SOM, hierarchical SOM, and growing hierarchical SOM. These schemes could dynamically expand the map; even generate hierarchical maps, during training. Encouraging results were reported. Basically, these schemes adapt the size and structure of the map according to the distribution of training data. That is, they are data-driven or data oriented SOM schemes. In this work, a topic-oriented SOM scheme which is suitable for document clustering and organization will be developed. Their proposed SOM will automatically adapt the number as well as the structure of the map according to identified topics. Unlike other data-oriented SOMs, our approach expands the map and generates the hierarchies both according to the topics and their characteristics of the neurons. The preliminary experiments give promising result and demonstrate the plausibility of the method.

In 2013, Yang et al. [21] two major deficiencies of classical SOM are the need of predefined map structure and the lack of hierarchy generation. Several approaches have been devised to tackle these deficiencies. They suggest that both structural and topical constraints which specified by the user could be used to guide the learning process. Preliminary experiments demonstrate improvements over previous algorithm on text categorization task.

In 2016, Pradip and Patil [22] suggested that the traditional clustering algorithms have some problems like instability of clusters, complexity and sensitivity. They have implemented a hierarchical and fuzzy relational eigenvector centrality-based clustering algorithm. The exploratory result demonstrates that hierarchical grouping will be valuable calculation for content records and gives better outcomes.

In 2016, Harish et al. [23] proposed a novel text categorization method. It is based on modified support vector clustering (SVC). The principle disadvantage of customary SVC is that it regards unclassified reports as anomalies. They have used fuzzy c-means (FCM) for this drawback. The changed (SVC-FCM) is connected to order message records. They have used FCM on the unclassified documents. To assess the execution of the proposed strategy, they led probes standard 20-NewsGroup dataset.

In 2017, Popat et al. [24] a test investigation of closeness based technique, HSC for estimating the similitude between information protests especially message archives is presented. It additionally gives a calculation which has an incremental approach and assesses group resemblance between archives that prompts much enhanced outcomes over other conventional techniques. It likewise centers on the determination of fitting likeness measure for examining comparability between the records.

In 2016, Dou and Liu [25] suggested that the visual text analysis is important in the real world scenario. They have proposed visual metaphors for visual text analysis. Thus, measures and benchmark datasets would significantly help the exploration network

assess, arrange, furthermore, think about visual content examination frameworks.

In 2016, Kohana et al. [26] suggested a distributed calculation scheme. It is helpful in scoring relationship among documents. This plan sorts reports by utilizing a calculation which computes a score an incentive for the connection between a classification and a word in a record. The more drawn out estimation time progresses toward becoming when expanding the quantity of reports. Subsequently, our plan utilizes different machines. An ace hub isolates a record set into a few subsets, and it circulates them to every estimation hubs. Utilizing this circulated estimation makes the figuring time short, and furthermore makes the memory utilization low.

In 2015, Nema and Sharma [27] proposed a feature optimization based multi-label text categorization. The procedure of highlight improvement is finished by subterranean insect settlement improvement. Their approach is based on ant colony optimization. For the procedure of order utilized bunch mapping characterization strategy. The element enhancement process lessens the loss of information amid the change of highlight mapping amid the order. For the approval of proposed calculation utilized some standard dataset, for example, site page information, medicinal inquiry information and RCV1 dataset. Our exact assessment demonstrates that proposed calculation is superior to fluffy pertinence procedure.

In 2015, Bide and Shedge [28] suggested that the need of faster categorization of documents for the forensic investigation. Along these lines, there is a need to discrete different accumulations of archives into comparable ones through bunching. Indicating number of groups is obligatory in existing apportioning calculations and the yield is absolutely reliant on given info. Over bunching is the significant issue in archive grouping. Their proposed calculation takes contribution as keywords found after extraction and takes care of the issue of over grouping by isolating the records into little gatherings utilizing divide and vanquish strategy. In this paper, an improved document bunching calculation is given which creates number of groups for any content reports and uses cosine comparability measures to put comparable reports in legitimate bunches. Test comes about demonstrated that exactness of proposed calculation is high contrast with existing calculation as far as measure and time complexity.

In 2016, Wandabwa et al. [29] suggested that the Content order involves settling on a choice on regardless of whether a record has a place with an arrangement of pre-indicated classes of different records. The key component of this calculation lies in the closeness estimation rule that is equipped for distinguishing neighbors of a specific archive to high correctnesses. The main disadvantage of this approach is in the weighting of all highlights to decide the separation among the reports being referred to. This isn't just tedious yet additionally abuses PC assets without including anything considerable to the general outcomes. Their approach has the improved performance as compared to the classical KNN.

3.Previous result analysis

Result analyses based on the related papers are as follows (*Table 1*):

Table 1 Results analysis based on the previous methods

S.No	Reference	Method	Approach
1	[30]	Generic log files and algorithms	They have presented a data capturing process in online games. They have focused on the structural components. They have introduce a suggested structure of the log files and the generic algorithms used to extract information. The term non specific calculation alludes the normal relevance of the calculations to log documents crosswise over diversions and understudies to give analysts rich and significant information about understudy practices.
2	[31]	Fuzzy based enhancement on prism and J48 classifier	They have proposed a modified computed aided design of experiments (MCADEX) using Kullback-Leibler divergence and modified principal component analysis (MPCA). It is used to improve the prediction of student's performance in prism and J48 classifiers. A fuzzy system is utilizing in acquiring information from human specialists can manage loose issues. These tenets are creating for portraying the relationship among the information quality space and classes. In fuzzification, Gaussian participation work is utilized. In this strategy, the weight estimation of each quality is ascertained utilizing neural

S.No	Reference	Method	Approach
			system. Fuzzy parameters are streamlined by utilizing cucko search algorithm. The trait with most extreme weight esteem and fuzzified estimation of highlights are utilized for building tree of crystal and J48 classifiers. The test results demonstrate that the proposed approach is giving better outcomes as far as exactness, better positive negative rate.
3	[32]	Text clustering approach using deep-learning vocabulary network	They presented a novel approach named deep-learning vocabulary network. Their vocabulary network has been constructed based on related-word set. It contains the "cooccurrence" relations of words.
4	[33]	Clustering of text documents by implementation of k-means algorithms	They have proposed clustering for use in browsing a collection of documents and the results fetch from the search engine by the relevant query.
5	[34]	Clustering articles based on semantic similarity	They have descried about the formation of the semantic representation for articles. They have applied two standard clustering methods, k-means and the Louvain community detection algorithm.

4.Problem analysis

The problem analyses based on the previous research work are shown below:

1. Document categorization based on the content and highlighted topic is missing in several research works.
2. Level wise categorization is needed to improve the searching.
3. Threshold based partitioning can improve in categorization.
4. Threshold based optimization may improve in appropriate document cluster ranking.

5.Conclusion

In this paper several aspects of document clustering and categorization have been analyzed. Different trends used in the previous technique have been explored along with the study and discussion. Pros and cons have been discussed. In future a framework based on optimization and classification technique can be developed for better searching and categorization.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Azizah A, Abraham J. Content analysis and exploratory factor analysis of relationship goals among young adults: converging data from instagram and offline surveys. International Journal of Advanced Computer Research.2017; 8(34):11-34.

[2] Rauber A, Dittenbach M, Merkl D. Towards automatic content-based organization of multilingual digital libraries: an English, French, and German view of the Russian information agency novosti news. In third all-Russian conference digital libraries: advanced methods and technologies 2001.

[3] Rauber A, Merkl D, Dittenbach M. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks. 2002; 13(6):1331-41.

[4] Bagajewicz M, Cabrera E. Pareto optimal solutions visualization techniques for multiobjective design and upgrade of instrumentation networks. Industrial & Engineering Chemistry Research. 2003; 42(21):5195-203.

[5] Berger W, Piringer H, Filzmoser P, Groller E. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. In Computer Graphics Forum. 2011; 30(3):911-20.

[6] Beume N, Naujoks B, Emmerich M. SMS-EMOA: multiobjective selection based on dominated hypervolume. European Journal of Operational Research. 2007; 181(3):1653-69.

[7] Dubey AK, Shandilya SK. A novel J2ME service for mining incremental patterns in mobile computing. In international conference on advances in information and communication technologies 2010 (pp. 157-64). Springer, Berlin, Heidelberg.

[8] Shrivastava P, Gupta H. A review of density-based clustering in spatial data. International Journal of Advanced Computer Research. 2012; 2(5):200-2.

[9] Chen K, Liu L. A random rotation perturbation approach to privacy preserving data classification. In proceedings of international conference on data mining 2005 (pp. 589-92).

[10] Liang SC, Lee YC, Lee PC. The application of ant colony optimization to the classification rule problem. In international conference on granular computing 2011 (pp. 390-2). IEEE.

[11] Sadh AS, Shukla N. Association rules optimization: a survey. International Journal of Advanced Computer Research. 2013; 3(9):111-5.

[12] Modiri A, Kiasaleh K. Permittivity estimation for breast cancer detection using particle swarm optimization algorithm. In annual international

conference of the engineering in medicine and biology society 2011 (pp. 1359-62). IEEE.
[13] Liu Y, Chung YY. Mining cancer data with discrete particle swarm optimization and rule pruning. In international symposium on IT in medicine and education 2011 (pp. 31-4). IEEE.
[14] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. CONSEG-2012 (pp.1-6). IEEE.
[15] Srihari S, Leedham G. A survey of computer methods in forensic handwritten document examination. In eleventh international graphonomics society conference 2003 (pp. 278-81).
[16] Oppliger R, Rytz R. Digital evidence: dream and reality. IEEE Security & Privacy. 2003; 99(5):44-8.
[17] Mehrbod A, Zutshi A, Grilo A. A vector space model approach for searching and matching product e-catalogues. In proceedings of the eighth international conference on management science and engineering management 2014 (pp. 833-42). Springer, Berlin, Heidelberg.
[18] Bai VM, Manimegalai D. An analysis of document clustering algorithms. In international conference on communication control and computing technologies 2010 (pp. 402-6). IEEE.
[19] Zuhtuogullari K, Allahverdi N. An improved itemset generation approach for mining medical databases. In International Symposium on Innovations in Intelligent Systems and Applications 2011 (pp. 39-43). IEEE.
[20] Yang HC, Lee CH, Ke KL. TOSOM: a topic-oriented self-organizing map for text organization. International Journal of Computer and Information Engineering.2010; 4(5):1013-7.
[21] Yang HC, Lee CH, Wu CY. Incorporating user constraints into topic-oriented self-organizing maps. In symposium on foundations of computational intelligence 2013 (pp. 91-7). IEEE.
[22] Pradip KG, Patil DR. Summarization of sentences using fuzzy and hierarchical clustering approach. In symposium on colossal data analysis and networking 2016 (pp. 1-7). IEEE.
[23] Harish BS, Revanasiddappa MB, Kumar SA. A modified support vector clustering method for document categorization. In international conference on knowledge engineering and applications 2016 (pp. 1-5). IEEE.
[24] Popat SK, Deshmukh PB, Metre VA. Hierarchical document clustering based on cosine similarity measure. In international conference on intelligent systems and information management 2017 (pp. 153-9). IEEE.
[25] Dou W, Liu S. Topic-and time-oriented visual text analysis. IEEE Computer Graphics and Applications. 2016; 36(4):8-13.
[26] Kohana M, Sakaji H, Kobayashi A, Okamoto S. A distributed calculation scheme for contents categorization. In international conference on advanced information networking and applications 2017 (pp. 614-20). IEEE.
[27] Nema P, Sharma V. Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique. In international conference on computers, communications, and systems 2015 (pp. 1-5). IEEE.
[28] Bide P, Shedge R. Improved document clustering using k-means algorithm. In international conference on electrical, computer and communication technologies 2015 (pp. 1-5). IEEE.
[29] Wandabwa H, Zhang D, Sammy K. Text categorization via attribute distance weighted k-nearest neighbor classification. In international conference on information technology 2016 (pp. 225-8). IEEE.
[30] Alom BM, Scoular C, Awwal N. Generic log files and algorithms developed for educational multiplayer games. International Journal of Advanced Computer Research. 2017; 7(33):223-32.
[31] Regha SR, Rani UR. A fuzzy based enhancement on prism and j48 classifier prediction of student performance. International Journal of Advanced Technology and Engineering Exploration. 2018; 5(42):89-95.
[32] Yi J, Zhang Y, Zhao X, Wan J. A novel text clustering approach using deep-learning vocabulary network. Mathematical Problems in Engineering. 2017:1-13.
[33] Singh H. Clustering of text documents by implementation of k-means algorithms. Streamed Info-Ocean. 2016; 1(1): 53-63.
[34] Wang S, Koopman R. Clustering articles based on semantic similarity. Scientometrics. 2017; 111(2):1017-31.

Chandni Sikarwar had completed her BE from Alpine Institute of Technology, Ujjain, M.P. in 2012 in the Department of Computer Science and Engineering. Currently she is pursuing M.Tech in Computer Science from SSSITS, Sehore.

Email: chandni.sikarwar90@gmail.com