Hierarchial Clustering in Data Mining: Technical and Analytical Review

Dr. Poonam Yadav
D.A.V College of Engineering & Technology,
Kanina, Haryana 123027, India

Abstract— Hierarchical clustering method is the combination of various procedures for identifying the web pages of cluster; thereby the information needs are satisfied for the user. Hierarchical classification is used widely for its efficiency and effectiveness. Regrettably, this hierarchical classification is not utilized directly, because of the datasets that are not organized explicitly in hierarchical forms. In this framework, the 20 papers are organized and made a research review over the Hierarchical clustering in data mining. Finally, the results are examined with the diagrammatic representation and tabulation as well.

Keywords—Clustering, Statistical Clustering, Flat Clustering, Hierarchical Clustering, Sensor Networks.

I. INTRODUCTION

Clustering means, the objects are divided into a group of related objects. This cluster's are correlated with the hidden prototype, e.g., in data mining [21] between the data attributes the correlation is discovered significantly. The data that are represented in Clustering method by some clusters may loss some definite fine information there by it attains simplification. The standard aim of these clustering techniques is for attaining a outcome that offers a large grouping quality and maintains the key data performance. The clustering with efficient documentation has a combined features of information retrieval (IR), natural language processing and machine learning (ML). The conventional algorithm in clustering is generally on the basis of bag-of-words (BOW) feature to represent a document. But still this BOW possessed some limitations such as (i) this method consider the entire words in similar way that neglects any semantic or syntactic connection between them, and (ii) the curse of dimensionality caused by the introduction of large vector space.

Statistical clustering is the first work that can be utilized for offering the set of classification in the multivariate information, on the basis of little relationship in the multivariate set of data (said as clusters). Particularly in flat clustering, each one of them allocated with clustering technique and it also identifies the given set of numbers. Flat clustering Algorithms are very quick. But the limitations of this flat clustering were the resulted quality relay on the prior choice. Further this model does not differentiate the closer and close patterns. Amongst the clustering techniques that are available hierarchical clustering is said to be the major model because of its capability in summarizing the data's summarize hierarchical structures on interpretable and intuitive mode. HC investigate the repeatedly the adjoined group of two clusters, then merged it up to the time of obtaining the group of objects in a single cluster. The HC context is based on the linkage method for calculating the distance among these two clusters. The total clustering procedure is recorded generally on the dendogram, which is a graphical representation from that which the individual clusters information has been obtained.

Some of the major limitation of data mining [22] in hierarchical clustering is it cannot undo the algorithm that was previously done. Time complexity is high so it has to be reduced. On the basis of the distance matrix types it undergoes some issues. They are (a) noise and outlier's sensitivity, (b) Large cluster breaking process and (c) while handling the clusters with different size and convex shape it possesses some difficulty. The main function has not been minimized directly. In some cases the correct number of cluster's identification by the dendogram may be difficult to perform.

This paper reveals the hierarchical clustering in data mining [23] techniques and its algorithmic analysis, performance measures and by its better performance. The paper is arranged as follows: The Literature review is explained in the Section II. The analysis review is made under the performance measure and the algorithm analysis in Section III. Section IV defines the analytical results and challenges of this concept. Section V concludes this paper

II. LITERATURE REVIEW

A. Related Works

In 2010, Gang and Chunwei [1] have stated that for the past two decades, the development in digital collection of data and the improvement in WWW permitted the organizations as well as the companies for sharing and storing the large amount of electronic credentials. Moreover the organization, analysis and presenting the credentials was manually hard and ineffective. Therefore to get the appropriate information, search engines were used that gave an answer to the queries from a list of web pages. The major challenge of this was to help the clients to get the applicable web page effectively from the enormous collection of text. The proposed study explained the hierarchical clustering model, which contained numerous factors for identifying the web page clusters thereby satisfies the information requirement of the user. Further for searching and navigation purpose this clusters were primarily predicted, as well as it was used potentially for few visualization. The investigation was done by the processional search engine with Clickstream data and the examined outcome showed that this clustering method was effectual and competent, with respect to the subjective and objective actions.

In 2012, Ceci, *et al*. [2] have explained that the drill-down and roll-up operations in traditional OLAP systems develops fixed hierarchy over data cubes that referred in discrete attributes. Sensor networks have a rising application over the OLAP systems with stimulated research. The earlier definitions of this ad-hoc discretization hierarchy next to every OLAP dimension has to be avoided, this was the main aim of this paper. This can be rectifies by implementing a narrative model named as density-based hierarchical clustering algorithm, thereby the roll-up and drill-down operations by continues dimensions over OLAP data cubes were supported. Here fact-table measures also have to be considered in this clusters dimension hierarchy model. The experimental result has shown that the clustering effect was enhanced regarding the performance analysis.

In 2004, Jiau, *et al.* [3] have declared that clustering was used widely in various applications such as data analysis, market research, pattern recognition, and image processing. The traditional clustering algorithm that utilized the distance-based dimensions for calculating the data differences were not adaptable for non-numeric attributes, while data mining was executed. This may creates some loss and resulted in low quality clusters. This proposed model used a narrative hierarchical clustering algorithm, named as MPM to cluster

the non-numeric attributes. The aim of this MPM was to maintain the features of data that clusters the group of data objects by maximum intra-similarity and minimum intra-similarity. This can be accomplished by two phases (a) implementing equality measures that captured the "characterized properties'' of data. (b) The matrix permutation and matrix participation application to allocate data in the suitable clusters. Further the heuristic-based algorithm named Heuristic MPM was proposed for the matrix permutation and partitioning in order to minimize the processing time.

In 2004, Ward, *et al.* [4] have stated that the investigative data analysis and visualization was a major issue in this framework. The traditional techniques were used to analyze lot of data sets, with respect to the variable number or the records number. This problem can be rectified by the use and development of various resolution schemes. The proposed model represented the improvement of multi-resolution investigative visualization environment of current activities for huge-scale multi variable information. The data set which includes hundreds of dimensions or millions of records can be effectively presented by interaction, visualization and data management techniques.

In 2011, Spanakis, *et al.* [5] have proposed a narrative model of theoretical clustering of hierarchal documents by extracting the knowledge from Wikipedia. The disadvantage of the conventional model (classic bag-of-words) was conquered by the proposed method by the utilization of link structure and textual content from Wikipedia. By deploying the programming application based on Wikipedia interface, a compact and robust representation of document was built in the real-time, in which there was no need to store the Wikipedia information locally. The process of clustering was hierarchal thereby it expands a design of common information through utilizing Wikipedia topics to select the labeled cluster.

In 2017, Jeon, *et al.* [6] have implemented a new method based on linkage as NC-link. The linkage choice not only affects the quality still the efficiency of HC also gets affected, in order to satisfy this issue the NC-link-based HC was proposed. The implemented algorithm retains the quadratic nature of time complexity for the correctness action. Thereby linear space complexity was shown to allow the address the million-object information from the personal computers. This approach can be extended by NC-link algorithmic nature that allowed sub-quadratic time approximation and SIMD parallelization in HC. The proposal was verified with the experimental test that includes the huge-scale synthetic and real number of dataset. The conventional algorithms were benefited by utilizing this proposed NC-link because of their linkage model that obtained a minimized time, space demands and better clustering results.

In 2018, Vignati, *et al.* [7] have presented an improved hierarchical bunch clustering. The accepted scaling of data and because of the function's arbitrary choice resulted in non-uniqueness that affects the classical clustering. These limitations can overcome by presented the weighted asymmetric function, for measuring the distance between the two components. The weighting of data dynamically relied with the advancement degree in the clustering process. Geometric performance of the clustering derives the narrative proximity measure; thereby the robustness of clustering was indicated in opposition to the uncertainty measure in primary data, and both the disengage outcome of the data scaling was allowed. The computational cost of this classical model was maintained and this model was applicable for both hierarchical and flat clustering.

In 2005, Zhao and Karypis [8] have focused to build hierarchal solution by executing the document clustering algorithms that (a) implements an inclusive studies on document clustering algorithms and partition algorithms, which utilizes the merging schemes and various criterion modules, and (b) implemented a clustering algorithm named as constrained agglomerative algorithms that mix the characteristics of agglomerative and partition functions. Thereby it reduced the starting stage of errors created by the agglomerative models and the quality of clustering function was thus improved. The outcome of the experiment shown that the partition algorithm has a best solution than any other conventional algorithms hence because of the less computational use it made perfect for clustering huge collections of documents and have high quality in clustering effect.

In 2013, M. Tabesh and H. Askari-Nasab [9] have described about the mining operation that includes various stages. Thereby the polygons were drawn by the mining engineers so that it can be used as operational guidelines. On the basis of the knowledge of deposit and the experience of the engineer, these polygons were designed manually. The shapes that formed by the automatic programming could minimize the efforts and maximize the quality. Huge polygons were used for mining cut in long-term planning while small shapes were required for mining cut in short-term planning. Additionally, the shape of desired polygons was affected by the mining direction. This problem was overcome by introducing a clustering algorithm with shape management mechanism, thereby the corrected guidelines for the entire abovementioned shapes were provided with standardized parameters. Various mining schemes were applied to get the algorithmic performance and were estimated on the basis of homogeneity in rock types, run times, grade, and determined destinations.

In 2010, Malik, *et al.* [10] have implemented IDHC, hierarchy cluster without mining for considerable worldwide patterns was constructed by the pattern-based hierarchical clustering algorithm. At first IDHC finds the local patterns to the responsible size 2 patterns thereby make sure a balance efficiently among the pattern implication and the frequency of the local pattern from a dataset. Then by utilizing the local promising patterns the Cluster hierarchy was directly designed. Initially every pattern formed a cluster, and then the remaining hierarchy cluster was offered using the distinctive iterative cluster modification procedure. Moreover, most descriptive cluster, and adaptive in soft clustering solutions were produced by this IDHC, which allowed the existence of adaptable nodes in different levels of hierarchy cluster. Here the experiment was carried out with text datasets that contained 16 standards, and the outcome shown as IDHC has a betterment than the hierarchical clustering algorithms with respect to FScore measures and average entropy.

In 2007, Lazzerini and Marcelloni [11] have assumed that the web portal pages were already arranged according to the quantity of various categories. A systematic approach was introduced in the user profile to decide the hierarchy in the user's access history. Initially, the access log was filtered by eliminating the users with poor interest and from the occasional users. After that to cluster the web portal users in accordance with their frequent interest, an algorithm was applied named Unsupervised Fuzzy Divisive Hierarchical Clustering (UFDHC) algorithm. Thereby each was characterized under a prototype that in which the profile of typical member of the group was defined. Then a narrative categorization method was implemented for identifying the profile of a particular user fits in. in conclusion, the obtained result was compared with the conventional profile application that depends on the fuzzy C-means modified version, and the outcome shows the better performance in the proposed one.

In 2011, Xiong, *et al.* [12] have presented a hierarchical clustering algorithm for data categorization namely DHCC. The data categorization clustering task in optimal perception was viewed to propose the efficient measure to ferine and initialize the cluster splitting. There were five advantages in proposed algorithm: at first, dendogram represented the similarity levels and nested grouped pattern at various granules because of its hierarchical nature.

Subsequently, this was fully automatic and parameters free that in case, no need on assumption with respect to the clusters number. The data that was processed was independent in order, this is the third merits. The fourth benefits were that the huge data sets have been scalable. And at last, it flawlessly discovers the embedded clusters in subspaces. Both the real and synthetic data performance showed the betterment of this procedure.

In 2012, Gillis and Francois [13] have considered two algorithms named as multiplicative updates and the hierarchical alternating least squares; thereby the NMF issues were solved. The proposed algorithm was on the basis of cautious analysis of the needed computational cost, using that the above schemes were considerably accelerated; hence the convergence properties were preserved. The project gradient model also used these acceleration techniques. The accelerated algorithm was demonstrated efficiently with the text data sets and image and was compared with the conventional nonnegative least squares algorithm.

In 2008, Mahmood, *et al.* [14] have stated about the requirement for improving numerous variation in clustering of network traffic flow records that was existed, thus the underlying traffic patterns were quickly deduced. The proposed model made an investigation on the utilization of the clustering mechanism, in which the attractive traffic patterns was identified efficiently from the network traffic data. Combination of attributes that includes categorical, hierarchical and numerical was taken care of in this implemented structure in one-pass hierarchical clustering algorithm. The resultant outcome demonstrated the enhancement in efficiency and accuracy when compared with the existing conventional model on the basis of network traffic.

In 2007, Wang, *et al.* [15] have developed a new two clustering algorithms named as PoissonHC and PoissonS on the basis of the development and adaptation of hierarchical clustering techniques and Self- Organizing Maps, for data analysis in SAGE. While testing the experimental and synthetic data on SAGE, this technique possesses numerous benefits on comparing with existed pattern techniques. While combining the PoissonS, SAGE data and PoissonHC statistical properties, along with the hybrid technique (neuro-hierarchical approach on the basis of the incorporated PoissonHC and PoissonS; the outcome shown that the visualization for SAGE data and pattern discovery has effectively improved. Further, it initiates the user-friendly environment with improved and accelerated data mining in SAGE.

In 2005, Gao, *et al.* [16] have developed a narrative algorithm, in that the hierarchical structure of data corpse in a taxonomy data was mined automatically from the framework that adopted for classification hierarchy. Specifically, the relation between the documents, terms and categories were represented by calculating the matrices. Moreover, the added cluster with three materials from reliable bipartite spectral graph co-partitioning at various scales, and it was created the comprehensive decomposition problem with singular value. Finally, the cluster category constructed the hierarchal taxonomy. Investigational outcome explained that the implemented strategy have found the hierarchy taxonomy there by aid to enhance the accuracy in classification.

In 2005, Lin and Chen, [17] have proposed a new correspondence measure to calculate the inter-cluster distance that was mentioned to as cohesion. Two- phase clustering algorithm was designed based on this new evaluation of cohesion named as cohesion-based self-merging (CSM) that in which the input dataset size operates with linear time. On the basis of this combined characteristics of partitioned and hierarchical clustering models, the CSM algorithm in the first phase divides the dataset inputs into various sub-divisions of clusters. Subsequently in the second phase, depends on cohesion function it merged all the sub-divided clusters in a hierarchical manner. CSM algorithms space and time complexities were evaluated. The experimental study shown as, CSM own an efficient tolerance for the outsiders with numerous workloads. Major advantage of this CSM was, it efficiently cluster the datasets in arbitrary shapes and offered betterment in clustering outcome over the other conventional models.

In 2005, Khan, *et al.* [18] have presented the study of enhancement of SVM training time, while specially deals on huge datasets by utilizing the hierarchical clustering process. The limitations in the traditional hierarchical clustering algorithms were rectified by this new algorithm named Dynamically Growing Self-Organizing Tree (DGSOT). The boundary points were analyzed with the help of clustering that trained the SVM among two classes. A new mixture of DGSOT and SVM were presented that by utilizing the clustering structure generated by the DGSOT algorithm was expanded and started with the early training set. Rocchio Bundling technique was compared with the proposed method and selection was made in random with respect to the accurate training time gain and loss utilizing the real data set single benchmark. The proposed algorithm enhanced the process of training in SVM with large accuracy generalization than the Rocchio Bundling technique.

In 2013, Gibert, *et al.* [19] have evaluated the numerical and categorical variables by presenting a mixed generalization Gibert's metrics also to include the semantic variables. The semantic variables were compared with the introduced super concept-based distance (SCD) by considering the instructions offered by the ontology reference. This standard has demonstrated a SCDs better performance regarding other method that was obtained from the journals for comparing the semantic characteristics. At last, on the basis of the touristic data, two real applications was shown Gibert's metrics general impact in clustering process. It also affects the reference ontology that was taken into account in clustering. The major outcome shown that while during the availability of the reference ontology, the final clusters meaningfulness was enhanced logically.

In 2018, Y. Jarraya, *et al*. [20] have introduced a narrative refinement process of efficient hierarchical for the automated modeling or multi-level fuzzy systems. The implemented method was based on the two standard techniques such as: at first, the multi-objective algorithm named Multi-Objective Extended Immune Programming algorithm (MOEIP) was implemented foe evaluating the Hierarchical Fuzzy System's architecture. In the second one, Hybrid Artificial Bee Colony algorithm (HABC) was applied for the Beta membership function limit and the consequential rules division were tuned. The implemented fusion procedure infuse the two studied schemes there by the parameter tuning and the architecture learning up to a close Hierarchical Fuzzy System optimization was created. The evaluation was done by the standard time series issues for the efficient methodology, a few large dataset's dimensional classification and nonlinear plant discovery issues. The proposed method was superior with respect to smaller rule-base, better convergence speed and high accuracy over the conventional methods.

III. Analytical Result

A. Algorithmic Analysis

The analysis of numerous algorithms that are used in this framework is illustrated in Fig.1. The algorithm that is implemented is as follows: the multiple factor hierarchical clustering model is used in [1]. In [2] the algorithm that is used is named as density based hierarchical clustering algorithm. The heuristic_MPM is the algorithm that deployed by the author in [3]. That author in [4] implemented the data management techniques. The CHC method is deployed in [5] by the author. The NC-link based HC is used in [6]. In [7], the static and dynamic model comparison is analyzed. The UPGMA is the algorithm that is used in [8]. The author in [9]

developed the shape controlled clustering algorithm. In [10], IDHC is utilized by the author. UHDFC is the algorithm that is deployed in [11]. In [12] DHCC is utilized by the author. Accelerated multiplicative updates and hierarchical ALS are the two algorithms that are used in [13]. In [14], ECHIDNA is used. The two algorithms named PoissonS and poissonHC are deployed by the author in [15]. In [16], co-clustering based hierarchical taxonomy mining is utilized. The CSM is the algorithm that is used in [17]. The author in [18] implemented the DGSOT algorithm. In [19], the SCD is utilized by the author. MOEIP algorithm is deployed by the author in [20].

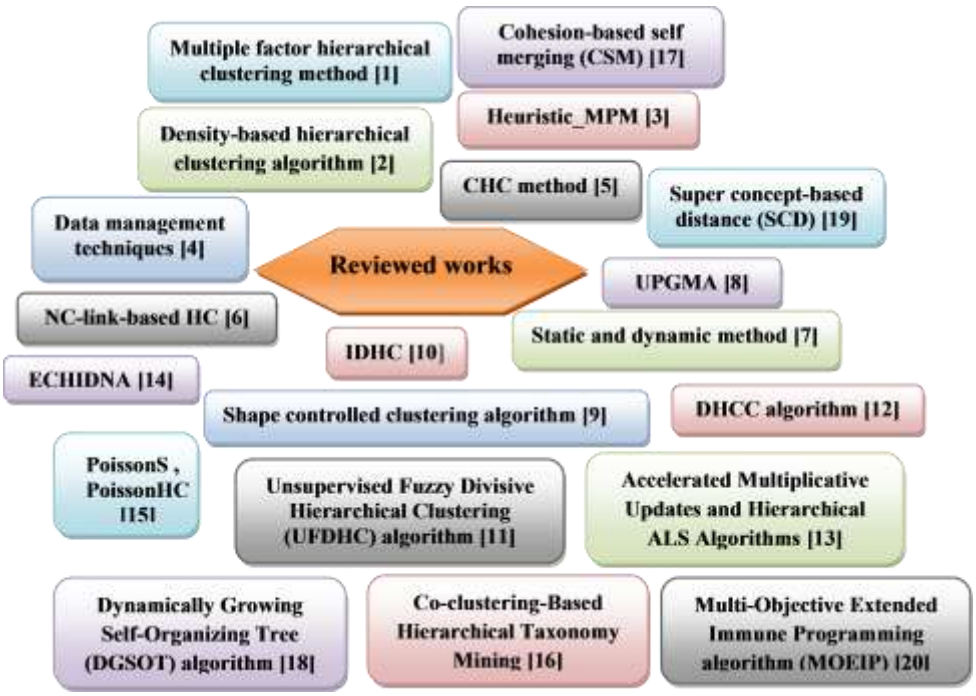

Fig. 1. Algorithmic Analysis of Reviewed Works

B. Performance analysis

The analyses of performance have been defined by using the performance measure, which is analyzed in this paper is given in Table I. the measures that are deployed in this framework are time, coherence, number of clusters, box width, std-deviation, data size, rate, accuracy and so on. Here, in this overall performance the contribution of time is 40%, which is the major measure utilized for analysis of performance. Only 5% is the contribution of the measure named Coherence. The contribution of measure named number of cluster is given as 10%. The other measures such as box width, std-deviation, and data size, rate, sparsity and frequency contributes only 5% of the overall percentage. The accuracy measure obtained a contribution of 10%. Moreover, there are some other measures used in this research work they are quality, utility, and number of users, radius and so on.

TABLE I. PERFORMANCE MEASURE OF REVIEWED WORKS

Citation	Time	Coherence	Cluster number	Box width	Std dev	Data size	Rate	Accuracy	Sparsity	Frequency	Others
[1]		□									
[2]	□										
[3]			□								
[4]											□
[5]	□										
[6]	□										
[7]				□							
[8]											□
[9]	□		□								
[10]					□						
[11]											□
[12]						□					
[13]									□		
[14]	□						□				
[15]											□
[16]	□										
[17]	□										
[18]	□							□			
[19]										□	
[20]								□			

C. *Attained Best Measure*

The best efficient measure that is used in this framework of hierarchical clustering is given in Table. II. Here the best value regarding the coherence and utility in [1] attains a measure as 4.46 and 4.90, respectively. The time that is utilized in most of these papers attains a best value as 15s. The quality measurement accomplishes 0.08. The number of cluster achieves a best measure as 320. The space value attains $1000*10^3$. The box width attains 5.75, and the std-deviation, number of users, data size, and sparsity attained a best performance measure of 0.749, 254, 1.5, and 99.92, respectively. The rate and radius that are used in [14] achieved a best measure of 0.97, and 0.95, correspondingly. The percentage of accuracy measure accomplishes the best value as 91.20% and in [19], the frequency measures the value of 12.

TABLE II. MAXIMUM ACHIEVED MEASURE

Measure	Best performance value	Citation
Coherence	4.46	[1]
Utility	4.90	[1]
Time	15	[2][5][6][9][14][16][17][18]
Quality	0.08	[2]
Cluster number	320	[3][9]
Space	$1000*10^3$	[6]
Box width	5.75	[7]
Std-deviation	0.749	[10]
Number of users	254	[11]
Data size	1.5	[12]
Sparsity	99.92	[13]
Rate	0.97	[14]
Radius	0.95	[14]
Accuracy	91.20%	[18][20]
Frequency	12	[19]

IV. RESEARCH GAPS AND CHALLENGES

The large set of datasets need the mixture of the clustering and sampling techniques in clustering. While using the data sample with clustering algorithm, the object allocation issue is acquired naturally; this is the major challenge in clustering techniques. These issues can be rectified by effectively utilizing the defined cluster centre of Chi-square distance and categorical data among each and every object in the categorical cluster. Another problem is the co-portioning and fast designing of algorithm. This has to be rectified to get an efficient optimal solution. In the hierarchical taxonomy generating procedure, the efficient way to select k automatically have yet to be found.

Another major challenge in the agglomerative hierarchical clustering is the computational cost, which is high. The agglomerative procedures complexity is given as O (n^2 log (n)), here n represent the number of observations. Moreover this method, for a huge number of observations could not scale well. Some of the drawbacks that have to be resolved in hierarchical clustering is classification/regression problems by utilizing the analytical learning framework in clustering.

V. CONCLUSION

This paper explains the analysis done on 20 papers with respect to the hierarchal clustering in data mining and their challenges that are has to be solved. The reviewed framework reveals the methodologies that are used and the performance analysis with maximum attained measures. The challenges and the future plan are also described in this paper. The reviewed work is evolved by the diagrammatic representation and the tabulation reviews.

REFERENCES

[1] Gang Kou, Chunwei Lou,"Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data,"Annals of Operations Research,vol. 197, no.1,pp.123-134, 2012J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Michelangelo Ceci, Alfredo Cuzzocrea, Donato Malerba,"Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering,"Journal of Intelligent Information Systems, vol.44,no. 3 pp. 309-333, 2015

[3] Hewijin Christine Jiau, Yi-Jen Su, Yeou-Min Lin,Shang-Rong Tsai,"MPM: a hierarchical clustering algorithm using matrix partitioning method for non-numeric data,"Journal of Intelligent Information Systems, vol.26, no. 2,pp. 185-207,2006

[4] Matthew Ward, Wei Peng, Xiaoning Wang," Hierarchical visual data mining for large-scale data,"Computational Statistics, vol.19,no.1,pp.147-158,2004

[5] G. Spanakis, G. Siolas and A. Stafylopatis, "Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents," The Computer Journal, vol. 55, no. 3, pp. 299-312, March 2012.

[6] Y. Jeon, J. Yoo, J. Lee and S. Yoon, "NC-Link: A New Linkage Method for Efficient Hierarchical Clustering of Large-Scale Data," IEEE Access, vol. 5, pp. 5594-5608, 2017.

[7] Federica Vignati, Damiano Fustinoni, Alfonso Niro,"A novel scale-invariant, dynamic method for hierarchical clustering of data affected by measurement uncertainty," Journal of Computational and Applied Mathematics,vol. 344, pp. 521-531,15 December 2018

[8] Ying Zhao,George Karypis,Usama Fayyad,"Hierarchical Clustering Algorithms for Document Datasets", Data Mining and Knowledge Discovery, vol.10,no.2,pp.141-168,2005

[9] M. Tabesh, H. Askari-Nasab,"Automatic creation of mining polygons using hierarchical clustering techniques",Journal of Mining Science, vol.49, no.3, pp.426-440, 2013

[10] Hassan H. Malik, John R. Kender, Dmitriy Fradkin, Fabian Moerchen, "Hierarchical document clustering using local patterns,"Data Mining and Knowledge Discovery, vol.21, no. 1,pp. 153-185, 2010

[11] Beatrice Lazzerini, Francesco Marcelloni, "A Hierarchical Fuzzy Clustering-based System to Create User Profiles,"Soft Computing, vol.11, no.2, pp. 157-168, 2007

[12] Tengke Xiong, Shengrui Wang, André Mayers, Ernest Monga,"DHCC: Divisive hierarchical clustering of categorical data,"Data Mining and Knowledge Discovery, vol.24, no.1,pp. 103-135, 2012

[13] N. Gillis and F. Glineur, "Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization," Neural Computation, vol. 24, no. 4, pp. 1085-1105, April 2012.

[14] A. N. Mahmood, C. Leckie and P. Udaya, "An Efficient Clustering Scheme to Exploit Hierarchical Data in Network Traffic Analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 6, pp. 752-767, June 2008.

[15] H. Wang, H. Zheng and F. Azuaje, "Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 2, pp. 163-175, April-June 2007.

[16] Bin Gao, Tie-Yan Liu, Guang Feng, Tao Qin, Qian-Sheng Cheng and Wei-Ying Ma, "Hierarchical taxonomy preparation for text categorization using consistent bipartite spectral graph copartitioning," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, pp. 1263-1273, Sept. 2005.

[17] Cheng-Ru Lin and Ming-Syan Chen, "Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 145-159, Feb. 2005.

[18] Latifur Khan, Mamoun Awad, Bhavani Thuraisingham,"A new intrusion detection system using support vector machines and hierarchical clustering,"The VLDB Journal,VOL.16, NO.4,PP. 507-521, 2007

[19] Karina Gibert, Aïda Vall, Montserrat Batet,"Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering",Knowledge and Information Systems, vol.40, no. 3,pp.559-593, 2014

[20] Y. Jarraya, S. Bouaziz and A. M. Alimi, "Hierarchical Flexible Beta Fuzzy Design by a Multi-Objective Evolutionary Hybrid Approach," in IEEE Access, vol. 6, pp. 11544-11558, 2018.

[21] R. Espinosa and J. A. Aguilar, "A Goal-oriented Requirement Analysis Approach for the Selection of Data Mining Techniques for Non-Expert Users," IEEE Latin America Transactions, vol. 16, no. 4, pp. 1180-1185, April 2018.

[22] Y. Guo and L. Lu, "Simulation research for telecommunication data mining based on mobile information node," IET Software, vol. 12, no. 3, pp. 245-250, 6 2018.

[23] K. Liu, Y. Liu, L. Hu, S. Du and J. Su, "Distribution network reliability investment effectiveness evaluation based on defect data mining," The Journal of Engineering, vol. 2017, no. 13, pp. 2062-2066, 2017.