International Journal of Soft Computing 12 (3): 156-163, 2017
ISSN: 1816-9503

Proposed Partitioning Approach and Hierarchical Approach for the Clustering of Web Users

Hadj-Tayeb Karima and Belbachir Hafida
Departement Mathematiques-Informatique,
Universite des Sciences et de la Technologie d'Oran Mohamed Boudiaf,
USTO-MB, BP1505, El M'naouar, Oran, 31000, Algerie

Abstract: In web usage mining, the cluster analysis is the most important technical. Based on this technical, the users groups provide insight into browsers behavior to refine the behavioral patterns to the website access and to identify frequently visited pages. Into the clustering context, the most popular approach is the partitioning approach but itprinciple as it stands seems inappropriate on sequential data. This work attempts to over come limitations and proposes a new model of users clustering. This approach is based on set ofmeasures which ensure rules quality and evaluate the interest of generated sequential rules. The experimental study implements the proposed algorithm, the k-medoids algorithm and hierarchical agglomerative approach in order to guarantee the good partitioning of the data interms of evaluation measures of the clustering quality and calculation time. Relative result sattempt to ensure a good partitioning data in terms of evaluation measures of clustering quality.

Key words: Web usage mining, clustering partitioning approach, clustering hierarchical agglomerative approach, k-medoids algorithm, evaluation measures of clusters, rulesquality measures

INTRODUCTION

To better understanding the behavior of web browsers and satisfy their needs, it isimportant to process and analyze this large data by applying the data mining techniques. Among these techniques, the cluster analysis is the most important technique applied to web data which represent a stream of sequential data where a time is primordial to follow visited pages. Among the existing methods of clustering, this study is interested to partitioning methods in the clustering of web users. It aims to provide solutions because this method presents some disadvantages when applied on web data. To this end, a new clustering model is based on sequential patterns and on grouping rules. This study wishes to ensure agood partitioning of data in terms of evaluation measures of the clustering quality and calculation time.

For a more deepened study in the experiments phase, the proposed approach is compared with the results obtained by the Hierarchical approach and k-medoid partitioning algorithm inorder to guarantee the good partitioning of data in terms of evaluation measures of the clustering quality in a minimal execution time.

Hierarchical and partitioning clustering approaches: Depending on how the clusters are formed, two types of approaches have been proposed: the first method is the partitioning method which subdivides objects into a number of classes by using an iterative optimization strategy whose principle is to generate the initial partition. The partitioning method improves progressively by allocating data from one class to another. Indeed, each individual is assigned to the closest center by calculating the distance measure with the proximity between the gravity center and the object. There after, the new centers representing groups are recalculated.

The algorithms for this method have the advantages to be: easy to understand and to implement to be rapid to have a low requirement in memory size and to be applicable toany type of data. The algorithms used in this method are: the k-means algorithm, the k-medoids algorithm and the EM algorithm (Karkkainen and Ayramo, 2006; Liu, 2007; Masseglia *et al*., 2000).

The second method is hierarchical method which constitutes classes gradually in the hierarchical form in which the most similar objects are grouped into clusters at low levels. This approach regroups nested sequences by producing a tree where clusters at an intermediate level include all clusters below them in the hierarchy. This method presents the advantage of being applicable on any type of attributes it is flexible for the granularity level and has the ability to treat any measure of similarity or

Corresponding Author: Hadj-Tayeb Karima, Departement Mathematiques-Informatique, Universite des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, BP1505, El M'naouar, Oran, 31000, Algerie

distance. It allows to the user to explore clusters at any level of the tree. Two methods of hierarchical clustering are identified: agglomeration clustering anddivision clustering. Among the algorithms used in this method, researchers Silva (2009) and Guha *et al.* (1999): the birch, cure and rock algorithms.

Division clustering: This method builds the tree from top to bottom. It starts with all data points in a root cluster. It divides the root into a set of clusters where each cluster is recursively divided again until the singletons clusters of individual data points remain. The process is repeated until each class contains a single point or if a desired number of classes is reached. In this technique, it is possible that each object alone forms a cluster.

Agglomerative clustering: This method builds the tree from bottom to up. Initially, each object forms a cluster and at each level it is necessary to merge the closest pairs of clusters to pass to the next level. The process is repeated until all points are in the same class or until a fixed number is reached. In hierarchical clustering, three methods are determined for the calculation of the distance between two clusters:

- Single link method (nearest neighb or clustering method): in single-link hierarchical clustering, the distance between two clusters is the distance between the two closest data points in the two clusters
- Complete link method (farthest neighbor clustering method): in complete-link hierarchical clustering, the distance between two clusters is the distance between the two farthest data points in the two clusters
- Average link method: in this method, the distance between two clusters is the average distance between all the pairs of distances between the data points of the two clusters

The agglomeration approach is the most commonly used in hierarchical clustering. It will be implemented for a comparative study with the proposed approach in the implementation part (Eduardo and Brea, 2013; Kashef, 2008; Poulose, 2013).

Problem: Among these two methods, the proposed approach is based on the principle of the partitioning method to build a clustering model of web users. However this method presents two disadvantages when applied on web data for following reasons.

The final partition depends on the initial partition due to the random selection of thek initial points that can generate a bad partitioning. These algorithms need to run multiple times with different initial states to obtain better results. In each initialization with a defined number of clusters, there are different solutions that can be away from the optimal solution. Then, it eventually becomes necessary to run thes ealgorithms multiple times with different initializations and retain the best regrouping data. The use of this solution is limited because of its very high cost in terms of time calculation, the memory space and the number of steps as the best partition can be obtained after several runs of the algorithm. These limitations make the application of these algorithms unsuitable on these large bases.

The calculated distance measure between the cluster center and the affected object in partitioning approach seems inappropriate in this case where the web data are manipulated. Indeed, this distance is insufficient since it doesn't take into account the similarity between calculated objects and nor does it reflect the order of the items in the sequence. These algorithms can generate groups at risk of becoming empty during the regrouping process.

This study aims to improve the partitioning approach which is the most popular method used in clustering applications. Among the partitioning algorithms, this study compares the proposed approach with k-medoids algorithm which is one of the most popular algorithms used in clustering applications.

The k-medoids algorithm is based on the partitioning technique that clusters the data set of n objects into k clusters with k known a priori. It could be more robust to noise and outliers as compared to k-means algorithm because it minimizes a sum of general pair wise dissimilarities instead of a sum of squared euclidean distances. The most common realisation of k-medoid clustering is as follows.

Initialization: It randomly selects k of the n data points as the medoids.

Assignment step: It associates each data point to the closest medoid.

Update step: For each medoid m and each data point o associatedto m swap m and o and compute the total cost of the configuration (that is the average dissimilarity of o to all the data points associated to m). It selects the medoid o with the lowest cost of the configuration. K-medoid algorithm repeats alternating Steps 2 and 3 until there is no change in the assignments (Eduardo and Brea, 2013).

Literature review: Several works clustering technique are interested on the on web user's data (Marascu, 2009) proposes navigators classification by a new centroid approach (sequence clustering in data streams) based on a technical of sequences alignment appliedon the cluster. The researcher also proposes a new measure to affect each sequence to the closest center based on the longest common sub-sequences between two sequences. After this, he compares this approach to the hierarchical clustering where he obtained a better clustering by measuring the quality of clusters.

Hay *et al*. (2004) propose a new method to extract navigations patterns by using an aligned sequence method. This method partitions a trace of navigations according to the order of required pages and administrates the problem of sequences clustering of different sizes on which is based the distance count. Experimental results are compared with a method based on the euclidian distance that does not contain any information on the sequence.

Poornalatha and Prakash (2013) propose an effective modification of k-means algorithm for the web sessions clustering. This clustering model uses the distance function based on the vector of varying length of sessions. The order of visited pages is not taken into account in session's comparison. In this algorithm, the alignment number isconsidered for determining the distance between sessions.

Fontaine (2008) proposes to study the behavior of users during their web interface. He compares each individual to other individuals at each change of state that is manifested by apparition of an element in the sequence. The researcher proposes to use the ascendant hierarchical clustering which indicates the evolution of proximity between in dividuals according to a distance measure.

In these reseaches, researchers applied a clustering algorithm to modify one or more parameters relating to the representation of the sequence or the distance measure. Through this work the proposed approach attempts to make other modifications which are based on sequential patterns. This technique has the characteristic to analyze the appearance order of items in asequential database according to the time in order to identify the sequence of items which represent in our application, the visited pages.

MATERIALS AND METHODS

Proposed approach: Proposed approach tries to improve the performance of the algorithms of the partitioning approach. This approach aims to eliminate random and iterative selection of the clusters number and to propose two new functions to calculate the gravity center of the cluster and to assign the object at the most appropriate cluster. To measure the quality of a clustering and achieve a good data partitioning, several functions have been suggested: the entropy, the intra-class inertia and the inter-class inertia used in (Koskela *et al*., 2004; Zhao, 2012; Rosenberg and Hirschberg, 2007).

A good data partitioning is obtained by minimizing the entropy function and by maximizing inter inertia (minimizing intra inertia). The proposed clustering model is based on the extraction of sequential rules from the sequential patterns technical. The treated web data represents a stream of sequential data where time is a primordial element in order to follow the footsteps of navigations. To extract these patterns, several algorithms have been proposed which are based on time parameters such as the time window, the minimum and maximum duration that offer the greater flexibility and significantly reduce the space of transactions (Kam and Fu, 2000; Meger, 2004; Saneifar *et al*., 2009).

Among these algorithms: the prefix tree algorithm (Meger, 2004), the generalized sequential patterns algorithm (Ayres *et al*., 2002), the prefix span algorithm (Pei *et al*., 2001), the sequential pattern mining algorithm and spade algorithm (Zaki, 2001).

In the extraction of sequential rules from these patterns, ordering of events must be taken into account to reflect the order of the events that compose the rule. Based on this set of generated rules, the study of quality measures is essential to evaluate the quality of the generated rules and assess the association rules having the same interest. An extracted rule must satisfy the support and confidence measures which are mostly used as basic measures. The support measure remains a heavily used measure for algorithmic reasons, it really difficult to do without the support measure during the exploration phase of the item sets trell is whose role is to restrict the search space in the trell is of item sets.

However, the confidence measure makes sense intuitively which is a major advantage when presenting a set of rules to expert because the conditional probability of the conclusion knowing the premise makes reading more directly. In addition to the great interest in the support and confidence measures as extraction criteria, these functions have a concrete sense of comprehension and are fully absorbed by the non-specialist user.

However, both measures have some limitations since the algorithms that use them generate avery large number of rules that are difficult to manage and many have less interest. The condition of the support removes the rules having a small support while some may have very strong confidence and can present a real interest. To sum up, the support and confidence measures appear in sufficient to measure and to evaluate the quality of the

generated rules. To overcome the weaknesses of these measures, it must consider other measures to evaluate the quality of rules. For this reason, several criteria to define a good quality measures have been proposed and subsequently different measures are available in the literature. Among these measures: lift, recall; conviction; pearl, Φ coefficient; pietetsky-shapiro, novelty, centered confidence, loevinger, sebag-schoenauer and index of involvement (Feno, 2007; Marascu, 2009; Pham *et al*., 2013).

Defined by Fournier (2010) and Pham *et al*. (2013), among the criteria that a good measure must satisfy in a rule; a measure must have a concrete sense and must be easy to interpret. It must be sensitive to the appearance of against-examples and favors the appearance of examples of the rule and allow the user to tolerate some against-examples while maintaining the interest of a rule. It must be sensitive to the size of the data as it should not lose its discriminatory power when the data size becomes large enough. It must also be used with a pruning threshold to remove all rules that do not interest the user and should be used with a pruning threshold.

None of these measures simultaneously satisfy all these criteria. This is why; the problem of research of quality measures to discover the most relevant rules remains largely open. The choice of measures specifically depends on the application domain and on processed data type. Based on these measures; the proposed approach attempts to resolve the following points.

Eliminates random and iterative selection of the clusters number: As mentioned previously for each partition, the partitioning algorithm must randomly select a set of fixed centers defining clusters and must iteratively seek the optimal partition. This algorithm converges in a significant number of iterations. The partitioning algorithms start repeatedly to choose the distribution that provides a better partitioning of data where a considerable loss of time and of a large number of calculations. To overcome this limit, this approach is based on the extraction of sequential patterns to fix the number of clusters. The technique of sequential patterns can extract a set of item sequences, commonly associated over a period of time well specified and highlights the inter-transaction associations to generating a set of sequential rules as following: generate a set of all the frequent patterns by the prefixs pan algorithm.

Filter patterns respecting the temporal parameters initially fixed including: the minimum time duration (MinGap), the maximum time duration (MaxGap) and the time window (WinSize) such as: the duration between two item sets must be strictly less than the winsize for these item sets are grouped in the same item set. The time between the item sets must belong to the interval (MinGap, MaxGap) to be in the same sequence. Generate all valid association rules respecting the confidence and the support measures.

As mentioned earlier the support and confidence measures are not sufficient to guarantee the quality of the detected rules because they allow generating a large number of rules that are very difficult to manage and many of them have little interest.

Based on the measures qualities, there are a large number of measures to characterize association rules and the choice of a measure depends largely on the application domain. These measures are used to identify identical rules in the same interest. Among all the properties (criteria) that a measure must satisfy, researchers have judged according to web application domain the importance of some properties than others. This choice has allowed us to remove some measures as: novelty measure, implication and index pietetsky shapiro measures for the following reasons: implication index and pietetsky-shapiro measures depend on the data size. Both measuresvary from increasing way with the size of the data and risk losing their discriminatory power when the size becomes large enough. In this case, the size of the treated web data should not inter vene in the evolution of the function.

It is desirable that the measure varies a nonlinear way according to the appearance of examples or against examples. This condition reflects that the user can tolerate little against-examples while maintaining the interest of a rule. But the novelty measure doesn't respect this condition.

After this, each chosen measure allows to calculate interest rules. The approach proposes to regroup the rules which have the same interest (quality) represented by a measure value which characterizes the generated group. Each cluster will be characterized by a single value obtained by this measure as along with a number of rules to verify having the same interest.

Proposes a new similarity measure to assign an object: The partitioning algorithms use several distance function (Euclidean, Jaccard) to assign an object to the nearest center that proves unsuited to web data. The similarity measure dedicated to sequential patterns must consider the following criteria:

- The sequential patterns are ordered sequences of item sets and not of items

- The positions of the item sets in temporal order when calculating the similarity should be taken into account
- The number of common items at the sequence should be taken into account

By taking into consideration these criteria, the proposed similarity measure follows these steps: the order of item sets when calculating the similarity is obtained through the verification of sequential rules that represent the item sets sequences and therefore the apparition order of the items compared to the temporal identifier. The similarity measure reflects the relationship between the items constituting the web object. The number of common items obtained by the intersection between the sequences representing the individual (S_1) and the sequences representing the cluster center (S_2).

Concerning the affectation of items to the most appropriate cluster, the first object is assigned to each cluster by verifying the most of its non-redundant rules to avoid getting the empty clusters at the end of the partitioning. The assignment of the remaining objects is achieved when respecting these criteria:

- The object must verify the maximum non-redundant rules of this cluster
- The similarity measure between the object and the cluster center must be the greatest

In case, an object is assigned to several clusters, this object is assigned to the least loaded cluster in terms of owned objects to establish a balance and avoid getting clusters more loaded than others at the end of the partitioning.

This algorithm stops when is no data exchange group or the iterations number is reached. At the end of its execution, each generated cluster is characterized by a set of sequential rulesdefined by the importance degree measuring their interest via the items set (visited pages bybrowsers) of this cluster.

Validates the obtained partition: At each end of the data partitioning, obtained partitions by proposed approach will become pared with those obtained by the hierarchical algorithm and by k-medoids algorithm through different measures to evaluate the quality of clusters: the entropy, intra-class and inter-class measures in a minimum execution time.

RESULTS

For a more deepened study in the experiments phase, the proposed approach was also compared with the hierarchical agglomerative approach in order to guarantee the good partitioning of the data in terms of evaluation measures of the clustering qualityin a short calculation time.

The first part: Aims to perform a comparative study of the quality evaluation measures of generated sequential rules and extracts the rule that offers a good clustering of data. The generation of frequent patterns requires fixing of several parameters including: minimal support, minimal confidence and temporal parameters. The values of these parameters are fixed after several tests and variations. In such a case, a good quality measure is the one that offers a good clustering of data through measures for evaluating obtained clusters.

The second first part: Serves to choose the good partition ensuring good clustering by thek-medoids algorithm which is run several times with different numbers of iterations and clusters. The partition verifying the quality evaluation measures of clusters is considered as the best partition.

The third part: Serves to choose the good partition ensuring good clustering by the hierarchical agglomerative approach which is implemented by using the three different methods (single-link method, complete-link method and average-link method). The obtained partition by these three methods verifying the quality evaluation measures of clusters is considered as the best partition

The fourth part: Serves to compare the results of clustering obtained by proposed model and those obtained by the Hierarchical approach and k-medoids algorithm. Once the best partitions are detected by these algorithms, a comparative study is applied to identify the best partition between them. The best partition must verify the quality evaluation measures of clusters in a time of minimal calculation. These algorithms try to minimize intra inertia (maximize inter inertia) and minimize the entropy in a short time. The maximal number of iteration is fixed to 20.

Selecting a good quality measure obtained by the proposed algorithm: Unlike the k-medoids algorithm that starts up several times, the proposed algorithm starts once with constant measures values that necessitates the gain of time. After several tests, the time constraints are fixed to: MinGap = 0, MaxGap = 7, WinSize = 2. The results are obtained from two studies: The calculation of the evaluation measures of quality clusters for min-support = 0. 3 and min-confidence = 0.6 is represented in Table 1.

Table1: The results of evaluation measures of quality clusters for min-support = 0.3 and min-confidence = 0.6 for proposed approach

Measures	No. of clusters	No. of iterations	No. of generated rules	No. of generated non redundant rules	Intra	Inter	Entropy	Time (MiliS)
Recall	4	5	90	62	549.56	1092855.60	0.61	6766
Lift	5	6	90	62	550.27	1098151.48	0.32	6820
Centered confidence	6	7	90	62	571.97	2336652.28	0.39	6641
Pearl	9	8	90	62	547.18	408008.60	0.33	6859
Φ coefficient	7	6	90	62	560.92	1092750.70	0.61	6750
Loevinger	6	7	90	62	572.01	1096696.36	0.39	7234
Sebag	8	7	90	62	565.15	956343.74	0.54	6547
Conviction	5	6	90	62	552.13	1098154.26	0.31	7003
Less contraction	10	7	86	58	7610.64	113.70	5.02	7315

Table 2: The evaluation measures of quality clusters for the Hierarchical approach

Methods	Under-clusters number	Intra	Inter	Entropy	Time MiliS (Ms)
Average link	50	10706.25	510848.83	43.65	4641
Complete link	50	12421.38	331614.25	59.23	4938
Single link	50	14706.25	200848.83	68.65	5697

Selecting the good partition obtained by hierarchical agglomerative approach: This algorithm is implemented by using these methods (single-link, complete-link andaverage-link). The Table 2 represents the evaluation measures of quality clusters through these measures.

Selecting the good partition obtained by the k-medoids algorithm: For a more detailed experiment, the number of clusters is varied for each launching. This algorithm is executed several times to the same number of clusters by taking the average of the results.

Selecting the best partition obtained by the three algorithms: This analytical study serves to compare the results of the best clustering partition obtained by these algorithms.

DISCUSSION

In proposed approach, the algorithm is executed for two different values of the threshold of minimum support and confidence: for a minimal-support = 0.3 and a minimal-confidence = 0.6, the best partition represented in (Table 1) is obtained by two measures: lift and conviction measures. They have been chosen as the best measures. They guarantee the good partitioning of the data in terms of evaluation measures of the clustering quality in a minimal execution time. In the k-medoids algorithm, the best partition represented in Table 2 and 3 is obtained by the number of clusters = 14 because it satisfies the criteria mentioned above.

However, in the hierarchical agglomerative approach, the best partition represented in Table 2 is obtained by the average link method because it satisfies the criteria mentioned above compared with complete link method and single link method. According to the obtained results by inter, intra and entropy measures, the Hierarchical approach presents a complexity in the calculation and is considered ineffective on a large data set.

Indeed, like in partitioning approach the euclidienne measure used in the different methods (single, complete and average link) to calculate the distance in order to regroup two clusters seems in appropriate on the web data. This distance is insufficient since it does not take into account the similarity between calculated objects and nor does it reflect the order of the items in the sequence.

Comparing the best partition obtained by the Hierarchical approach, the k-medoids algorithm and the proposed approach for a min-confidence = 0.8, the results show that the proposed approach verifies these following criteria:

- The inertia intra calculated by proposed approach is smaller than those calculated by hierarchical agglomerative approach and k-medoids algorithm
- The inertia inter calculated by proposed approach is greater than those calculated by hierarchical agglomerative approach and k-medoids algorithm
- The entropy calculated by proposed approach is smaller than those calculated by hierarchical agglomerative approach and k-medoids algorithm
- The time classification calculated by proposed approach is less than those calculated by hierarchical agglomerative approach and k-medoids algorithm

The built clustering model depends on the dataset. For these implemented data, the clustering model that guarantees the best data partitioning is obtained by lift and conviction measure for a number of clusters = 8. Each cluster is characterized by a measure value that groups the rules having the same interest. These rules will verify by a new user before the assignment.

Table 3: Represents the evaluation measures of quality clusters for the k-medoids algorithm

Clusters	Iterations number	Intra	Inter	Entropy	Time (MiliS)
Nb-clusters = 3	12	1815.48	23768.76	3.36	17656
Nb-clusters = 4	12	1731.99	2266.56	3.29	17989
Nb-clusters = 5	12	1716.27	2344.35	3.51	17907
Nb-clusters = 6	12	1645.85	43847.71	3.75	18985
Nb-clusters = 7	12	1600.52	46719.79	3.84	17675
Nb-clusters = 8	12	1595.38	49578.56	3.99	17715
Nb-clusters = 9	12	1589.05	48963.56	2.04	17972
Nb-clusters = 10	12	1430.35	46833.45	2.11	18775
Nb-clusters = 11	12	1465.00	45789.03	2.15	17629
Nb-clusters = 12	12	1459.31	55932.36	2.19	18614
Nb-clusters = 13	12	1487.52	51705.99	2.21	18651
Nb-clusters = 14	12	1438.80	55939.44	2.06	17633
Nb-clusters = 15	12	1494.03	49452.54	2.47	17592
Nb-clusters = 16	12	1466.33	48456.68	2.51	18650

The value measure that identifies the cluster means that users have the same behavior from visited pages via all rules.

CONCLUSION

This study talks about a proposed partitioning approach for the clustering of web users based on extraction of sequential rules. In the partitioning of web data, the performance of any clustering algorithm is based on the similarity measure and on the calculation of the gravity center of generated clusters that are the key to success of any clustering algorithm.

Proposed approach develops a new clustering algorithm based on the sequential rules in a web environment. Proposed algorithm is based on the partitioning method of the clustering technical and also based on the prefix span algorithm for the generation of sequential patterns. These patterns allow generating a set of sequential rules to be subsequently grouped according to their interest and evaluated through the quality evaluation measures of association rules.

The aim of this research is to improve the performance partitioning method. A first modification is proposed to resolve the problem of the clusters number which is done randomly and iteratively. This study proposes a new similarity measure of the affected objects and thus reduces the execution time. To validate this approach, proposed approach, the partitioning k-medoids algorithm and hierarchical agglomerative approach have been implemented for a comparative study of the performance of two algorithms according to the quality evaluation measures of obtained clusters for each partition. This study has compared the quality of obtained clusters by partitioning proposed algorithm and hierarchical algorithm and it concluded that the proposed approach offers a better partition of data in a shorter running time.

IMPLEMENTATIONS

The proposed algorithm and k-medoids algorithm of the partitioning approach are implemented on Java by using a log file of 800 records. The implementation process of the proposed algorithm is based on three major elements where in the two algorithms attempt to minimize intra inertia (maximize inter inertia), minimize the entropy, in a minimum execution time.

REFERENCES

Ayres, J., J. Flannick, J. Gehrke and T. Yiu, 2002. Sequential pattern mining using a bitmap representation. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, 2002, New York, USA., pp: 429-435.

Eduardo, M. and A. Brea, 2013. Constrained clustering algorithms: Practical issues and applications. Ph.D Thesis, University of Coruna, Coruna, Spain.

Feno, D.R., 2007. Quality measures of the association rules: Standardization and characterization of the bases. Ph.D Thesis, University of La Réunion, Saint-Denis, France.

Fontaine, M.D., 2008. From artificial learning to human learning: From trace harvesting to user modeling. Ph.D Thesis, Universite Pierre et Marie Curie-Paris, Paris, France.

Fournier, V.P., 2010. A hybrid model for learning support in procedural and ill-defined domains. Ph.D Thesis, Universite du Quebec a Montreal, Montreal, Quebec.

Guha, S., R. Rastogi and K. Shim, 1999. ROCK: A robust clustering algorithm for categorical attributes. Proceedings of the 15th International Conference on Data Engineering, March 23-26, 1999, Sydney, Australia, pp: 512-521.

Hay, B., G. Wets and K. Vanhoof, 2004. Mining navigation patterns using a sequence alignment method. Knowl. Inf. Syst., 6: 150-163.

Kam, P.S. and A.W.C. Fu, 2000. Discovering Temporal Patterns for Interval-Based Events. In: Data Warehousing and Knowledge Discovery, Kambayashi, Y., M. Mohania and A.M. Tjoa (Eds.). Springer, Berlin, Germany, pp: 317-326.

Karkkainen, T. and S. Ayramo, 2006. Introduction to partitioning-based clustering methods with a robust example. Master Thesis, University of Jyvaskyla, Jyvaskyla, Finland.

Kashef, R., 2008. Cooperative clustering model and its applications. Ph.D Thesis, University of Waterloo, Waterloo, Ontario.

Koskela, M., J. Laaksonen and E. Oja, 2004. Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval. Proceedings of the 17th International Conference on Pattern Recognition, Vol. 2, August 26, 2004, IEEE, New York, USA., ISBN: 0-7695-2128-2, pp: 1005-1008.

Liu, B., 2007. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin, Germany, ISBN:13-978-3-540-37881-5, Pages: 531.

Marascu, A., 2009. Extracting sequential patterns in data streams. Ph.D Thesis, Universite Nice Sophia Antipolis, Nice, France.

Masseglia, F., P. Poncelet and R. Cicchetti, 2000. An efficient algorithm for web usage mining. Networking Inf. Syst. J., 2: 571-604.

Meger, N., 2004. Automatic search for optimal time windows in sequential patterns. Ph.D Thesis, University of Louisville, Louisville, Kentucky.

Pei, J., J. Han, M.A. Behzad, P. Helen, Q. Chen and M.C. Hsu, 2001. PrefixSpan,: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany, pp: 215-224.

Pham, T.T., J.W. Luo, T.P. Hong and B. Vo, 2013. An efficient algorithm for mining sequential rules with interestingness measures. Intl. J. Innov. Comput. Inf. Control., 9: 4811-4824.

Poornalatha, G. and S.R. Prakash, 2013. Web sessions clustering using hybrid sequence alignment measure (HSAM). Soc. Network Anal. Min., 3: 257-268.

Poulose, J., 2013. Development of hierarchical clustering techniques for gridded. Master Thesis, University of CUSAT, South Kalamassery, India.

Rosenberg, A. and J. Hirschberg, 2007. V-Measure: A conditional entropy-based external cluster evaluation measure. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Association for Computational Linguistics, Prague, Czech Republic, pp: 410-420.

Saneifar, H., S. Bringay and T.M. Maguelonne, 2009. S2MP: A similarity measure for sequential patterns. Proceedings of the International Conference on Evaluation Methods of Knowledge Extraction in Data, February 27, 2009, Montpellier 2 University, Montpellier, France, pp: 5-46.

Silva, D.A.G., 2009. Scalable data analysis: Application to web usage data. Ph.D Thesis, Paris Dauphine University, Paris, France.

Zaki, M.J., 2001. SPADE: An efficient algorithm for mining frequent sequences. Mach. Learn. J., 42: 31-60.

Zhao, Q., 2012. Cluster validity in clustering methods. Ph.D. Thesis, University of Eastern Finland, Finland.