American-Eurasian Journal of Scientific Research 11 (1): 28-31, 2016
ISSN 1818-6785

DOI: 10.5829/idosi.aejsr.2016.11.1.22708

Informative Review on Data Partition and Node Stability in Mining

[1]Akey Sungheetha, [2]R. Rajesh Sharma,
[1]R. Indu Poornima, [1]N. Vaitheeka and [1]G. Ashokkumar

[1]Karpagam College of Engineering , Coimbatore 641032, India
[2]Hindusthan College of Engineering and Technology, Coimbatore 641032, India

Abstract: The web information available is vast and expanding to meet the current challenges of intelligent techniques. There are billions of pages in the web, where pages huge in quantity are not the problem; its disorganized state leads the user to face the conflicts when searching information in order to reach the goal on developing a project. Taking a glance at the categories of web mining, the techniques operated on the types of web transaction (i.e. Protection, Monitoring, Storage and Nominal) enable the billions of pages to function in an organized state. The spine clustering is the technique proposed for the web transaction execution. The rate of partition, volume of transaction, quantity of transaction, delivery, collision, delay and energy consumption are the level metrics calculated to measure the data capacity to address the problem of pages in disorganized state. The results are carefully analyzed and presented for the simulated evolutionary systems.

Key words: Data Capacity · Partition · Accuracy · Clustering · Web mining

INTRODUCTION

Clustering is the technique applied to group the data either randomly or based on the type on user requirement [1]. The node failure due to population of data in a node also uses the six types of level metrics. The Rate of Partition is elaborated for the proposed technique called spine clustering [2]. The dataset is usually grouped by different techniques. The proposed technique is one of the methods for grouping the dataset. In the spine clustering once the standard state is realized, the capacity of the node is evaluated to accommodate the traced information [3]. The information is traced from the web and metrics are evaluated for all the nodes in the network, which updates the network node information [4].

The data can be distributed based on the nodes capacity with respect to the clustering technique for the particular transaction. The clustering technique for the nodes is a unique feature in terms of distance. The distance is an important factor that can be considered during cluster formation. The algorithm preserves the log of distance between the nodes in all the other nodes of the network even, which in turn allows those set of nodes to form a number of clusters [5]. The goal of the technique presented in the paper enable the evaluation of delivery base on the distance between the nodes. The maximum delivery rate depends only on short distance between the nodes. If the nodes face increased distance between them, the delay rate is naturally increased during the communication between the nodes. So the prediction method illustrates the nearness of neighbor nodes that reduce the distance between nodes grouping based on types allowing each node to maintain the information of other nodes [6].

Data Capacity Measure: The clustering type (C_t) and node capacity (N_c) are the two methods for computation of cluster space (C_s) and cluster data (C_d) using prediction method. Every node is provided with degree, ID and location [7]. This will be remarkable to identify the order of nodes.

$$C_s = N_c - A(C_t) \quad (1)$$

$$C_d = C_{tn} \text{ X } C_{td} \quad (2)$$

Where,
C_{tn} – Number of type data
C_{td} – Number of data sets
$A(C_t)$-Represents the availability

Corresponding Author: Akey Sungheetha, Karpagam College of Engineering, Coimbatore 641032, India.

The N_c can be measured by the result on number of data types (D_t) which can be represented as N_c=CD(N)+TS(N)+DD(N)+ND(N) (for the proposed paper) where N denotes the number of data present in that particular type and $Nc=D_{t+1}$ (Increment) if in case of additional data type [8].

This approach enables four data types on different modes as follows:

- Continuous Data (CD) pertain sequenced indexing and partitioning to classify various modes to evaluate the metrics.
- Time Series (TS) data are multilayered indexing and partitions to resolve the clustering complexity with respect to the time interval.
- Discrete Data (DD) obtain finite value with a periodical difference between the set of values.
- Nominal Data (ND) express the count in form of label for set of data values observed.

These different data types take a part of role in spine clustering by evaluating the rate of partition for each measure. The sample table for accuracy and error data is illustrated below. The occurrence of each cluster for its type is found and similar type is indexed [9].

Once indexed, the nearest indexed neighbors are combined to form a cluster which improves the result of computation. When indexed, the similar cluster type quickly processes any number of updating and transaction due to request on demand between nodes that become easier. This technique is entirely different from all the other clustering methods. The nodes capacity can be measured based on the indexed clusters with respect to their type. In each node cluster type C_t is essential to calculate the cluster space, C_s and cluster data C_d for each node.

The clustering group the like records in a document together, which profits the developer with high level implementation [10]. Now there will be availability of different set of cluster where rate of partition is measured for each set. At most situations clustering also describe partitioning. Some trademark industries grouped the populated information into partitions and extract the product information for business sales and marketing [11]. Here the prediction method for nearest neighbor increases the rate of partition.

The Table 1 highlights the clustering phenomenon while measuring the accuracy on data capacity of a node. The nearness is described in most clustering algorithm, but in this paper nearness facilitates the prediction on measuring rate of partition.

Table 1: Rate of Partition for Accuracy of Data

	Priority						
Cluster							
Techniques	1	2	3	4	5	6	7
Hierarchical						△	
Two-way		△					
co-clustering					△		
Agglomerative					△		
Fuzzy c-means						△	
QT_clustering		△					△
Data clustering		△					
isolate index			△				
Graph methods						△	

Table 2: Rate of Partition for Error Data

	Priority						
Cluster							
Techniques	1	2	3	4	5	6	7
Partitional			△				
Spectral						△	
Two-way							△
Agglomerative		△					
Fuzzy c-means		△					
Pooling					△		
hashing		△					
Graph methods							△

The Table 2 shows error measure for the node in addition to accuracy.

The prediction method illustrates the rate of partition measure for spine clustering. Hence the spine clustering depends on sequence of dataset formation after the execution of all above measures and logic. The important partitions of the industrial product information can be represented as high-level clusters (H_c), (product with minimum cost), low-level clusters (L_c), (product with maximum cost). Some other records with any other constraints have no possibility to match with these clusters. In this case any of the data with the effect to form another cluster with an important partition of the information can be allocated to cluster level called middle level clusters (M_c) [12,13]. The cost for all these clusters based on accuracy and error can be measured.

Hence the volume of transaction can be estimated based on these three different level clusters where the quantity of transaction can be measured individually for low level, high level and middle level which takes minimum time, T_{min} and after integration of data by indexing and partitioning the maximum time T_{max} is evaluated.

For the volume of data, the field is divided into ranges using deep similar data partition method, where the values of field are clustered without removing the weakly dominated clusters [7]. The solution of this cluster is large in dimension that is equal to the number of fields in the table. Then this large cluster is broken into P equal parts where P is the required number of partition.

Transaction on Clustering Operation: The cluster operation in each partition updates the integrated data and ensures the RPUP properties of transaction as follows:

- Reflexive operation represents the atomicity of the clustered transactions.
- Preservative operation insists only the specific transaction execution for a cluster to represent the consistency.
- Unawareness operation is the lack of carefulness for which to and fro transaction is guaranteed inside a cluster even without the knowledge of other concurrent transaction that are executed in the system representing isolation.
- Persistence operation define the changes made to the data inside a cluster exist in case of system failure after transaction completeness representing durability.

Each node indexes the data and then partitions the cluster. The reflexive operation maintains the atomicity if the similar kind cluster types match with similar kind data. The preservative operation is the method to maintain a single copy of the similar data to simplify the transaction for next time. In a node the unawareness of other transaction assist in concentrating only in the particular transaction that is currently executed on respective cluster type irrespective of any other transaction on other cluster types. The indexing and partitioning simultaneously function for each specific cluster type transaction.

The persistence results in durability of the modified steps during indexing and partition based on cluster type techniques. The durability stands even if system fails or system is disabled. The web transaction over internet is common to all internet users. The web transaction concepts use different algorithm [14]. Probably at present there is no algorithm to address the problem of this particular mechanism that use cluster node and cluster data for web transaction [15].

Table 3: Clustering Accuracy

	Accuracy (%)						
Cluster Techniques	Priority						
Hierarchical	82	54	67	59	73	95	64
Two-way	84	92	74	78	77	43	72
co-clustering	73	84	82	81	93	86	89
Agglomerative	52	57	80	82	94	75	87
Fuzzy c-means	47	49	47	46	48	95	89
QT_clustering	85	94	79	83	81	84	96
Data clustering	90	95	94	96	94	91	89
isolate index	68	57	91	75	79	85	87
Graph methods	54	65	66	69	67	92	90
Spine	95	98	97	96	99	98	97

Table 4: Clustering Error

	Error (%)						
Cluster Techniques	Priority						
Partitional	1	3	0.3	2	4	3	2
Spectral	4	3	1	0.9	1	0.4	2
Two-way	4	2	3	3	1	0.9	0.5
Agglomerative	1	0.2	5	7	5	6	4
Fuzzy c-means	3	0.6	2	4	3	1	2
Pooling	0.9	0.9	1	2	0.4	1	2
Hashing	1	0.8	2	4	5	7	3
Graph methods	3	5	4	2	1	2	0.7
Spine	0.3	0.2	0.4	0.2	0.5	0.1	0.2

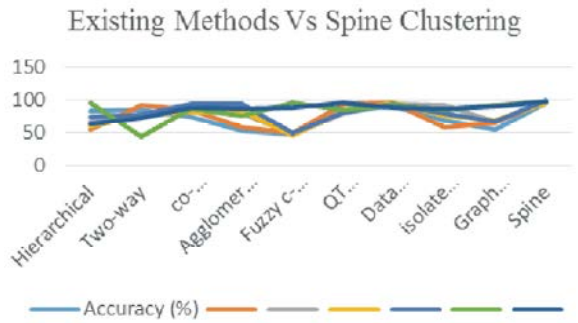

Fig. 1: Accuracy Comparison of Spine Clustering with existing methods

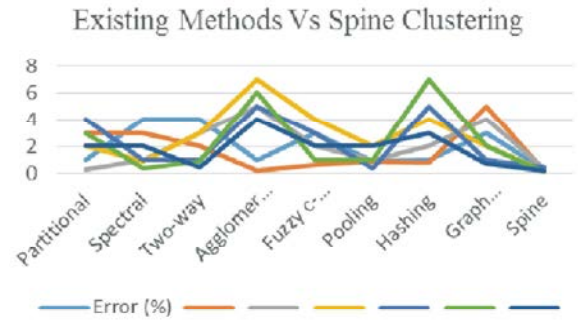

Fig. 1: Error Comparison of Spine Clustering with existing methods

Table 3 shows the accuracy after the data partition and transaction in the web.

The proposed spine clustering is improved in its accuracy when compared to the existing clustering methods as shown in Figure 1.

Table 4 shows the error after the data partition and transaction in the web.

The proposed spine clustering shows very minimum error occurrence when compared to existing clustering methods.

CONCLUSION

Evaluating selected transaction exchange data in order to exchange data, the collection of data in the node is split or partitioned for the exchange process between the nodes. During the exchange, the conversions on spine are based on the modes such as Faulty mode, Active mode and Inactive mode.

The approach can be constructed with the scalability options on approximating other data types also.

REFERENCES

1. Estivill-Castro, V. and J. Yang, 2000. A Fast and robust general purpose clustering algorithm. Pacific Rim International Conference on Artificial Intelligence, pp: 208-218.
2. Richard, C. Dubes and Anil K. Jain, 1988. Algorithms for Clustering Data, Prentice Hall.
3. Fraley, C. and A.E. Raftery, 1998. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, Technical Report No. 329. Department of Statistics University of Washington.
4. Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang and Lei Hua, 2004. An Introduction to Cluster Analysis for Data Mining.
5. Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data Clustering: A Survey, ACM Computing Surveys, 31(3).
6. King, B., 1967. Step-wise Clustering Procedures, J. Am. Stat. Assoc., 69: 86-101.
7. Khaled, M. Alzoubi, Peng-jun Wan and Ophir Frieder, 2002. Distributed Heuristics for Connected Dominating Sets in Wireless Ad Hoc Networks, Journal of Communications and Networks, 4(1).
8. Sneath, P. and R. Sokal, 1973. Numerical Taxonomy, W.H. Freeman Co., San Francisco, CA.
9. Guha, S., R. Rastogi and K. Shim, 1998. CURE: An efficient clustering algorithm for large databases. In Proceedings of ACM SIGMOD, International Conference on Management of Data, New York, pp: 73-84.
10. Ward, J.H., 1963. Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, 58: 236-244.
11. Mukras, R., N. Wiratunga, R. Lothian, S. Chakraborti and D. Harper, 2007. Information gain feature selection for ordinal text classification using probability re-distribution,Proc. of IJCAI Textlink Workshop.
12. Martin-Valdivia, M.T., M.C. Diaz-Galiano, A. Montejo-Raez and L.A. Urena-Lopez, 2008. Using information gain to improve multi-modal information retrieval systems, Inf. Process. Manage, 44: 1146-1158.
13. Frank, E., M. Hall, L. Trigg, G. Holmes and I.H. Witten, 2004. Data mining in bioinformatics using Weka, Bioinformatics, 20: 2479-2481.
14. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
15. Crammer, K. and Y. Singer, 2000. On the learn ability and design of output codes for multiclass problems, Proc. of the Thirteenth Annual Conf. on Computational Learning Theory, pp: 35-46.