IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. I (Sept - Oct. 2016), PP 84-89
www.iosrjournals.org

A study on Clustering Algorithms for XML Data Clustering

S.Saranya[1],B.S.E.Zoraida[2]
Research Scholar[1], Assistant Professor[2] Department of Computer Science and Engineering, Bharathidasan University,Trichy-24.

Abstract: *Nowadays mining meaningful information from large scale web documents is more important to satisfy the user demand. XML and RDF documents are supporting the semantic information retrieval to interpret and extract meaningful information for user query. XML documents have light weight code and logical structure, which facilitate easy exchange of data values and structure information in terms of knowledge. Many mining techniques and algorithms are used to enhance the performance of XML information Retrieval. Classification (Supervised Learning) and Clustering (Unsupervised Learning) are the preprocessing techniques used to grouping up the similar data objects based on similarity criteria. This paper presents the study on three clustering algorithms (k-means, EM, Tree Clustering) and its similarity measures on XML datasets. The three clustering algorithms are compared and tested with the same xml datasets for finding the best one to cluster XML documents.*

Keywords: *Clustering, XML, Data Clustering*

I. Introduction

With the development of information technology, digital information grows very fast and has more and more kinds of types.[1] Web data have different formats; therefore about 90% of data remain without use and are not represented in user mining. [2] XML is a W3C standard structured language.XML is used to provide meaningful information about the stored content. An XML document can be modelled as a rooted, ordered, and labeled tree. [3] The XML page will be consisted of built-in and user defined tags. The metadata information of the pages is extracted from the XML. User defined tags will help the system in getting answers from reliable sources. [4] To get meaningful information from XML document there are different techniques and mechanisms were incorporated. But finding the best among them is a difficult task.[4,5] Data classification and Clustering techniques are used to extract and summarize data into similar groups. As web is migrating from HTML to XML, large amount of data is accumulating day by day. This huge amount of data on the websites is needed to be managed.[6,7] For the same purpose, many data mining techniques are available to manage the datasets.[7] Supervised learning is one of the technique used to discovers the patterns in the data, which is used to predict the values of the class attribute of future data instances. These classes indicate some real-world predictive or classification tasks such as determining whether a news article belongs to the category of sports or politics, or whether a patient has a particular disease.[8] Clustering is one technology for finding intrinsic data that has no class attributes.. It organizes data instances into **similarity groups**, called **clusters** such that the data instances in the same cluster are similar to each other and data instances in different clusters are very different from each other.[8] Clustering is also known as **unsupervised learning.** This paper presents the study on various clustering algorithms.

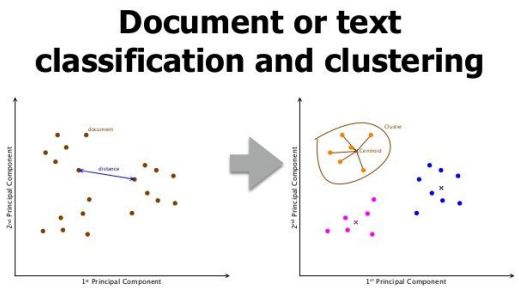

Figure 1: Supervised Vs Unsupervised Learning

II. Cluster

Cluster is the process of grouping up the similar data objects in to groups. A Cluster group of objects are different from other cluster group of objects. Cluster analysis is an important technique which is used for many practical applications. Clustering is the process of partitioning a given set of objects into disjoint clusters. [9] This is done in such a way that objects in the same cluster are similar while objects belonging to different

clusters differ considerably, with respect to their attributes.[9,10] To get meaningful information from clusters, then clusters should get the real structure of data objects.

A)Similarity-based Cluster definition: A cluster is a set of objects that are "similar", and objects in other clusters are not "similar." A variation on this is to define a cluster as a set of points that together create a region with a uniform local property, e.g., density or shape.[11]
Generally Cluster Algorithm classified as five types. The most commonly used Cluster types are Partition and Hierarchal cluster.

i. K-means Algorithm

K-means Algorithm is a most popular and productive algorithm for cluster formation. K-means is partition clustering technique. But the time and computational complexity of k-means algorithm increases when the size of XML dataset increases. Moreover, this algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers for improving the performance of the k-means clustering algorithm.[10,12]The phenomena of the k-means algorithm is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists two phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid.[10] Euclidean distance is calculated for every cluster based on the distance relies between the data points and the centroid. The pseudo code of k-means algorithm is given below.[9,12]

K-means Algorithm
Input: D = {d1, d2,.....,dn} //set of n data items.
k // Number of desired clusters
Output: A set of k clusters.
Steps 1*:* Arbitrarily choose k data-items from D as initial centroids; ***Step 2:*** Repeat Assign each item d_i to the cluster which has the closest centroid; Calculate new mean for each cluster; Until convergence criteria is met. [9,10,12]

ii. EM Algorithm

EM is a Cluster algorithm stands for Expectation and Maximization. The EM algorithm computes the cluster based on likelihood nature. EM algorithm has two phases. The Expectation is for assignment of data items to the centers and it's calculated in first step. The Maximization is the second step for update of centers. [13,11]

Pseudo code
E-step: estimate E(z) for each z, given θ
M-step: estimate θ maximizing E(log likelihood) given E(z) [where "E(logL)" is wrt random z ~ E(z) = p(z=1)]
In E step, Computes Estimate distribution over labels given a certain fixed model and M step Choose new parameters for model to maximize expected log-likelihood of observed data and hidden variables.[13,10]

iii. Tree Clustering Algorithm

Hierarchical clustering is another major clustering approach. Hirerichal clustering method is a popular clustering technique because of its versatile properties. It clusters by producing a nested sequence of clusters like a **tree. [10]**Tree clustering is used to joining the similar objects into large groups in form of tree structure. Tree cluster is also known as joining cluster algorithm.[,14]

Algorithm Agglomerative(*D*)
Step 1: Make each data point in the data set *D* a cluster,
Step 2: Compute all pair-wise distances of **x**1, **x**2, …, **x***n* □ *D*;
2 **repeat**
Step 3: find two clusters that are nearest to each other;
Step 4: merge the two clusters form a new cluster *c*;
Step 5: compute the distance from *c* to all other clusters;
12 **until** there is only one cluster left [14,15]

B)Comparison of Clustering Algorithms

Tree cluster analysis is Hierarchical cluster technique. Hierarchical clustering has several advantages compared to the *k*-means and other partitioning clustering methods. Tree cluster algorithm has the ability to

computes distance between cluster based on similarity measures like portioning cluster analysis. Moreover, the *k*-means algorithm produces *k* clusters at the end, whereas in hierarchy of clusters method enables the user to explore clusters at any level of detail (or granularity).[16,17]

For example, in XML text document clustering, the cluster hierarchy may represent a topic hierarchy in the documents. Some studies have shown that agglomerative hierarchical clustering often

Table 1: Comparison of Cluster Algorithms

Characteristics	**K-Means Algorithm**	**EM Algorithm**	**Tree clustering Algorithm**
Methodology	Partitioning Method based on similarity Euclidean distance calculation	Partition and distribution method based likelihood.	Hierarchal categorization and joining methodology based on top down-bottom up approach using divide and conquer technique.
Type of Dataset	Real Dataset & Random Dataset	Real Dataset & Random	Real Dataset & Random
Advantages	Flexibility, Efficient and productive algorithm, Order independent.	Robust to noisy data, desire no.of cluster as input is possible, fast coverage, high dimensionality.	Applicable to any attribute type, flexibility, Ease of Handling high dimensionality, more versatile, produce better cluster for topic hierarchy.
Limitations	Difficult for arbitrary shape hierarchy. Produce cluster at the end of iteration.	Complex for large datasets. Only best for linear database.	Sensitive to outliers, inefficient compared to partition clustering algorithm. Space and computation complexities.
Repeatability	Yield different results on different runs. Lack consistency.	Different Result.	Same Results. Consistent
Shape of Cluster	Good for hyper Spherical and 2d Cluster.	Good for high dimension cluster	Also Good for non-spherical dataset
Complexity	Linear Time complexity	Complex than k-means algorithm	Quadratic time Complexity. Requires more space than other partitioning algorithm.

produces better clusters than the *k*-means method.[18] It can also find clusters of arbitrary shapes, e.g., using the single-link method. Hierarchical clustering also has several weaknesses. The time complexity of K-means algorithm is linear O(n). EM algorithm takes much time than k-means algorithm for same input. The time complexity of Tree clustering algorithm is quadratic.[15,17,18,19]

Tree clustering methods are sensitive to outliers. The main shortcomings of all hierarchical clustering methods are their computation complexities and space requirements, which are at least quadratic. Compared to the *k*-means algorithm, this is very inefficient and not practical for large data sets. [15,18,19]

III. Experiment & Result

Clustering algorithms work well with all kinds of data including categorical, numerical, and textual data. In this Experiment section, three algorithms(K-means, EM and Tree Cluster Algorithm) are compared for same XML datasets which is simulated using STATISTICA software. STATISTICA is a software tool supports a data mining technique which is developed by DELL. In Figure2. Sample dataset for organism is shown.[14,20]

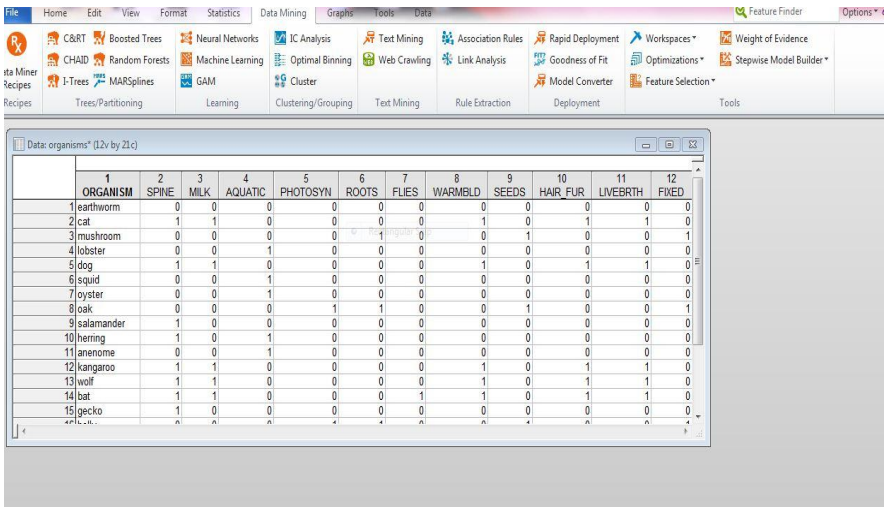

Figure 2: Sample XML dataset

k-Means Clustering is very different from Joining (Tree Clustering) and is widely used in real-world scenarios.[12] .For the given XML data set the K-means algorithm computes cluster using STATISTICA.

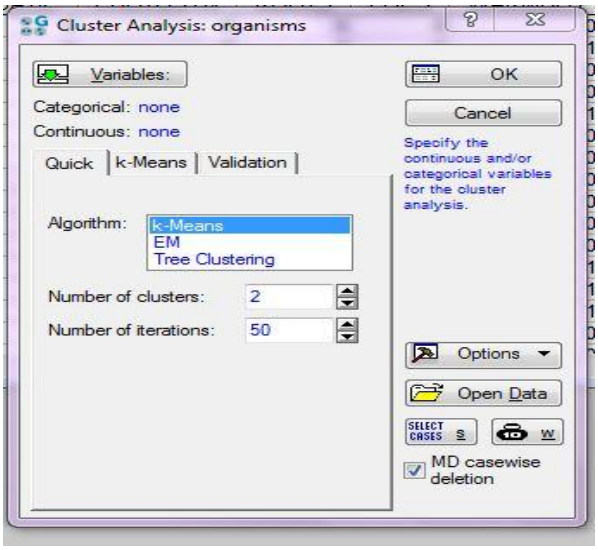

Figure 3: Selection of Cluster Algorithm

STATISTICA generates 5 clusters for 50 iteration using k-means algorithm is shown in figure4.

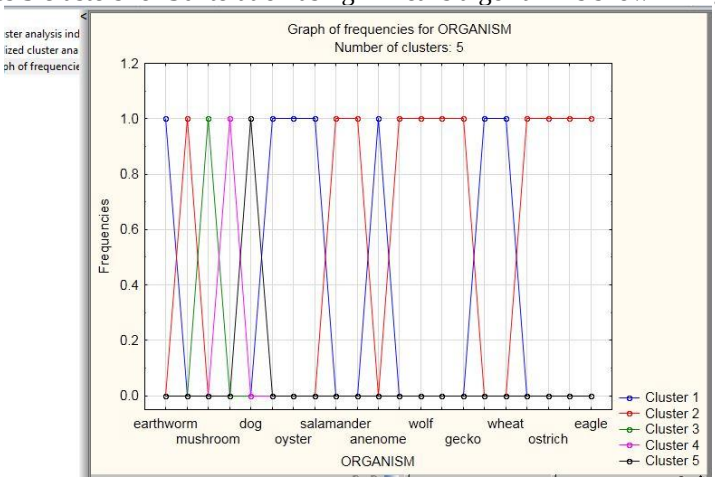

Figure 4: K-means Cluster formation

Frequency table for categorical variable: ORGANISM (organisms)
Number of clusters: 5
Total number of training cases: 21

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
earthworm	1	0	0	0	0	1
cat	0	1	0	0	0	1
mushroom	0	0	1	0	0	1
lobster	0	0	0	1	0	1
dog	0	0	0	0	1	1
squid	1	0	0	0	0	1
oyster	1	0	0	0	0	1
oak	1	0	0	0	0	1
salamander	0	1	0	0	0	1
herring	0	1	0	0	0	1
anenome	1	0	0	0	0	1
kangaroo	0	1	0	0	0	1
wolf	0	1	0	0	0	1
bat	0	1	0	0	0	1
gecko	0	1	0	0	0	1
holly	1	0	0	0	0	1
wheat	1	0	0	0	0	1
robin	0	1	0	0	0	1
ostrich	0	1	0	0	0	1
platypus	0	1	0	0	0	1
eagle	0	1	0	0	0	1

Figure 5: Frequency Table for K-means Algorithm

The EM (Expectation Maximization) Clustering technique is another tool offered in STATISTICA. The general purpose of this technique is also to detect clusters in observations (or variables) and to assign those observations to the clusters[20].

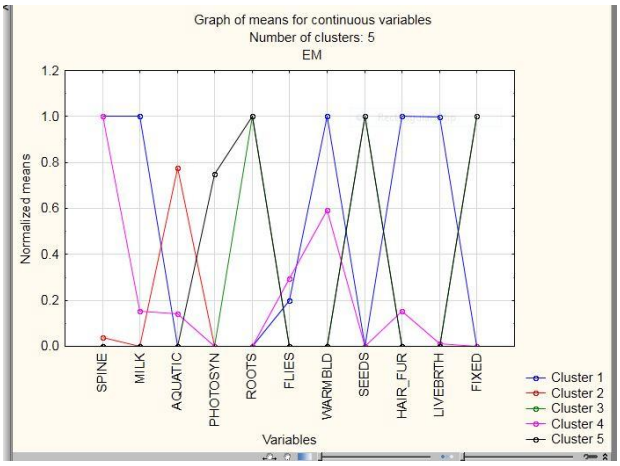

Figure 6: EM Cluster Formation

Classification probabilities (weights) for EM clustering (organisms)
Number of clusters: 2
Total number of training cases: 21

	Cluster 1	Cluster 2	Final classification
1	1.000000	0.000000	1
2	1.000000	0.000000	1
3	1.000000	0.000000	1
4	1.000000	0.000000	1
5	1.000000	0.000000	1
6	1.000000	0.000000	1
7	1.000000	0.000000	1
8	0.000000	1.000000	2
9	1.000000	0.000000	1
10	1.000000	0.000000	1
11	1.000000	0.000000	1
12	1.000000	0.000000	1
13	1.000000	0.000000	1
14	1.000000	0.000000	1
15	1.000000	0.000000	1
16	0.000000	1.000000	2
17	0.000000	1.000000	2
18	1.000000	0.000000	1
19	1.000000	0.000000	1
20	1.000000	0.000000	1
21	1.000000	0.000000	1

Figure 7: Classification Probability Table

The output for EM Cluster Analysis will be quite similar to that of k-Means Clustering. Along with the final cluster classification output, EM Cluster Analysis provides probabilities of cluster membership. In the output spreadsheet below, these probabilities are shown for the first several cases in the data set.
The purpose of tree cluster algorithm is to join together objects (e.g., animals) into successively larger clusters, using some measure of similarity or distance. A typical result of this type of clustering is the hierarchical tree**. [20]**

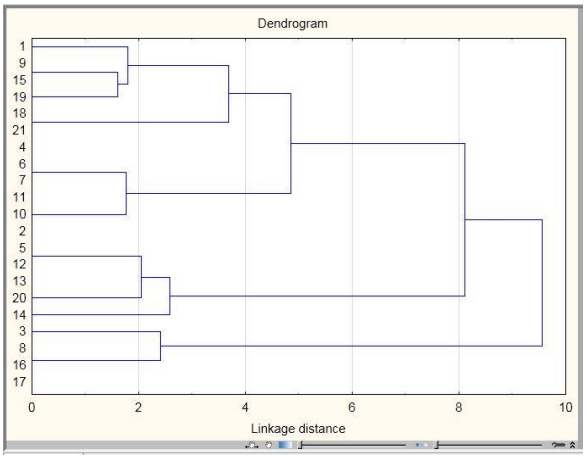

Figure 8: Tree Cluster

	Summary for tree clustering (organisms) Number of clusters: 3 Total number of cases: 21
Algorithm	Tree Clustering
Amalgamation method	Ward
MD casewise deletion	Yes
Total number of cases: %d	21
Number of clusters	3

Figure 9: Summary of E tree clustering Algorithm

Tree cluster algorithm produced different results compare to other portioning algorithms. Hierarchical clustering doesn't need the number of clusters to be specified. Hierarchical Clustering can give different partitioning depending on the level-of-resolution. Tree cluster algorithm provides three no. of clusters for the given data set. From the analysis of three cluster algorithms first two algorithms (K-means and EM algorithms) yield same result for the given data sets. As we discussed earlier in the previous section (comparison of Clustering algorithms) EM algorithm consume more time than K-means algorithm for computation. So k-means algorithm is the best among portioning algorithm. As the number of records increase the performance of hierarchical algorithm goes decreasing and time for execution increased [15]. Now comparing the K-means algorithm with hierarchal algorithm, K-mean algorithm also increases its time of execution but as compared to hierarchical algorithm its performance is better. Hierarchical algorithm shows more quality as compared to k-mean algorithm.

IV. Conclusion

This paper presented the comparative study on three clustering algorithms (K-mean, EM and Tree clustering Algorithm). Three algorithms are compared using STATISTICA tool on XML datasets. By comparing the results of three algorithms on same XML datasets K-means algorithm works best among three. But Tree clustering algorithm produces best quality cluster. From this study, the results shown that combing the K-means algorithm with Tree cluster algorithm will produce best cluster and have better performance for XML Data clustering.

References

[1]. Yuekui Yang, Yajun Du, Yufeng Hai, Zhaoqiong Gao, "A Topic-Specific Web Crawler with Web Page Hierarchy Based on HTML Dom-Tree", 2009 IEEE , Asia-Pacific Conference on Information Processing.
[2]. Dragos Arotaritei,Sushmita,Web mining: a survey in the fuzzy framework, Fuzzy Sets and Systems 148,p.5-19,2004.
[3]. LI Guoliang , FENG Jianhua, ZHOU Lizhu ," Keyword Searches in Data Centric XML Documents Using Tree Partitioning, Tsinghua Science and Technology, February 2009, 14(1): 7-18
[4]. Ritu Khatri, Kanwalvir Singh Dhindsa, Vishal Khatri, " Ivestigation and Analysis of New Approach for Intellignet Semantic Web Search Engines, IJRTE April 2012.
[5]. Amit Mishra, Sanjay Kumar jain, " A Survey on question answering system with classification, Journal of King Saud University - Computer and Information Sciences
[6]. Megha Gupta[1], Naveen Aggarwal[2], "Performance Analysis of Classification Techniques
[7]. on XML Dataset", IJCST Vol. 1, Iss ue 1, September 2010
[8]. Gurpreet Kaur and Naveen Aggarwal, "Exploiting Hierarchal Structure of XML Data Using Association Rule Analysis", International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012
[9]. Web Data Mining-Part of the series Data-Centric Systems and Applications pp 117-150
[10]. N. Koteswara Rao, G. Sridhar Reddy, "Discovery of Preliminary Centroids Using Improved K- Means Clustering Algorithm",IJCSIT 2012.
[11]. K. A. Abdul Nazeer, M. P. Sebastian," Improving the Accuracy and Efficiency of the k-means Clustering Algorithm" Proceedings of the World Congress on Engineering 2009 Vol I.
[12]. Micheal Steinbach, Levent Ertoz and Vipin Kumar, " The Challenges of Clustering High Dimensional Data".
[13]. K. Rajeswari, Omkar Acharya, Mayur Sharma, Mahesh Kopnar, Kiran Karandikar, "Improvement in k-Means Clustering Algorithm Using Data Clustering", IEEE Computer Society Washington, DC, USA ©2015
[14]. https://en.wikipedia.org/wiki/Expectation–maximization_algorithm
[15]. http://documents.software.dell.com/Statistics/Textbook/cluster-analysis
[16]. Archana, "Study and Comparison of Partition Based and Hierarchical Clustering", International Journal of Advanced Research in Computer Science and Software Engineering 4(5), May - 2014, pp. 583-587
[17]. https://wwwusers.cs.umn.edu/~kumar/dmbook/ch8.pdf
[18]. https://en.wikipedia.org/wiki/Cluster_analysis
[19]. Ssujata Kolhe, Dr.Sudhir Sawarkar, " Review of Document Clustering Techniques: Issues, Challlenges and FeasibleAvenue" International Journal of Advanced Research in Computer Science and Software Engineering 5 (4), April- 2015, pp. 1-5
[20]. https://en.wikipedia.org/wiki/DBSCAN
[21]. http://www.statsoft.com/portals/0/products/data-mining/clustering.pdf