International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

An Overview of Clustering Algorithm and Collaborative Filtering Method through E-Commerce Data Perspective

R. Gowri, Associate Prof.,
Dept. of IT, SMVEC, Puducherry,
India,
Ashish Kumar, UG student,
Dept. of IT, SMVEC, Puducherry,
India

Arvind M. J., UG student,
Dept. of IT, SMVEC Puducherry,
India
Jeric Rajan K., UG student,
Dept. of IT, SMVEC, Puducherry,
India,

***Abstract* - Innovations in technology and greater increase of data sets have presided over today's age of marketing. An umbrella term for the explosion in the quantity and diversity of users and their need have invoked Big Data technology. Rapidly growing challenges among ecommerce dealers to adopt greater number of customer into their folk has attended a presence of new technology. As a result there is large numbers of technology and method has been adopted. This paper has been divided into two parts. First part delivers a glances of different Clustering Algorithm which has been adopted up to now for easy access of Big Data. The second has been towards Collaborative filtering method which has been invoked for recommender system in ecommerce market. We propose a theoretical survey of these two mechanism with regards to e-commerce data.**

Keywords - clustering, collaborative filtering, service similarity, recommender system, E-Commerce Data

I. INTRODUCTION

Current digital era has involved with masses which interacts with large volume of data throughout the network in 24*7 manner. Involvement of masses have fetched a large volume of structured or nonstructural digital data. These data comes from different sources and services which were not actually available a few decades ago. Massive quantities of data are produced by and about people, things, and their interactions. This data comes from available different online resources and services which have been established to serve their customers. Generally speaking, Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time" is on the rise. Services and resources like Sensor, Networks, Cloud Storages, Social Networks and etc., produce big volume of data and also need to manage and reuse that data or some analytical aspects of the data.

The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. A class of emerging high-end applications need to perform efficient data computing on massive datasets. This kind of applications require the data processing technology to have both data-intensive and computation-intensive abilities.

A challenge for data clustering resides in the substantial growth of data generated in many fields over the years. This growth requires the distribution of large data sets in separate repositories, also called data sites. In many scenarios, the data are naturally distributed, i.e., have been generated and stored in different data sites. Large distributed data sets demand computational techniques that are able to extract relevant information with good computational performance and scalability. In order to accomplish these requirements, low data transmission cost is also necessary.

The main goal of this paper is to study different clustering algorithm used in Big Data as well as a glance of how it will suit to E-Commerce data. The second section involves with different criteria for surveying clustering technique .The third section is comparative study of different clustering algorithm in terms of big data.The fourth is Characterizing E-Commerce Data in terms of Big Data**.**Fifth section shows the conclusion of surveyfollowed by sixth section as reference section.

II.CRITERIA FOR SURVEYING CLUSTERING TECHNIQUE IN TERMS OF BIG DATA

While categorizing Big Data, we need to consider the following three categories: Volume, Velocity & Variety. Different clustering methods could be evaluated under comparison of these three category. Each category too is associated with some criterion. The criterion could be explained while keeping in view of 3V characteristics of Big Data. The 3V is- volume, velocity and variety. Volume is the foremost category while surveying clustering in terms of Big Data. Volume could be termed as amount of data. The second category which has to take in account is Variety which refers to number of types of data. The last category is Velocity which refers to speed of data processing.

Notwithstanding a vast number of surveys for clustering algorithm is available as live in various domain (such as machine learning, information retrieval, pattern recognition, bio-informatics & information retrieval), it is difficult for developer to decide an appropriate one which would be almost accurate result oriented. This is due to some of the following barrier which arose from existing clustering mechanism in terms of large data sets. (i). the characteristics of clustering algorithm did not check well; (ii) It did not well suited to retrieve all expected information from data sets; (iii) No comparative analysis has been performed keeping in view of large data sets. The situation moves too complicated while we subcategorize Big Data into growing sets of data in e-commerce. Spurred by these reasons the following survey paper has attempted to review the field of basic clustering algorithm, in terms of their possibility in e-commerce. After completion of surveying with atheoretical evaluation the paper reached out to the following objectives.

- To explore a framework which would describe different evaluation factor of clustering e-commerce data.
- To describe an efficient clustering way which could be tied up with collaborative filtering method for an accurate recommendation to the users.
- To finalize an accurate clustering algorithm which would suit to our proposed project.
- To provide a collection of characteristics of large data sets which could be used by other to develop new technology and terminology or to enhance pre-existing technology.

Therefore as a whole the paper has focused on surveying clustering mechanism used in e-commerce data sets. In order to reach this goal, we first summarize an introduction to basic clustering mechanism, followed by displaying survey table of [1] Fahad et al. in large data sets and finally explaining our proposed survey will an experimental evaluation.

Basic Clustering Mechanism:

- *Hierarchical-Based:* Data are organized in a hierarchical manner depending on the nature of nearness. Nearness are obtained by the intermediate nodes. Hierarchical clustering methods can be agglomerative or divisive. An Agglomerative is known as bottom-up approach and Divisive clustering is top-down. An agglomerative clustering starts with one object for each cluster and recursively merges two or more of the most appropriate clusters. A divisive clustering starts with the datasets as one cluster and recursively splits the most appropriate cluster. The process continues until k cluster is obtained. The k cluster refers to finally reduced set of cluster from a given starting set of cluster as n.
- *Partitioning-Based:* Here data objects are classified into number of groups. Each group is known as a cluster. Each object must be associated with a single group or cluster. The length of group could be self-definable at an initial stage. These clusters should fulfil the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group.
- *Density-Based:* Here, data objects are separated based on their regions of density, connectivity and boundary. The boundary could be defined as- (1) Non-Convex Shape and (2) Arbitrary Shape. They are closely related to point-nearest neighbors. Created cluster grows towards relative density of data.
- *Grid-Based:* The space of data object is quantized into grids. It is used to discover cluster of any shape. The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid.
- *Model-Based:* Data are clustered according to predefined mathematical assumptions and theorem. Generally a probability operation is performed for given data sets to cluster it. Before applying this algorithm we must ensure that the given data set has strong statistical basis. The clustering process using Model Based is fast coverage.

Clustering in terms of large data sets.

Figure 1 shows basic clustering mechanism and different algorithm developed under these mechanism in terms of capability of handling large data sets.

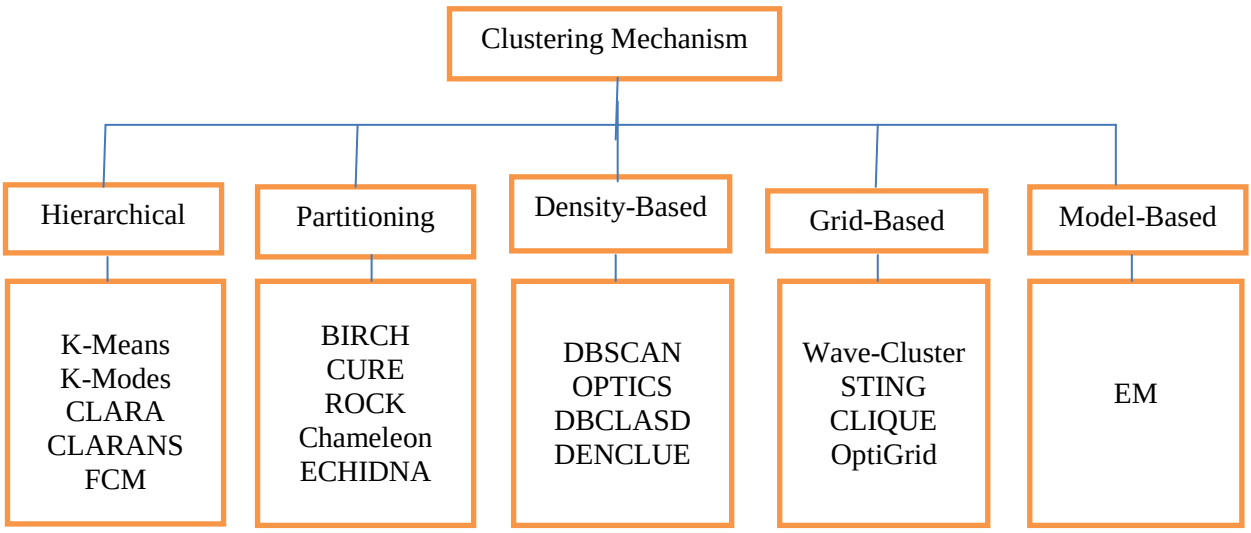

Fig. 1. Tabulation of Clustering Mechanism & their example Algorithm

The figure has been derived from [2].Fahad et al. while escaping some of the algorithm (K-Means, PAM, COBWER, CLASSIT, & SOMs) due to their handling of small scale of data-sets. As Big Data is associated with large content of data-sets, we understood the ignorance of these all algorithm. Though K-Mean is said to be as basic clustering algorithm, which uses mean value of numerical data sets as a key and then find the mechanism to cluster it.

III. COMPARATIVE STUDY OF DIFFERENT ALGORITHMS IN TERMS OF BIG DATA

Thetable shown by [2] Fahad. Et al. has been derived under different criterion which falls under 3Vs of Big Data. Keeping in Mind of these criteria we have shortlisted them with respect to e-commerce data. While moving with e-commerce data, it is necessary to check the type of dataset must not be numerical. As the product which could be categorized under e-commerce data may be character or string. These data sets have majority of English-alphabet despite of presence of some special character (@,#,$ etc.). Special character could be ignored because of its minor presence (0-5%).

We explain different criteria which falls under 3Vs of Big Data and relatively suits for e-commerce data sets. The criteria as a whole delivers an efficient evaluation mapping method for large data sets of e-commerce data. While sub-dividing these criteria, they deliver 3 different approach of large data sets, namely- volume, velocity & variety.

IV. CHARACTERIZING E-COMMERCE DATA IN TERMS OF BIG DATA:

E-commerce data could be characterized with 3V criterion of Big Data. In this section we define these property and shows the key criterion with related to Volume, Variety and Velocity characteristics of Big Data.

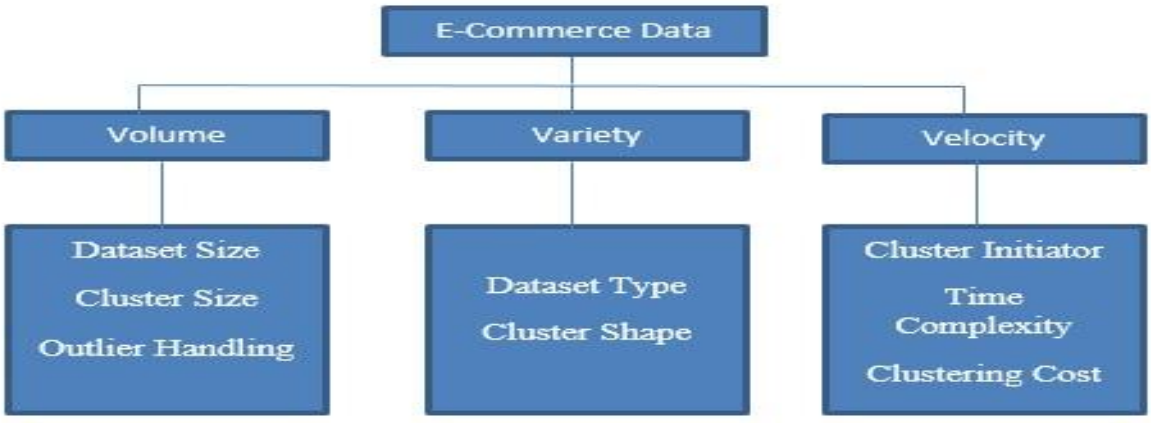

Fig.2. Tabulating properties of 3Vs of E-Commerce Data

- **Volume:** It refers to ability of clustering large content of products and customer details. To select an appropriate algorithm with respect to volume the following criteria is used: (i) Dataset size, (ii) Cluster size, (iii) Outlier Handling.
- **Variety:** It refers to different shape or appearance of products in terms of data. To consider an appropriate algorithm with respect to Variety the following criteria is used: (i) Dataset Type, and (ii) Cluster Shape.
- **Velocity:** It refers to processing speed of clustering algorithm with regards to e-commerce data. To take an appropriate clustering algorithm with regards to velocity we consider the following criteria: (i) Cluster Initiator, (ii) Time Complexity, (iii) Clustering Cost.

Table 1

Algorithms	Dataset size	Cluster Size	Outlier/Noise Handling	Cluster Shape	Data Set Type	Clustering Initiator	Time Complexity	Clustering Cost
K-Means	Large	Pre defined k	No	Non-Convex	Ideal	Centroid	O(nkd)	Low
K- Medoids	Large	Pre defined k	Yes	Non-Convex	Random	Medoid	O (n^2dt)	High
K-Mode	Large	Pre defined k	No	Non-Convex	Random	Top of stack	O(n)	Low
Chameleon	Large	Threshold value	No	Arbitrary	Hierarchical	Centroid	O (n^2)	High
Echidna	Large	Pre defined k	No	Non-Convex	Hierarchical	Any data	O(N*B(1+log m))	High
Wave Clustering	Large	Threshold value	Yes	Arbitrary	Hierarchical	Single grid data	O(n)	Low
Sting	Large	Pre defined k	Yes	Rectangular	Hierarchical	Single grid data	O(k)	Low
Clique	Large	Pre defined k	No	Rectangular	Ideal	Single grid data	O(ck+mk)	Low
OptiGrid	Large	Threshold value	Yes	Rectangular	Hierarchical	Single grid data	O(nd)	Low
EM	Large	Threshold value	No	Non-Convex	Random	Random data	O(knp)	High

- ***Dataset size:***It implies to data sets which do we need to store. It could be large or small. In Big Data perspective it must be more. So the algorithm dealing with smaller amount of data has not been taken in count.
- ***Cluster size:*** It implies reduced range of data sets. It could be- Pre Defined K(Before starting clustering this is defined in terms of getting k cluster.) or Threshold value (The at-least reduction from which next reduction is not possible.) The K value or threshold value must be lesser to total number of data-set n. The lesser the K value will be, the Compaq the Cluster efficiency will be.
- ***Outlier/Noise Handling:*** Different clustering algorithm has its own capability of handling noisy data or the data which could be said as an outlier.
- ***Cluster Shape:*** The reduced data set can reside in any shape of cluster among these all- Non-Convex, Arbitrary and Rectangular. A Non-Convex clustering shape is named as after an arbitrary number of iteration [10] the data-sets start reduction and make a cluster. An arbitrary cluster shape can be concave, convex or nested. This cluster shape does not have any pre-defined shape. It could be adopted as in web stream data clustering. This is because this shape is obtained at live process in any shape with regards to type and amount of data. A rectangular shape could be found only in grid-based clustering, as in this data-sets are divided into smaller grids and then we follow clustering by following reduction mechanism. After reduction a grid becomes same in previous shape (rectangular) but in reduced size.
- ***Data Set type:*** The data set defined for a product can be of different type. It can be Ideal (Numerical) - whose value could not change, or Random (Categorical) - which shows alphabetical data with some special character data, or Hierarchical (Multivariate)- It implies both the categorical and numerical data.
- ***Clustering Initiator:*** It implies to a single value which is used to start the clustering process. This could be used with regards to "term reduction matrix" [16]. It varies from iteration to iteration and algorithm to algorithm.
- ***Time Complexity:*** Most common characteristics used to evaluating any algorithm. It is calculated in terms of- no. of variables used and no. of iteration etc.
- ***Clustering Cost:*** According to the uses of clustering mechanism it could be low or high.

Table 1 shows the property as a whole in terms of big data.

Algorithm Comparison of Clustering with regards to Collaborative Filtering (CF):

Algorithm	Proposed By	Clustering	CF	Remedy
ClubCF	[1] Rong Hu et al.	Hierarchical	Item Based	Service Similarity has not been used
LSA	[16] Haytham et al.	Partitional	Not used	No change in result compare to traditional recommender system
ECT using K-Mean	[17] Xuan & Zhinjun et al.	Partitional	Not used	Used only for e-commerce transaction
FTCA	[18] Sudhamathy et al.	Partitional	User Based	No item characteristics has mined

Fig:3. Collaborative Filtering Algorithm comparing with Clustering Algorithm.

V. CONCLUSION

In this paper we have studied different clustering method which has been used in different domain for mining data characteristics. Thus this survey provides a depth in clustering algorithm as a whole and at a place which could be used by researchers to further explore knowledge and characteristics from large data sets. We proposed a categorized framework for clustering mechanism with regards to the property of Big Data. To use more precisely in E-Commerce data, which is rapidly growing, the categorized framework developed from theoretical view point will fetch more researcher interest to find more accurate solution. Thus future birth of new clustering algorithm as well new mechanism used in e-commerce data along with clustering, could be inherited from this survey.

VI. REFERENCES

1. Rong Hu, Member, IEEE, Wanchun Dou*, Member, IEEE, Jianxun Liu, Member, IEEE "ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application", IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING-2014.
2. A. Fahad, N. Alshatri, Z. Tari, Member, IEEE , A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING-2014.
3. Dheeraj Kumar, MarimuthuPalaniswam, iSutharshanRajasegarar, "clusiVAT: A Mixed Visual/Numerical Clustering Algorithm for Big Data", IEEE International Conference on Big Data-2013
4. Timothy C. Havens, Senior Member, IEEE, James C. Bezdek, Life Fellow, IEEE, Christopher Leckie, Lawrence O. Hall, Fellow, IEEE, and MarimuthuPalaniswami, Fellow, IEEE, "Fuzzy c-Means Algorithms for Very Large Data", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 20, NO. 6, DECEMBER 2012
5. Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-2012, 06-08DECEMBER, 2012
6. G. Anuradha, Bidisha Roy, "Suggested Techniques for Clustering and Mining of Data Streams", International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)-2014
7. Sachchidanand Singh, Nirmalasingh, "Big Data Analytics", International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India-2012
8. S. VikramPhaneendra, E. Madhusudhana Reddy, "Big Data - Solutions for RDBMS Problems - A Survey", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013
9. White Paper, UN Global Pulse, "Big Data for Development-Challenges and Opportunities", May-2012
10. ZHEXUE HUANG, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, Kluwer Academic Publishers, 1998.
11. RaghaviChouhan, Abhishek Chauhan, "An Ameliorated Partitioning Clustering Algorithm for Large Data Sets ", International Journal of Advanced Research in Computer and Communication Engineering -2014.
12. Shehroz S Khan, "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation", IJCAI-2007
13. George KarypisEui-Hong (Sam) Han Vipin Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", IEEE-2007
14. AbdunNaser et al. "Critical infrastructure protection-Resource efficient technique to improve detection of less frequent pattern in network traffic", ELSEVIER-2010
15. Preeti Baser et al. "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets", International Journal of Computer Science & Communication Networks-2013
16. Haytham et al., "Automatic Clustering of e-Commerce Product Description", Journal of Applied Computer Science & Mathematics-2012
17. Xuan &Zhijun, "Clustering Analysis on E-commerce Transaction Based on K-means Clustering", JOURNAL OF NETWORKS-2014
18. Sudhamathy, "Fuzzy Temporal Clustering Approach for E-Commerce Websites", International Journal of Engineering and Technology-2012