Enhancement of Big Data Using Clustering Mechanism

Authors
Ankita Sehgal¹, Dr. Anuj Kumar Sharma²

¹P.G. Student, Department of Computer Science & Engineering, Om Institute of Engineering & Technology, Hisar, Haryana, India
²Associate Professor, Department of Computer Science & Engineering, Om Institute of Engineering & Technology, Hisar, Haryana, India
Email: ankitasehgal.hsr@gmail.com, deanoitm@gmail.com

ABSTRACT:
Distributed Networking is a distributed computing network system, said to be distributed when computer programming & data to be worked on are spread out over more than one computer. Usually, it is implemented over a network. In charge to make clever decisions together for people and for thoughts in IoT, folder mining technologies are integrated with IoT technologies for decision making support & system optimization. Data mining involves discovering novel, interesting, & potentially useful patterns from data & applying algorithms to extraction of hidden information

Keyword: Clustering, Data Warehouse, Association, Data Mining.

INTRODUCTION
Distributed Networking is a distributed computing network system, said to be distributed when computer programming & data to be worked on are spread out over more than one computer. Usually, it is implemented over a network.

Prior to emergence of low-cost desktop computer power, computing was generally central to one computer. Such that centers still exist, distribution networking applications & data operate very powerful over a mix of desktop workstations, local area network servers, regional servers, Web servers, & other servers. This is principle that a client computer could provide certain capabilities for a user & request others from other computers that provide services for clients. The Web's Hypertext Transfer Protocol is an example of it idea. Enterprises that have grown within scale over years & those that are continuing to increase are finding it extremely challenging to manage their distributed network within traditional client/server computing model. Recent developments within area of cloud computing had been opened up new possibilities. Cloud-based networkings vendors have been grow to sprout solve solutions for firm distributed networking needs. Whether it turns out to revolutionize distributed networking area or turns out to be another craze remains to be seen.

LITERATURE REVIEW
Several techniques for learning statistical models have been developed recently by researchers in machine learning & data mining. All of these secrets must address a similar set of representational algorithmic choices & must face a set of statistical challenges unique to learning from relational data.

However, little this work had been made good use of research in other areas, such as social network analysis & statistics. Cross-disciplinary efforts & joint research efforts should be encouraged to promote rapid
development & dissemination of useful algorithms & data representations. This work should focus on unique statistical challenges raised by relational data.

Data mining based on Neural Network & Inherited Algorithm is researched in detail & key technology & ways to achieve data mining on Neural Network & Genetic Algorithm are also surveyed. This paper also conducts a formal review of area of rule extraction from ANN & GA. If conception of computer algorithms being based on evolutionary of organism is surprising, extensiveness within which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new & important area of research & ANN itself is a very suitable for solving problems of data mining because its characteristics of better robustness, self-organizing adaptive, distributed storage & high degree of fault tolerance. Commercial, educational & scientific applications are increasingly dependent on these methodologies.

Data mining is a powerful & a new field having various techniques. It converts raw data into useful information in various research fields. It helps in finding patterns to decide future trends in medical field.

A grouping of data objects such that objects within a group are similar to one another & different from (or unrelated to) objects in other groups. This paper intends to study & compare different clustering algorithms. These algorithms include K-Means, Farthest First, CURE, Chameleon algorithm. All algorithms are compared to each other on basis of their pros & cons, similarity measure & working, functionality & time complexity. In his paper they present brief & easy comparison between different clustering algorithms. They also evaluate these algorithms on different datasets & present results through tables.

OBJECTIVES
User should be able to organize a huge data set.
Investigation should be done the limitations of existing clustering mechanism.
Study of existing clustering mechanism such as K-MEAN.
The customized algorithm must be made in order to overcome limitation of existing clustering algorithm.
Comparative analysis should be made of traditional and proposed algorithm.

RESEARCH METHODOLOGY
Data mining techniques for effective automated discovery of earlier unknown, valid, novel, useful & understandable patterns in large databases. Patterns must be actionable so that they might be used in an enterprise’s decision making process. It is generally used to business intelligence firm, & financial analysts, but this is more used in sciences to extract information from enormous data sets generated by modern experimental & observational methods.

An example for a data mining scenario might be In context of a market, if a mining analysis observes that people who buy pen tend to buy pencil too, then for better business results seller could place pens & pencils together.

Classification- Here data example had to be classified into one of mark classes which are already known or defined. One of examples could be a client had been to be classified or a defaulter with in a credit card transaction data base, given his various demographic & previous purchase Characteristics.

Estimation- Like classification, purpose of an estimation model is to determine a value for an unknown output attribute. Moreover, unlike classification, output power for an important problem is numeric rather than categorical. An example could be “Estimate salary of an individual who owns a sports car.
Prediction- It is not easy to make different prediction from classification or approximation. Only dissimilarity is somewhat determining current behavior, predicts a future outcome. Output attribute could be categorical or numeric. An example could be “Predict next week’s closing price for Dow Jones Industrial Average explains construction of a decision tree & its predictive applications.

Association rule mining -Here interesting hidden rules called association rules with in a large transactional data base is mined out. For e.g. rule {milk, butter->biscuit} provides information that whenever milk & butter are purchased together biscuit is also purchased, such that these items could be placed together for sales to increase overall sales of each of items.

Clustering- Clustering is a special type of classification with in which target classes are unknown. For e.g. given 100 customers they have to be classified based on certain similarity criteria & this is not preconceived which are those classes to which customers should finally be grouped into.

RESULTS AND FINDINGS

K-Means algorithm converges to local minimum. Before k-means converges, centroid computed number of times, & all points are assigned to their nearest centroid, i.e., complete redistribution of points according to new centroid, this takes $O(nkl)$, where $n$ is number of points, $k$ is number of clusters & $l$ is number of iterations. In existing enhanced k-means algorithm, to obtain initial clusters, this process requires $O(nk)$. In our research cluster generated previously is rechecked clusters where no data points are allocated to a cluster under consideration during assignment phase are eliminated.

The consists of results and discussion in several circumstances. K-Means algorithm converges to local minimum. K-means clustering is very Fast, robust & easily understandable. If data set is separated from one other data set, then it gives best results. Clusters do not having overlapping character & are also non-hierarchical within nature. One more problems with K-means clustering is that empty clusters are generated during execution, if within case no data points are allocated to a cluster under consideration during assignment phase.

REFERENCES

10. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research 11: 2533–2541. original title, "Practical machine learning", was changed ... term "data mining" was [added] primarily for marketing reasons.
17. Proceedings, International Conferences on Knowledge Discovery & Data Mining, ACM, New York.
18. SIGKDD Explorations, ACM, New York.