IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Decisive Mining for Heterogeneous Data

Prakash R.Andhale[1], S.M.Rokade[2]

M.E.Student, Department Of Computer Engineering, SVIT, Nasik, India [1]

Head of Department, Computer Engineering, SVIT, Nasik, India [2]

Abstract: Heterogeneous data is the term concern with the data from any number of sources largely known, unknown, unlimited, and in many varying format. The heterogeneous data are now rapidly expanding in all technical, biological, physical and medical science with the help of fast development of storage system, networking and the collection in capacity of data. The paper presents the characteristics of HACE theorem which provides the features of heterogeneous data and proposes a model processing on heterogeneous data from the view data mining. This information extraction model involves the information extraction, data analysis and provides the security and privacy mechanism to the data.

Keywords: Data Mining, Information Extraction, Big Data, Heterogeneous Data, HACE Theorem.

I. INTRODUCTION

The term heterogeneous data is widely used in everywhere in the form of online and offline fashion. It is not limited for only the computer system but also comes under the information technology which is now part of almost all technology and various fields of study and business.

The Facebook application is the social media site on this number of people post their photos, videos, text messages and the assuming size of these photos and videos is more than 2 Megabytes(MB),this requires the more than 4 terabytes(TB) storage for single day.

The above example illustrate that the tremendously growth in heterogeneous data. By exploring the large volume of heterogeneous data, the useful; information and knowledge will be extracted for the future required actions.

The main motivation of this paper is to understand the importance of huge, complex and information rich-data set in science , business and engineering field, also the discovering the knowledge and information from massive data to improve the efficiency of information extraction methods.

A. *Objectives:*

1) Building prediction models from heterogeneous Data streams. Such models can adaptively adjust to the dynamic changing of the data.
2) To propose Model for discover the useful knowledge and information from heterogeneous data.
3) A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.
4) To clustering the data object fromprocessed heterogeneous data using K-means algorithm.

II. LITERATURE SURVEY

Dynamic networks have recently being recognized as a powerful abstraction tomodel and represent the temporal changes and dynamic aspects of the data underlying many complex systems. Significant insights regarding the stable relationalpatterns among the entities can be gained by analyzing temporal evolution of thecomplex entity relations. This can help identify the transitions from one conservedstate to the next and may provide evidence to the existence of external factorsthat are responsible for changing the stable relational patterns in these networks.

This paper presents a new data mining method that analyzes the time-persistentrelations or states between the entities of the dynamic networks and captures allmaximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real-world applications show that themethod is ancient and scalable [2].

HACE Theorem:

1) User Perspective huge data collection from various data resources.
The heterogeneous data are comes from various types of sites like Facebook, Orkut, Twitter, Snapdeal, Gamiletc .Because every information collector has his own views to collect or represents the data recording.

2) Autonomous Sources with Unstructured Control.
Autonomous sources with unstructured control are the main characteristics of heterogeneous data application. Being autonomous, each data sources have their own ability to collect and produce the information without having any centralized control.

3) Complex and Evolving Linkage.
As the rapidly growth in heterogeneous data need to decrease the complexity and relationship among data, which leads to focus on finding the each observation.

III.PROPOSED SYSTEM

A. *Architecture Of Proposed System*

DOI 10.17148/IJARCCE.2015.412101

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

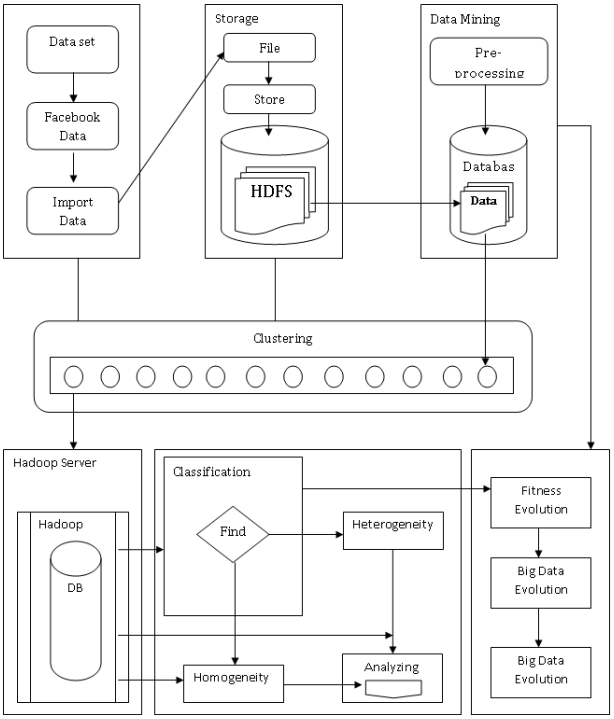

Fig. 1 Proposed System Architecture

B. Methodology and Workflow:

In the proposed system, we collect the dataset from the UCI (University Of Callifornia, Irvine) machine learning website. UCI is the collection of data; it contains the 'n' number of data. After selection of the dataset for the process, this has to be pre-processed.

In this pre-processing, insert the dataset into the database by using to tokenization concept. The tokenization process is elimination of unwanted symbol or extra spaces in the dataset. After preprocessing process done on dataset, next process is a clustering. The aim of clustering process is to find out the similarities between the data object clustering dataset using K-means algorithm .After done with clustering process , clustered data set are moved to the database for generating the graph and searching dataset.

C.Algorithms

The set of n object are portioning into k cluster using k means algorithm.

Step:

1. Choose the number of cluster object From D as initial cluster centers;
2. Repeat.
3. Reassign each object to the cluster to which the object is most similar based on the mean value of this object in this cluster;
4. Update the cluster mean; i.e. calculate the mean value of the objects for each cluster.
5. Until no change

IV. CONCLUSION

To explore the heterogeneous data, we have to analyze the various challenges at the data application and system levels. To support the information extraction from the heterogeneous data .The challenge data level in to handle the unstructured data. At the model level the key challenge in models requires carefully designed algorithms to analyze model relationship between various sites. At the system level the required challenge that the information extraction framework needs to consider complex association between data sets, models, and data sources, along with their evolving challenge with possible factor. We regard the information extraction from the heterogeneous data as an engineering trend and the essential for information extraction in all science and engineering domain.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member, IEEE "Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1 97-107,Jan-2014.

[2] Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 603-630 .

[3] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," Science, vol. 336, no. 6077, p. 22, 2012.

[4] A. Rajaraman and J. Ullman, Mining of Massive Data Sets. Cambridge Univ. Press, 2011.

[5] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[6] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[7] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.