Commercial usage using Big Data

Smita Dhawane,Shadab Khan,Zainulabedin Shaikh

Prof. Pavan Kulkarni, Dept. Of Computer Engineering, Trinity College of Engineering and Research, Pune University, Pune, India

---***---

Abstract - *Big data is concern with huge amount of data which includes complexity,multiple data sets as we know technologies are increasing rapidly so the data usage is expanded day by day on social media every seconds 1 million or 1 billion posts or data are updated so it has become very difficult to manage data by traditional data base than concept of Data Mining arises. In data mining there are different methodologies which are used to manage such as clustering, frequent patterns etc. This paper represents HACE theorem which is used to manage different type of data like organizational, educational, industrial, social data along with security and accuracy.*

***KeyWords*:BigData,Hadoop,Hive,DataMining,Clustering ,Hace.**

1.INTRODUCTION

Data mining ,is the operation of examine data from various views and summaries it into meaningful data - data that can be used to gain revenue, cuts costs, or both. Data mining software is one of a number of analytic tools for recognizing data. It allows users to study data from many different attribute or angles, categories it, and summaries the relation identified. Data mining is the activity of uncovering correlative or patterns among heaps of fields in large relational databases.

1.1 LITERATURE SURVEY

1 "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Dec. 2012.

Author:. R. Ahmed and G. Karypis

Dynamic networks have just being identify as powerful idea to model and denote temporal changes and dynamic aspects of core data in difficult system. To recognize the transitions from one preserved to the next and it give confirmation to previous of external factors that are accountable for changing relational patterns in network. This paper presents a new data mining technique that analyzed states between entities of dynamic network and identify maximum non redundant path of stable relational states.[1]

2. "Novel Approaches to Crawling Important Pages Early"

Author: M.H. Alam, J.W. Ha, and S.K. Lee

In data mining web crower is used web application like web search engine, web archives and directories which maintain web page designed algorithms utilized different quality including title of page, and topic significance. The trial using openly available data sets to study the result of every feature on crawl ordering and estimate the performance of different algorithms.[2]

3. "Identifying Influential and Susceptible Members of Social Networks"

Author : S. Aral and D. Walker

Recognize social power in networks is critical to understanding how behaviors spread. We present a method that uses in randomized test to recognize influence and receptiveness in networks while avoiding the biases inbuilt in traditional estimates of social contagion. Interference in a envoy sample of 1.3 million Facebook users showed that younger users are more tendency to influence than older users, men are more significant than women, women significant men more than they influence other women, and married individuals are the least susceptible to influence in the decision to adopt the product offered. Analysis of influence and susceptibility mutually with network structure exposed that influential individuals are less disposed to influence than non influential single and that they cluster in the network while susceptible individuals.[3]

4. "Analyzing Collective Behavior from Blogs Using Swarm Intelligence,"

Author : S. Banerjee and N. Agarwal

With the fast growth of the availability and quality of social and activity-rich resources such as blogs and other social media avenues, rising possibility and questioning arise as people now can, and do, actively use computational ability to find out and realize the opinions of others. The study of cooperative behavior of individuals has expressed to business intelligence , predictive calculus, customer relation management, and analyzing online joint action as manifested by various expressive mobs, In this article, we introduce a nature-inspired theory to model co operative behavior from the observed data on blog using group ability, where the goal is to correctly model and forecast for upcoming behavior of a large group of people after observing their connections during a training phase. Generally, an ant colony optimization model is trained with activity trend from the blog data and is tested over real-world blogs. [4]

5. Twitter Mood Predicts the Stock Market

Author: J. Bollen, H. Mao, and X. Zeng

Behavioral finances tell us that emotions can greatly affect individual behavior and decision making. can this also apply to societies at large, cooperative decision making? By extension is the public mood correlated or even analytical of economical indications? Here we examine if the measurements of collective mood states derived from large-scale Twitter feeds are connected to the value of the Dow Jones Industrial Average (DJIA) over time. We tally the resulting mood time series by comparing the resulting mood time series by comparing their skill to find the public's response to the presidential voting and Thanksgiving Day in 2008. A G ranger causality analysis and a Self-Organizing Fuzzy Neural circuit are then used to examine the premise that public mood states, as measured by the suggestion Finder and GPOMS mood time series, are cooperative of changes in DJIA closing values. Our results shows that the correct of DJIA predictions can be significantly increased.[5]

6. "Network Analysis in the Social Sciences"

Author: S. Borgatti, A. Mehra, D. Brass, and G. Labianca

Social scientists rapidly identify the prospective of social network examine, which enriches the clarification of human behavior by openly taking its social structure into account. In particular for the science of groups, public circuit analysis has reached a point of logical refinement that makes it a valuable tool for examining some of the central mechanisms that trigger intra- and intergroup behavior. The present article highlights the general significance of this scientific approach and describes the back-ground, generation, and application of cross-sectional as well as longitudinal network statistics that are of specific attention to group researchers. In doing so, we aim to offer a general preface for researchers new to this approach, while demonstrating the potential and limit of public network analysis for different areas in this area.[6]

7. "The Spread of Behavior in an Online Social Network Experiment"

Author: D. Centola

Social networks affect the distributed of activity .A popular proposal states that networks with many gregarious and a high degree of state will be less impressive for activity natural process in which locally excess that are rewired to give shortcuts cross ways the societal space. A competing proposal represent that when behaviors need social support, a network with more clustering may be more beneficial, even if the network as a whole has a big diameter. Analyze the personal property of network structure on natural process by studying the distributed of health conduct through artificially organized online communities. Single approval was much more likely when participants received social group support from multiple in the social network. The behavior distributed farther and faster cross ways clustered-lattice networks than across related to random networks system.[7]

8."Parallel Algorithms for Mining Large-Scale Rich-Media Data"

Author: E.Y. Chang, H. Bai, and K. Zhu

The sum of online photos and videos is now at tens of billions. To make, index, and recover these large-scale rich-media data, A system must employ ascendible data management and mining algorithms. The research communities necessarily to consider finding ample measure question instead of finding problems with small data sets that do not reflect real life script. This tutorial present key difficulties in large-scale rich-media data mining, and presents parallel algorithms for challenges. We instant our parallel implementations of Spectral Clustering (PSC), FP-Growth (PFP), Latent Dirichlet Allocation (PLDA), and Support Vector Machines (PSVM).[8]

9. "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data"

Author: R. Chen, K. Sivakumar, and H. Kargupta

We present a joint approach to studying a Bayesian network from distributed heterogeneous data. We first study a localized Bayesian network at all sites using local data. Then each site determines the observances that are mostly to be information of difference between local and non-local variables and convey a subset of these observances to a central site. Other Bayesian network is studied at the central site using the data inherited from the public site. The public and central Bayesian networks are clubbed to get a collective Bayesian network that models every data. Experimental outputs and theoretical consideration that shows the correctness of our approaches are shown.[9]

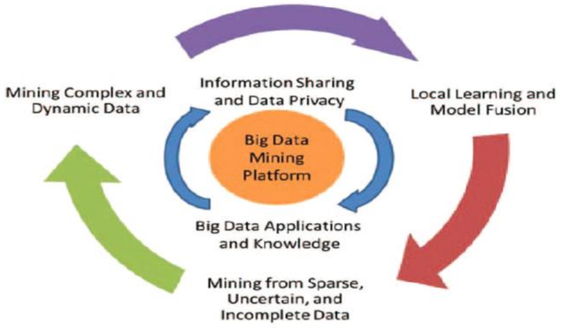
10. "Efficient Algorithms for Influence Maximization in Social Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

Author: Y.-C. Chen, W.-C. Peng, and S.-Y. Lee

In recent years, due to the flow in popularity of public-networking web sites, considerable involvement has shown because of maximum *usage of* social circuit. Given a public circuit structure, the problem of influence maximization is to find a minimum set of nodes that could maximize the distributed of influences. With a big-scale social network, the ratio and utility of such algorithms are depreciative. Although many recent studies have focused on the problem of influence maximization, these works are time-consuming when a social network is big-scale. In this paper, we propose two novel algorithms, CD H-Kcut and Community and power Heuristic on Kcut/SHRINK, to solve the influence maximization problem based on a graphic model. The community structure, which significantly reduces the number of candidates of authoritative nodes, to avoid knowledge intersection. The experimental results on both synthetic and real data sets indicate that our algorithms not only outmatch the state-of-the-art algorithms in efficiency but also possess graceful scalability.[10]

SYSTEM ARCHITECTURE :

Fig: Big data processing framework


Mining Complex and Dynamic Data
Information Sharing and Data Privacy
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

HACE THEORM

THE HACE stands for:

H: Heterogeneous

A: Autonomous

C: Complex

E: Evolving.

Heterogeneous: Data is a plan often used in the science and statistics relating to the quality in a substance .A photo that is homogeneous remaining the same in all cases and at all times in characters shape , size , height , weight, texture , distribution , disease , temperature , radioactivity , design , etc. one that is heterogeneous in a way that is readily distinguishable by the senses constant in one of these qualities.

Autonomous: Sources with distributed and decentralized authority main feature of Big Data.

Complex: Unstructured Data which is raw data yet to be processed.

Evolving: The day to day data is increasing with new type of data.

CONCLUSION

Because of Increase in the amount of data in the field of genomics, meteorology, biology, environmental research, it gets hard to take care of data, to find connections, patterns and to analyze the large data sets. As an organization rolls up much more data at this scale, validating the process of big data analysis will become paramount. The paper describes different methods of algorithms used to manage such large data sets and it gives an overview of architecture and algorithms used in large data sets.

ACKNOWLEDGEMENT

This research has been detailed in the structure of the project "Commercial Usage Usage of Big Data". Expanding our sincere and heartfelt thanks to our guide, Prof. Pavan Kulkarni, for providing us with the right guidance and for showing the right way, precious suggestions and guidance. Lastly we would like to thank our family& friends for the assist and confidence they have given us during the course of our work.

REFERENCES

1) "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Dec. 2012.Author:. R. Ahmed and G. Karypis.

2) "Novel Approaches to Crawling Important Pages Early" Author:M.H. Alam, J.W. Ha, and S.K. Lee.

3) "Identifying Influential and Susceptible Members of Social Networks" Author :S. Aral and D. Walker.

4) "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Author :S. Banerjee and N. Agarwal.

5) "Twitter Mood Predicts the Stock Market" Author :J. Bollen, H. Mao, and X. Zeng.

6) "Network Analysis in the Social Sciences" Author :S. Borgatti, A. Mehra, D. Brass, and G. Labianca.

7) "The Spread of Behavior in an Online Social Network Experiment" Author :D. Centola.

8) "Parallel Algorithms for Mining Large-Scale Rich-Media Data" Author :E.Y. Chang, H. Bai, and K. Zhu.

9) "Collective Mining of BayesianNetworks from Distributed Heterogeneous Data" Author :R. Chen, K. Sivakumar, and H. Kargupta.

10) "Efficient Algorithms for Influence Maximization in Social Networks," Author :Y.-C. Chen, W.-C. Peng, and S.-Y. Lee.