International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-7, Issue-4, April 2017

An Approach to Analysis and Classification of Data from Big Data by Using Apriori Algorithm

Mr. Nagesh Sharma, Mr. Ram Kumar Sharma

***Abstract*— A small amount of data can be easy to manage and straightforward analysis can be gleaned from it. After all, there's only so much data to consider. But when data gets big, big problems can arise. Most companies are swimming in more data than they know what to do with. Unfortunately, too many of them associate that drowning phenomenon with big data itself. Big data has rapidly developed into a hot topic that attracts extensive attention from academia, industry, and governments around the world. Data has become an indispensable part of every economy, industry, organization, business function and individual. Big Data is a term used to identify the datasets that whose size is beyond the ability of typical database software tools to store, manage and analyze. The Big Data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors. Enormous amount of data are generated every minute.**

***Index Terms*— Apriori Algoritthm, EMV, HIVE, Heterogeneity KDD, MangoDB, NoSQL, PIG, C4.5, SLIQ, Hadoop, Volume, Velocity, Variety**

I. INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data.

Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories.

The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules

Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making. The data may be enterprise specific or general and private or public. Big data are characterized by 3 V's: Volume, Velocity, and Variety.

Volume -the size of data now is larger than terabytes and peta bytes. The large scale and rise of size makes it difficult to store and analyze using traditional tools.

Velocity – big data should be used to mine large amount of data within a pre defined period of time. The traditional methods of mining may take huge time to mine such a volume of data.

Variety – Big data comes from a variety of sources which includes both structured and unstructured data. Traditional database systems were designed to address smaller volumes of structured and consistent data whereas Big Data is geospatial data, 3D data, audio and video, and unstructured text, including log files and social media. This heterogeneity of unstructured data creates problems for storage, mining and analyzing the data.

New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment.

II. ISSUES AND CHALLENGES

Big Data: Emerging Challenges of Big Data and Techniques for handling Big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. Big Data Data comes mainly in two forms- **1. Structured Data 2. Unstructured Data** (there are also **semi-structured data** – eg. XML) structured data has semantic meaning attached to it whereas unstructured data has no latent meaning. The growth in data that we are referring is most unstructured data. Below are few examples of unstructured data.

The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

2.1 Heterogeneity and Incompleteness

The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the

Mr. Nagesh Sharma, Assistant Professor, Noida Institute of Engineering and Technology Greater Noida (201306)
Mr. Ram Kumar Sharma, Assistant Professor, Noida Institute of Engineering and Technology Greater Noida (201306)

properties of the patterns vary greatly. Data can be both structured and unstructured. 70% of the data generated by organizations are unstructured.

2.2. Scale and complexity

Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, Computer Science & organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analysed.

2.3 Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. As the size of the data sets to be processed increases, it will take more time to analyse. In some situations results of the analysis is required immediately.

2.4 Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources.

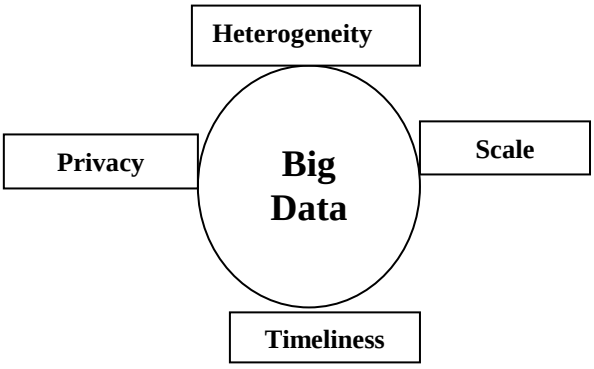

Fig 1 . Challenges and Issues of Big Data

III. III PROPOSED ALGORITHM FOR BIG DATA CLASSIFICATION AND ANALYSIS

There are many techniques available for data management. The Big Data handling techniques and tools include Hadoop, MapReduce, Simple DB, Google BigTable, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort. Out of these, Hadoop is one of the most widely used technologies.

i. Predictive analytics: software and/or hardware solutions that allow firms to discover, evaluate, optimize, and deploy predictive models by analyzing big data sources to improve business performance or mitigate risk.

ii. NoSQL databases: key-value, document, and graph databases.

iii. **Search and knowledge discovery**: tools and technologies to support self-service extraction of information and new insights from large repositories of unstructured and structured data that resides in multiple sources such as file systems, databases, streams, APIs, and other platforms and applications.

iv. **Stream analytics**: software that can filter, aggregate, enrich, and analyze a high throughput of data from multiple disparate live data sources and in any data format.

v. In-memory data fabric: provides low-latency access and processing of large quantities of data by distributing data across the dynamic random access memory (DRAM), Flash, or SSD of a distributed computer system.

vi. Distributed file stores: a computer network where data is stored on more than one node, often in a replicated fashion, for redundancy and performance.

vii. Data virtualization: a technology that delivers information from various data sources, including big data sources such as Hadoop and distributed data stores in real-time and near-real time.

viii. Data integration: tools for data orchestration across solutions such as Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couchbase, Hadoop, and MongoDB.

ix. Data preparation: software that eases the burden of sourcing, shaping, cleansing, and sharing diverse and messy data sets to accelerate data's usefulness for analytics.

x. Data quality: products that conduct data cleansing and enrichment on large, high-velocity data sets, using parallel operations on distributed data stores and databases.

So, Various Algorithms are used to analysis and classification of data for above parameter.

A. ID3 algorithm

The ID3 algorithm (**Quinlan**86) is a decision tree building algorithm which determines the classification of objects by testing the values of the their properties. It builds the tree in a top down fashion, starting from a set of objects and a specification of properties. At each node of the tree, a property is tested and the results used to partition the object set. This process is recursively done till the set in a given subtree is homogeneous with respect to the classification criteria - in other words it contains objects belonging to the same category. This then becomes a leaf node. At each node, the property to test is chosen based on information theoretic criteria that seek to maximize information gain and minimize entropy. In simpler terms, that property is tested which divides the candidate set in the most homogeneous subsets.

B. C4.5 algorithm

This algorithm was proposed by **Quinlan** (1993). The C4.5 algorithm generates a classification-decision tree for the given data-set by recursive partitioning of data. The decision is grown using **Depth-first** strategy. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the

entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes.

C. SLIQ algorithm

SLIQ (Supervised Learning In Quest) developed by IBM's Quest project team, is a decision tree classifier designed to classify large training data [1]. It uses a pre-sorting technique in the tree-growth phase. This helps avoid costly sorting at each node. SLIQ keeps a separate sorted list for each continuous attribute and a separate list called class list. An entry in the class list corresponds to a data item, and has a class label and name of the node it belongs in the decision tree. An entry in the sorted attribute list has an attribute value and the index of data item in the class list. SLIQ grows the decision tree in **breadth-first** manner. For each attribute, it scans the corresponding sorted list and calculates entropy values of each distinct values of all the nodes in the frontier of the decision tree simultaneously. After the entropy values have been calculated for each attribute, one attribute is chosen for a split for each nodes in the current frontier, and they are expanded to have a new frontier. Then one more scan of the sorted attribute list is performed to update the class list for the new nodes.

While SLIQ handles disk-resident data that are too large to fit in memory, it still requires some information to stay memory-resident which grows in direct proportion to the number of input records, putting a hard-limit on the size of training data. The Quest team has recently designed a new decision-tree-based classification algorithm, called SPRINT (**S**calable **Par**allelizable **In**duction of decision **T**rees) that for the removes all of the memory restrictions.

D. Apriori Algorithm

An association rule mining algorithm, Apriori has been developed for rule mining in large transaction databases. A *itemset* is a non-empty set of items.

They have decomposed the problem of mining association rules into two parts

- Find all combinations of items that have transaction support above minimum support. Call those combinations frequent itemsets.
- Use the frequent item sets to generate the desired rules. The general idea is that if, say, ABCD and AB are frequent item sets, then we can determine if the rule AB CD holds by computing the ratio r = support(ABCD)/support(AB). The rule holds only if r >= minimum confidence. Note that the rule will have minimum support because ABCD is frequent. The Apriori algorithm used in Quest for finding all frequent item sets is given below

Procedure AprioriAlg()
begin
L_1 :={frequent1-itemsets};
for (k := 2; $L_{k-1\ 0;\ k++}$) **do**
{ C_k= apriori-gen(L_{k-1}) ; // new candidates
for all transactions t in the dataset **do**
{ **for** all candidates c C_k contained int **do**
c:count++ }
L_k = { c C_k | c:count >= min-support}
}
Answer := $_k$ L_k
end

It makes multiple passes over the database. In the first pass, the algorithm simply counts item occurrences to determine the frequent 1-itemsets (itemsets with 1 item). A subsequent pass, say pass k, consists of two phases. First, the frequent itemsets L_{k-1} (the set of all frequent (k-1)-itemsets) found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the apriori-gen() function. This function first joins L_{k-1} with L_{k-1}, the joining condition being that the lexicographically ordered first k-2 items are the same. Next, it deletes all those itemsets from the join result that have some (k-1)-subset that is not in L_{k-1} yielding C_k.Frequent itemset generation after classification of the Apriori algorithm

k =1
Lk = {i|i ∈ I ∧ σ({i}) ≥ N × minsup}.
{Find all frequent 1-itemsets}
repeat k=k+1.
Ck = apriori-gen(Lk-1)).
{Generate candidate itemsets}
for each transection t ∈ T do
Ct = subset(Ck , t)
{Identify all candidates that belong to t}
for each candidate itemset c ∈ Ct do
σ(c) = σ(c)+1.
{Increment support count}
end for
end for Lk = {c|c ∈ Ck ∧ σ(c) ≥ N x minsup
{Extract the frequent kitemsets}
until Lk = ϕ
Result = ∪Lk

IV Comparison and Result Analysis

.There are currently 4 kinds of classical algorithm, frequently of use is the most widely used of these two algorithm, the following two algorithms will be compared With above given 10 parameters :-

Comparison Algorithm Name	Application	Processing	Aims
Apriori	Association rule	Frequent item set	Search the database to obtain candidate set of suppose item sets
C4.5	Machine Learning and data mining classification	Tuple	Supervised learning

Table1. The Comparison of Algorithms

The **Fig. 2** shows the all above parameter it would appear from this graph and comparison between to algorithm Apriori and C4.5 that are used in the Big Data Analytics and classification.

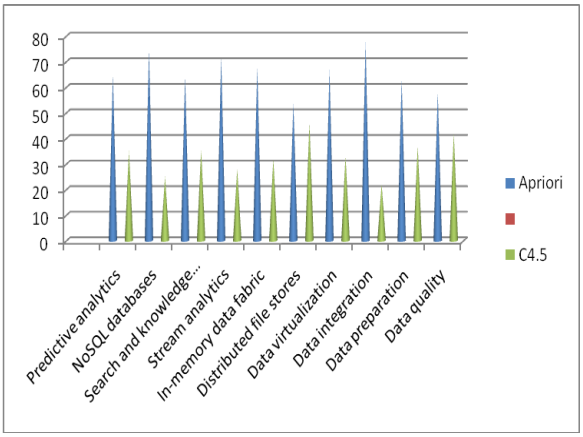

Fig 2 . Comparison Graph of Apriori and C4.5 Algorithm

V. CONCLUSION

These concepts include Big Data characteristics, challenges and techniques for handling big data and Big Data Mining. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. We are show the comparison of two different algorithm for classifying and analysis of the data with above ten parameters and find the Apriori algorithm is best on any data itemset.

VI. FUTURE SCOPE

As there are huge volumes of data that are produced every day, so such large size of data it becomes very challenging to achieve effective processing using the existing traditional techniques. Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big Data are presented. Apriori Algorithm is implemented with Map reduce for Classification of Structured, semi structured and unstructured data. We can also implement Apriori Algorithm for analysis of terabyte data with frequent itemset.

REFERENCE

[1].Jianwei Dai, Zhaolin Wu, Mingdong Zhu, Jianhua Gong, et al.Data engineering theory and technology.National defence industry press, 2010, p.180-220.

[2]. Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li," Big Data Processing in Cloud Computing Environments", International Symposium on Pervasive Systems, Algorithms and Networks,2012

[3].Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop"

[4].S.Vikram Phaneendra and E.Madhusudhan Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network Operations & Management Symposium (NOMS 2010),Osaka Japan, Apr 19-23 2013.

[5]. Ashish R. Jagdale, Kavita V. Sonawane & Shamsuddin S. Khan, Data Mining and Data Pre-processing for Big Data", Vol 5, Issue 7, July 2014.

[6]. M. Chen, S. Mao, and Y. Liu, "Big data: a survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[7].V.Jude Nirmal and D.I. George Amalarethinam, Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data", IJFMA Vol. 6, No. 2, 2015, 149-159, ISSN: 2320 –3242 (P), 2320 –3250 (online)Published on22 January 2015.

Nagesh Sharma, Department Name Information Technology, Noida Institute of Engineering & Technology Greater Noida MobileNo.-9999100436,(e mail:iamnageshsharma@gmail.com

Ram Kumar Sharma, Department Name - Information Technology, Noida Institute of Engineering & Technology Greater Noida MobileNo.-9654624322