International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-9,December 2015

ISSN: 2395-3470

www.ijseas.com

A Research on Big Data Analysis & Processing With Data Mining Techniques

Bharati Punjabi , Prof. Sonal Honale

Scholar - MTech. CSE,Abha Giakwad Patil College of Engineering,Nagpur.

bharati.ki.p@gmail.com

Prof. Sonal S. Honale – Deptt. Of Mtech CSE Abha Giakwad Patil College of Engineering,Nagpur.

sonalhonale@gmail.com

ABSTRACT

Today increasing number of organizations are facing the problem of explosion of data and the size of the databases used in today's enterprises has been growing at exponential rates. Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form. Today's business applications are having enterprise features like large scale, data-intensive, web-oriented and accessed from diverse devices including mobile devices. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data found in current scenario includes web logs, sensor networks, satellite and geo-spatial data, social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research, military. Surveillance, medical records, photography archives, video archives, and large-scale ecommerce.

Keywords— Big Data Problem, Hadoop cluster, Hadoop Distributed File System, Parallel Processing, MapReduce

INTRODUCTION

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Applications where data collection has grown tremendously and is beyond the capability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful

information or knowledge for future actions . In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

Data is being produced at an ever increasing rate. There has also been an acceleration in the proportion of machine-generated and unstructured data (photos , videos, social media feeds and so on) compared to structured data such that 80% or more of all data holdings are now unstructured and new approaches and technologies are required to access, link, manage and gain insight from these data sets.

The commonly accepted definition of big data comes from Gartner who define it as high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making, and process optimization. These are known as the “three Vs”. Some analysts also discuss big data in terms of value (the economic or political worth of data) and veracity (uncertainty introduced through data quality issues). Government agencies hold or have access to an ever increasing wealth of data including spatial and location data, as well as data collected from and by citizens. Experience suggests that such data can be utilised in ways that have the potential to transform service design and delivery so that personalised and streamlined services, that accurately and specifically meet individual’s needs, can be delivered to them in a timely manner.

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.

Improved service delivery could cover areas as diverse as remote medical diagnostics, major infrastructure management, personalised social security benefits delivery, improved first responder and emergency services, reduction of fraudulent or criminal activity across both government and private sectors, and the development of innovative new services as the growth and availability of Public Sector Information (PSI) becomes more prevalent.

The private sector holds huge amounts of data about its customers and in many cases leads the way in how this data is analysed and used to create new business models and services. Agencies have the opportunity to learn from the innovations occurring in the private sector to operate more efficiently and deliver services more effectively while ensuring that privacy and security matters are carefully considered.

Apache Hadoop - The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's Map Reduce and Google File System (GFS).

HDFS (Hadoop Distributed File System)- The Hadoop Distributed File System (HDFS) is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster.

HBASE- HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database

systems, HBase does not support SQL.In fact, HBase isn't a relational database at all. HBase applications are written In Java much like a typical MapReduce application.

Map Reduce - Map reduce is a software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers. Map Reduce is a programming model for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key.

"Map" step: The master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain: Map (k1, v1) → list (K2, v2)

"Reduce" step: The master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain: Reduce (K2, list (v2)) → list (v3)

RELATED WORK

Various research work in this area are made by the authors in the following national projects that all involve Big Data components:

Integrating and mining biodata from multiple sources in biological networks, sponsored by th US National Science Foundation, Medium Grant No. CCF-0905337, 1 October 2009 - 30 September 2013.

Issues and significance. We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

Big Data Fast Response. Real-time classification of Big Data Stream, sponsored by the Australian Research Council (ARC), Grant No. DP130102748, 1 January 2013 - 31 Dec. 2015.

Issues and significance. We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as accurately predict the trend of the data in the future; and - a knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications. .

Pattern matching and mining with wildcards and length constraints, sponsored by the National Natural Science Foundation of China, Grant Nos. 60828005 (Phase 1, 1 January 2009 - 31 December 2010) and 61229301 (Phase 2, 1 January 2013 - 31 December 2016).

Issues and significance. We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows: - exploration of the NP-hard complexity of the matching

and mining problems, - multiple pattern matching with wildcards, - approximate pattern matching and mining, and - application of our research onto ubiquitous personalized information processing and bioinformatics.

Key technologies for integration and mining of multiple, heterogeneous data sources, sponsored by the National High Technology Research and Development Program (863 Program) of China, Grant No. 2012AA011005, 1 January 2012 - 31 December 2014.

Issues and significance. We have performed an investigation on the availability and statistical regularities of multisource, massive and dynamic information, including cross-media search based on information extraction, sampling, uncertain information querying, and cross-domain and cross-platform information polymerization. To break through the limitations of traditional data mining methods, we have studied heterogeneous information discovery and mining in complex inline data, mining in data streams, multigranularity knowledge discovery from massive multisource data, distribution regularities of massive knowledge, quality fusion of massive knowledge. Group influence and interactions in social networks, sponsored by the National Basic Research 973 Program of China, Grant No. 2013CB329604, 1 January 2013 - 31 December 2017.

Issues and significance. We have studied group influence and interactions in social networks, including - employing group influence and information diffusion models, and deliberating group interaction rules in social networks using dynamic game theory, studying interactive individual selection and effect evaluations under social networks affected by group emotion, and analyzing emotional interactions and influence among individuals and groups, and - establishing an interactive influence model and its computing methods for social network groups, to reveal the interactive influence effects and evolution of social networks.

Why doing?

The simple idea that an organization should retain data that result from carrying out its mission and exploit those data to generate insights that benet the organization is of course not new. Commonly known as business intelligence, among other monikers, its origins date back several decades.

In this sense, the big data hype is simply a rebranding of what many organizations have been doing all along. Examined more closely, however, there are three major trends that distinguish insight-generation activities today from, say, the 1990s. First, we have seen a tremendous explosion in the sheer amount of data / orders of magnitude increase. In the past, enterprises have typically focused on gathering data that are obviously valuable, such as business objects representing customers, items in catalogs, purchases, contracts, etc. Today, in addition to such data, organizations also gather behavioral data from users. In the online setting, these include web pages that users visit, links that they click on, etc. The advent of social media and user generated content, and the resulting interest in encouraging such interactions, further contributes to the amount of data that is being accumulated.

Second, and more recently, we have seen increasing sophistication in the types of analyses that organizations perform on their vast data stores. Traditionally, most of the information needs fall under what is known as online analytical processing (OLAP). Common tasks include ETL (extract, transform, load) from multiple data sources, creating joined views, followed by altering, aggregation, or cube materialization.

Statisticians might use the phrase descriptive statistics to describe this type of analysis. These outputs might feed report generators, front-end dashboards, and other visualization tools to support common \roll up" and \drill down" operations on multi-dimensional data. Today, however, a new breed of \data

scientists" want to do far more: they are interested in predictive analytics. These include, for example, using machine learning techniques to train predictive models of user behavior|whether a piece of content is spam, whether two users should become \friends", the likelihood that a user will complete a purchase or be interested in a related product, etc. Other desired capabilities include mining large (often unstructured) data for statistical regularities, using a wide range of techniques from simple (e.g., k-means clustering) to complex (e.g., latent Dirichlet allocation or other Bayesian approaches). These techniques might surface \latent" facts about the users|such as their interest and expertise|that they do not explicitly express. To be fair, some types of predictive analytics have a long history|for example, credit card fraud detection and market basket analysis. However, we believe there are several qualitative diferences. The application of data mining on behavioral data changes the scale at which algorithms need to operate, and the generally weaker signals present in such data require more sophisticated algorithms to produce insights. Furthermore, expectations have grown|what were once cutting-edge techniques practiced only by a few innovative organizations are now routine, and perhaps even necessary for survival in today's competitive environment. Thus, capabilities that may have previously been considered luxuries are now essential.

THE BIG DATA MINING CYCLE

In production environments, effective big data mining at scale doesn't begin or end with what academics would consider data mining. Most of the research literature (e.g., KDD papers) focus on better algorithms, statistical models, or machine learning techniques |us ually starting with a (relatively) well-de ned problem, clear metrics for success, and existing data. The criteria for publication typically involve improvements in some figure of merit (hopefully statistically significant): the new proposed method is more accurate, runs faster, requires less memory, is more robust to noise, etc.

In contrast, the problems we grapple with on a daily basis are far more \messy". Let us illustrate with a realistic but hypothetical scenario. We typically begin with a poorly formulated problem, often driven from outside engineering SIGKDD Explorations Volume 14, Issue 2 Page 7 and aligned with strategic objectives of the organization, e.g., \we need to accelerate user growth". Data scientists are tasked with executing against the goal|and to operationalize the vague directive into a concrete, solvable problem requires exploratory data analysis. Consider the following sample questions:

When do users typically log in and out?

How frequently?

What features of the product do they use?

Do diferent groups of users behave diferently?

Do these activities correlate with engagement?

What network features correlate with activity?

How do activity roles of users change over time?

Before beginning exploratory data analysis, the data scientist needs to know what data are available and how they are organized. This fact may seem obvious, but is surprisingly difficult in practice.

DATA MINING AND BIG DATA

Big data and data mining are two different things. Both of them relate to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients. However, the two terms are used for two different elements of this kind of operation.

Big data is a term for a large data set. Big data sets are those that outgrow the simple kind of database and data handling architectures that were used in earlier times, when big data was more expensive and less feasible. For example, sets of data that are too large to be easily handled in a Microsoft Excel spreadsheet could be referred to as big data sets.

Data mining refers to the activity of going through big data sets to look for relevant or pertinent information. This type of activity is really a good example of the old axiom "looking for a needle in a haystack." The idea is that businesses collect massive sets of data that may be homogeneous or automatically collected. Decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business.

Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, or it can be largely labor-intensive, where individual workers send specific queries for information to an archive or database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific operating year.

In short, big data is the asset and data mining is the "handler" of that is used to provide beneficial results.

BIG DATA MINING ALGORITHMS

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods [11], designing a data mining mechanism from a multisource perspective [11], [12], as well as the study of dynamic data mining methods and the analysis of stream data [7], [12]. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data stream or a characteristic flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multisource data provide essential differences between single-source knowledge discovery and multisource data mining. Wu et al. [11], [12], [10] proposed and established the theory of local pattern analysis, which has laid a foundation for global

knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find. Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing.

Data streams are widely used in financial analysis, online trading, medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining [7]. Knowledge evolution is a common phenomenon in real world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by dynamics and context in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features.

CONCLUSION

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors.

We have examined the design and architecture of Hadoop's MapReduce framework. For Bigdata Processing Hadoop Mapreduce is the tool available that can be used for processing data and its distributed, column-oriented database, HBase which uses HDFS for its underlying storage, and support provides more efficiency to the system.

A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

REFERENCES

[1] 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.

[2] 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.

[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," Proc. 17th ACM Int'l Conf. Multimedia, (MM '09,) pp. 917-918, 2009.

[12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.