Brahmini Togiti ,INDIA / International Journal of Research and Computational Technology, Vol.7 Issue.2
ISSN: 0975-5662, April, 2015 www.ijrct.org

Analysis of Mining on Big Data

B.Srinivas[1], Asst. Professor, CSE Department ,Email: bsvsjbm@gmail.com
Brahmini Togiti[2], M.Tech, CSE Department,E mail: tbrahmini@gmail.com
Christu Jyothi Institute of Science and Technology, Warangal, T.S , INDIA

Abstract— Large amounts of data, a variety of sources, high speed production, but also high speed processing - these are the basic characteristics of Big Data. The amount of data that is generated and collected in each second grows exponentially. The management of Big Data, the intelligent use of large, heterogeneous data sets, is becoming increasingly important for competition. It is affecting all sectors - industry and academia but also the public sector. The data can be from health care and scientific sensors, user-generated data, Internet and financial companies, and supply chain systems. Big data is growing in terms of Volume, Velocity and Variety, also known as 3Vs which include data sets with sizes beyond the ability of commonly-used software tools to capture, manage, and process the data within a tolerable elapsed time". Volume is sizes of the data sets consumed by today's Web applications can be extraordinarily. Velocity can be speed with which new data is created or existing data is updated. Handling data from various sources which can be of different formats and models and capture the different types of data to correlate their meanings is Variety.

Index Terms— Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations, Volume, Variety, Velocity.

1. INTRODUCTION

1.1 Big Data:

Big data is used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

1.2 Data Mining:

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.3 Difference between Big data and Data mining:

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.

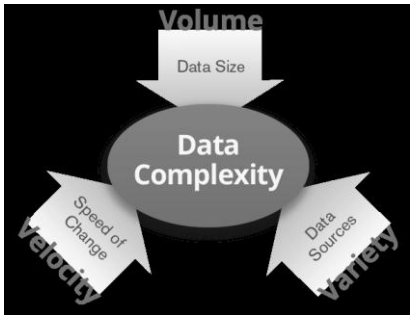
2. LITERATURE SURVEY

2.1 Big Data Characteristics

Volume (Scale of data): The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety (Different forms of data): The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data.

Velocity (Analysis of streaming data): The term 'velocity' in this context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.



Veracity (Uncertainty of data): This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Complexity: Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

2.2 HACE Theorem

The basic features of big data are it is huge in size. The data keep on changing time to time. Its data sources are from different phases. It is free from the influence, guidance, or control of anyone. It is too much complex in nature, thus hard to handle.

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data

1. **Huge Data with Heterogeneous and Diverse Dimensionality**

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, each single human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on

2. **Autonomous Sources with Distributed and Decentralized Control**

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data-related applications, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/ regions.

3. **Complex and Evolving Relationships**

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background, and so on, to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections, which is one of the most important factors of the human society.

3. PROBLEM DEFINITION

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. As another example, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, there is an old saying states: “a picture is worth a thousand words,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data. Here we are concerned about time taken for processing the huge amount of diverse data, analyzation techniques, and displaying useful data.

3.1 Big Data Challenges:

Big data presents a number of challenges, the challenges in Big Data are usually the real implementation hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the

technology implementation and some unpleasant results. There are numerous challenges, from privacy and security to access and deployment such as

3.1.1 Collecting data privately:

A tremendous amount of data about individuals – e.g., demographic information, internet activity, energy usage, communication patterns and social interactions – are being collected and analyzed by many national statistical agencies, survey organizations, medical centers, and Web and social networking companies. Wide dissemination of microdata (data at the granularity of individuals) facilitates advances in science and public policy, helps citizens to learn about their societies, and enables students to develop skills at data analysis. Often, however, data producers cannot release microdata as collected, because doing so could reveal data subjects' identities or values of sensitive attributes. Failing to protect confidentiality (when promised) is unethical and can cause harm to data subjects and the data provider.

3.1.2 Access control in web and social networking applications:

Ensuring privacy of linked data, e.g. social networks, where people are linked to other people, and relational data, where different types of entities maybe linked to one another. Reasoning about privacy in such data is tricky since information about an individual may be leaked through links to other individuals. Another interesting problem is that of releasing sequential releases of the same data over time. Attackers may link individuals across releases and infer additional sensitive information that they could not have from a single release. Finally, as the data we deal with become extremely high dimensional, we need to develop techniques that can protect privacy while guaranteeing utility. Understanding theoretical trade-offs between privacy and utility is an important open area for research.

3.1.3 Displaying meaningful results:

Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information.

3.1.4 Other Challenges:

The main sectors at which the challenges for Big Data arrive are Mining Platform, Privacy, Design of mining algorithms. Mining Platform: includes Big data reduction, Big data integration and cleaning, Big data indexing and query, Big data analysis and mining. Big privacy: concerns can arrive from Systems for collecting data privately, Access control in web and social networking applications, Data security and cryptography, Protocols for secure computation**.**

4. SOLUTION APPROACH AND METHODOLOGY

The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously and is beyond the ability 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. For example, the square kilometer array (SKA) in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented

data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

Distributed and parallel data management solutions such as OceanBase have been tried to address the scalability problem for online processing large scale relational data. In-memory databases such as VoltDB and HANA have been recently commercialized to response the performance challenges for OLTP and OLAP applications of big data. For big graph data, many solutions on efficient processing large graphs (e.g., Neo4j and Pregel) that cannot be held in memory have been proposed recently. Pregel is a famous one of them.

4.1 Big Data Mining Techniques

1. Big data mining platform
2. Big data semantics and application knowledge
3. Big data mining algorithms
4. Extract useful and meaningful data

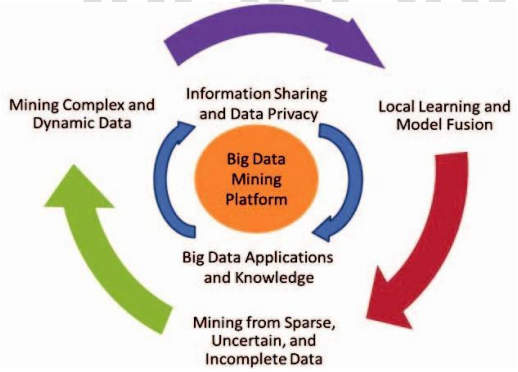


Fig.2. A Big Data processing framework: The research challenges form a three tier structure and center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms

4.1.1 Tier I: Mining Platform for Big Data

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals. Indeed, many data mining algorithm are designed for this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing or collective mining to sample and aggregate data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as Map Reduce or Enterprise Control Language (ECL), on a large number of computing nodes (i.e., clusters). The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes. For example, as of this writing, the world most powerful super computer Titan, which is deployed at Oak Ridge National Laboratory in Tennessee, contains 18,688 nodes each with a 16-core CPU.

Big Data mining offers opportunities to go beyond traditional relational databases to rely on less structured data: weblogs, social media, e-mail,

sensors, and photographs that can be mined for useful information. Major business intelligence companies, such IBM, Oracle, Teradata, and so on, have all featured their own products to help customers acquire and organize these diverse data sources and coordinate with customers' existing data to find new insights and capitalize on hidden relationships.

4.1.2 Tier II: Big Data Semantics and Application Knowledge

Semantics and application knowledge in Big Data refer to numerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include

- Data sharing and privacy; and
- Domain and application knowledge.

The former provides answers to resolve concerns on how data are maintained, accessed, and shared; whereas the latter focuses on answering questions like "what are the underlying applications?" and "what are the knowledge or patterns users intend to discover from the data ?"
Information sharing is an ultimate goal for all systems involving multiple parties. While the motivation for sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not resolve privacy concerns. For example, knowing people's locations and their preferences, one can enable a variety of useful location-based services, but public disclosure of an individual's locations/movements over time can have serious consequences for privacy. To protect privacy, two common approaches are to

- Restrict access to the data, such as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and
- Anonymize data fields such that sensitive information cannot be pinpointed to an individual record .

For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconducted by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indistinguishable from k -1 others. Common anonymization approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data.

One of the major benefits of the data annomization-based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. This naturally leads to another research area namely privacy preserving data mining , where multiple parties, each holding some sensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the data. This privacy preserving mining goal, in practice, can be solved through two types of approaches including

- Using special communication protocols, such as Yao's protocol, to request the distributions of the whole data set, rather than requesting the actual values of each record, or
- Designing special data mining methods to derive knowledge from anonymized data (this is inherently similar to the uncertain data mining methods).

4.1.3 Tier III: Big Data Mining Algorithms

1. Local Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to

the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models. Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthetized by aggregating patterns across all sites. At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

2. Mining from Sparse, Uncertain, and Incomplete Data

Spare, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues,where data in a high-dimensional space (such as more than 1,000 dimensions) do not show clear trends or distributions. For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection to reduce the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For example, data produced from GPS equipment are inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recording location is represented by a mean value plus a variance to indicate expected errors. For data privacy related applications, users may intentionally inject randomness/errors into the data to remain anonymous.

This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k]. For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. For example, error aware data mining utilizes the mean and the variance values with respect to each single data item to build a Naive Bayes model for classification. Similar approaches have also been applied for decision trees or database queries. Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission).

While most modern data mining algorithms have in-built solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data). Many imputation methods exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance.

3. Mining Complex and Dynamic Data

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving. For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly real-time speed and very high accuracy. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node network may be subject to one trillion connections. For a large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily user actions/interactions into consideration, the scale of difficulty will be even more astonishing.

Inspired by the above challenges, many data mining methods have been developed to find interesting knowledge from Big Data with complex relationships and dynamically changing volumes. For example, finding communities and tracing their dynamically evolving relationships are essential for understanding and managing complex systems. Discovering outliers in a social network is the first step to identify spammers and provide safe networking environments to our society.

If only facing with huge amounts of structured data, users can solve the problem simply by purchasing more storage or improving storage efficiency. However, Big Data complexity is represented in many aspects, including complex heterogeneous data types, complex intrinsic semantic associations in data, and complex relationship networks among data. That is to say, the value of Big Data is in its complexity.

Complex heterogeneous data types. In Big Data, data types include structured data, unstructured data, and semistructured data, and so on. Specifically, there are tabular data (relational databases), text, hyper-text, image, audio and video data, and so on. The existing data models include key-value stores, bigtable clones, document databases, and graph databases, which are listed in an ascending order of the complexity of these data models. Traditional data models are incapable of handling complex data in the context of Big Data. Currently, there is no acknowledged effective and efficient data model to handle Big Data.

Complex intrinsic semantic associations in data. News on the web, comments on Twitter, pictures on Flicker, and clips of video on YouTube may discuss about an academic awardwinning event at the same time. There is no doubt that there are strong semantic associations in these data. Mining complex semantic associations from “text-image-video” data will significantly help improve application system performance such as search engines or recommendation systems. However, in the context of Big Data, it is a great challenge to

efficiently describe semantic features and to build semantic association models to bridge the semantic gap of various heterogeneous data sources.

Complex relationship networks in data. In the context of Big Data, there exist relationships between individuals. On the Internet, individuals are webpages and the pages linking to each ther via hyperlinks form a complex network. There also exist social relationships between individuals forming complex social networks, such as big relationship data from Facebook, Twitter, LinkedIn, and other social media including call detail records (CDR), devices and sensors information GPS and geocoded map data, massive image files transferred by the Manage File Transfer protocol, web text and click-stream data, scientific information, e-mail, and so on. To deal with complex relationship networks, emerging research efforts have begun to address the issues of structure-and-evolution, crowds-and-interaction, and information-and-communication.

The emergence of Big Data has also spawned new computer architectures for real-time data-intensive processing, such as the open source Apache Hadoop project that runs on high-performance clusters. The size or complexity of the Big Data, including transaction and interaction data sets, exceeds a regular technical capability in capturing, managing, and processing these data within reasonable cost and time limits. In the context of Big Data, real-time processing for complex data is a very challenging task.

Advantages

1. Fast response
2. Extract useful information
3. Prediction of required data from large amount of data.

5. CONCLUSION

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values.

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.

In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. 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. 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.

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data

technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at realtime. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

6. REFERENCES

[1] Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.

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

[3] 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

[4] Xindong Wu,Xingquan Zhu,Gong-Qing Wu, and Wei Ding "Data Mining with Big Data" Knowledge and Data Engineering vol. 26, no. 1, January 2014