GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data

J.Josepha Menandas
Assistant Professor(Grade-I),
Panimalar Engineering College,
Chennai, India

J.Jakkulin Joshi
Assistant Professor,
Panimalar Engineering College,
Chennai, India

ABSTRACT

Big data concern extremely large data sets with multiple that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions including physical, biological and biomedical sciences. Big data describes the exponential growth and availability of data both structured and unstructured, includes large-volume, complex, autonomous sources. With the fast development of networking, data storage and the capacity of data collection, Big data are now rapidly expanding in all science and engineering domain. This paper presents a PARALLEL PROCESSING TECHNIQUE (PPT) that characterizes the features of big data revolution, reduces complexity, and proposes a big data processing model from the data mining perspective. This model involves data accessing and computing, data privacy and domain knowledge, and big data mining algorithms, and also the big data revolution.

Keywords

Big data, complexity, data mining, heterogeneity, autonomous sources, volume, velocity, variety, variability.

1. INTRODUCTION

In recent years, the amount of data in our world has been increasing explosively, and analyzing large data sets-so called “Big Data”- becomes a key basis of competition underpinning new waves of productivity growth, innovation and consumer surplus[1]. Then, what is “Big Data”?,Big data refers to data sets whose size is beyond the ability of current technology, method and theory to capture, manage, and process the data within the tolerable elapsed time.

As far back as 2001,a mainstream definition of big data spans three dimensions : Volume, Velocity, and Variety, in addition to Variability and Complexity. **Volume:** Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data. **Velocity:** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations. **Variety:** Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with. **Variability:** In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved. **Complexity:** Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. For example, the square kilometer array (SKA) [17] 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 [41] 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.

The remainder of the paper is structured as follows: In Section 2, we given a Parallel Processing Technique to model and characterize Big Data. Section 3 summarizes the key challenges for intelligent learning with Big Data. Related work is discussed in Section 4, and we conclude the paper in Section 5.

2. BIG DATA CHARACTERIZATION

The characteristics of Big Data are heterogeneous, autonomous sources with distributed and decentralized control, and seek to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Exploring the Big Data is equivalent to aggregating heterogeneous information from different sources to create a best possible in real time fashion.

2.1 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.

2.2 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.

2.3 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. For reducing complexity, most frequently accessed data should be kept in a separate servers, so that to make active fully. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process on the data. In a dynamic world, the features used to represent the indivi-duals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the complex (nonlinear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

3. INTELLIGENT LEARNING WITH BIG DATA

For an intelligent learning database system [52] to handle Big Data, the essential key is to scale up to the exceptionally large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem[58]. Fig. 1 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (phase I), data privacy and domain knowledge (phase II), and Big Data mining algorithms (phase III).

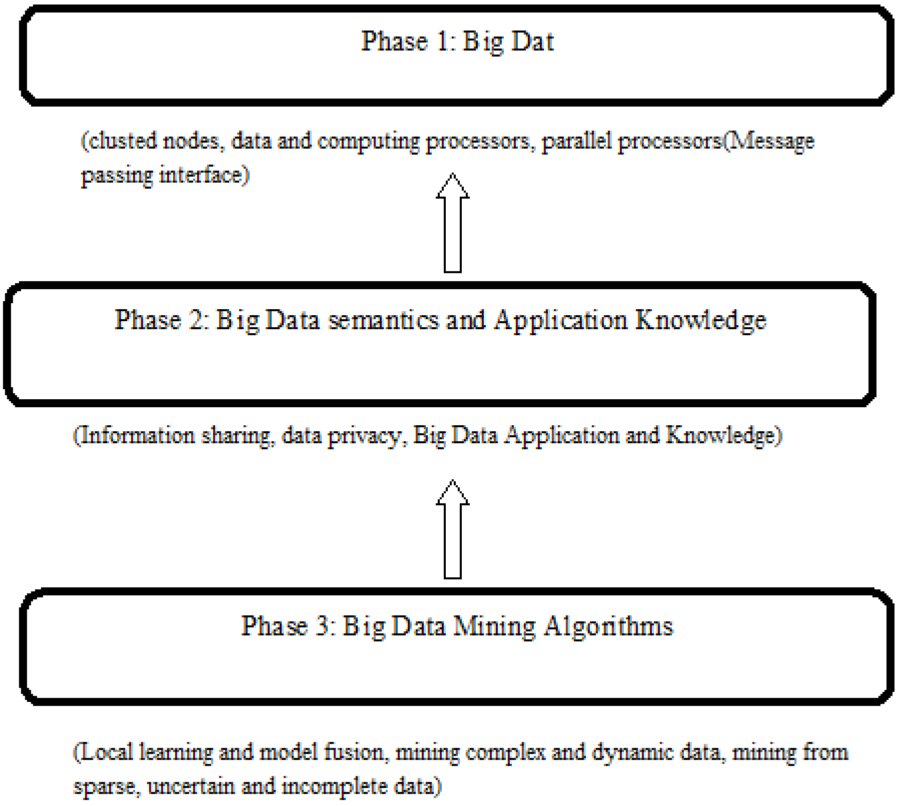
Phase 1: Big Dat
(clusted nodes, data and computing processors, parallel processors(Message passing interface)
Phase 2: Big Data semantics and Application Knowledge
(Information sharing, data privacy, Big Data Application and Knowledge)
Phase 3: Big Data Mining Algorithms
(Local learning and model fusion, mining complex and dynamic data, mining from sparse, uncertain and incomplete data)

Fig 1:Big Data processing Framework

3.1 Big data mining platform (Phase I)

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 [43], [33] or collective mining [12] 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 MapReduce 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.

Such a Big Data system, which blends both hardware and software components, is hardly available without key industrial stockholders' support. In fact, for decades, companies have been making business decisions based on transactional data stored in relational databases. 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.

3.2 Semantics of Application Knowledge in Big data (Phase II)

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 1) data sharing and privacy; and 2) 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 under-lying applications ?" and "what are the knowledge or patterns users intend to discover from the data ?"

3.2.1 Information Sharing and Data Privacy

Information sharing is an ultimate goal for all systems involving multiple parties [24]. While the motivation for applications are related to sensitive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not resolve privacy concerns [19], [25], [42]. 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 1) 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 2) anonymize data fields such that sensitive information cannot be pinpointed to an individual record [15]. 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 [30], 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 1) using special communication protocols, such as Yao's protocol [54], to request the distributions of the whole data set, rather than requesting the actual values of each record, or 2) designing special data mining methods to derive knowledge from anonymized data (this is inherently similar to the uncertain data mining methods).

3.2.2 Domain and Application Knowledge

Domain and application knowledge [28] provides essential information for designing Big Data mining algorithms and systems. In a simple case, domain knowledge can help identify right features for modeling the underlying data (e.g., blood glucose level is clearly a better feature than body mass in diagnosing Type II diabetes). The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques. For example, stock market data are a typical domain that constantly generates a large quantity of information, such as bids, buys, and puts, in every single second. The market continuously evolves and is impacted by different factors, such as domestic and international news, government reports, and natural disasters, and so on. An appealing Big Data mining task is to design a Big Data mining system to predict the movement of the market in the next one or two minutes. Such systems, even if the prediction accuracy is just slightly better than random guess, will bring significant business values to the developers [9]. Without correct domain knowledge, it is a clear challenge to find effective matrices/measures to characterize the market movement, and such knowledge is often beyond the mind of the data miners, although some recent research has shown that using social networks, such as Twitter, it is possible to predict the stock market upward/downward trends [7] with good accuracies.

3.3 Algorithm for mining Complex and dynamic data (Phase III)

3.3.1 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, just like the elephant and blind men case. 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 [50]. 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.

3.3.2 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 [48] 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 [36], 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 [49] utilizes the mean and the variance values with respect to each single data item to build a Naıve 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 [20] 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.

3.3.3 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 [6]. Documents posted on WWW servers, Internet back-bones, 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 [23]. 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 [3], [10]. Discovering outliers in a social network [8] 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, big table 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 award-winning 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 web pages and the pages linking to each other 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 [5], [13], [56], including call detail records (CDR), devices and sensors information [1], [44], GPS and geocoded map data, massive image files transferred by the Manage File Transfer protocol, web text and click-stream data [2], scientific information, e-mail [31], 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.

4. RELATED WORK

4.1 Big Data mining platform

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristic of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing the parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality".

Data mining by testing series of standard data mining tasks on medium size clusters. Papadimitriou and Sun [38] proposed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduce project showed that DisCo has perfect scalability and can process and analyze massive data sets (with hundreds of GB).

To improve the weak scalability of traditional analysis software and poor analysis capabilities of Hadoop systems, Das et al. [16] conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis capabilities for Hadoop. Wegener et al. [47] achieved the integration of Weka (an open-source machine learning and data mining software tool) and MapReduce. Standard Weka tools can only run on a single machine, with a limitation of 1-GB memory. After algorithm parallelization, Weka breaks through the limitations and improves performance by taking the advantage of parallel computing to handle more than 100-GB data on MapReduce clusters. Ghoting et al. [21] proposed HadoopML, on which developers can easily build task-parallel or data parallel machine learning and data mining algorithms on program blocks under the language runtime environment.

4.2 Big Data semantics and application knowledge

In privacy protection of massive data, Ye et al. [55] proposed a multilayer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization/refinement order for classification. A recent paper on confidentiality protection in Big Data [4] summarizes a number of methods for protecting public release data, including aggregation (such as k-anonymity, I-diversity, etc.), suppression (i.e., deleting sensitive values), data swapping (i.e., switching values of sensitive data records to prevent users from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions.

For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data. To carry out Big Data mining, having an efficient and effective data access mechanism is vital, especially for users who intend to hire a third party (such as data miners or data auditors) to process their data. Under such a circumstance, users' privacy restrictions may include 1) no local data copies or downloading, 2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many others. In Wang et al. [48], a privacy-preserving public auditing mechanism for large scale data storage (such as cloud computing systems) has been proposed. The public key-based mechanism is used to enable third-party auditing (TPA), so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy.

For most Big Data applications, privacy concerns focus on excluding the third party (such as data miners) from directly accessing the original data. Common solutions are to rely on some privacy-preserving approaches or encryption mechanisms to protect the data. A recent effort by Lorch et al. [32] indicates that users' "data access patterns" can also have severe data privacy issues and lead to disclosures of geographically co-located users or users with common interests (e.g., two users searching for the same map locations are likely to be geographically colocated). In their system, namely Shround, users' data access patterns from the servers are hidden by using virtual disks. As a result, it can support a variety of Big Data applications, such as microblog search and social network queries, without compromising the user privacy.

4.3 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 [50], [51], as well as the study of dynamic data mining methods and the analysis of stream data [18], [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. [50], [51], [45] 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 [18], [57].

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 of cardiovascular and other chronic epidemiological changes with the passage of time. 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 [53].

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 [27].

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 real-time. 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. REFERNCES

[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. Multi-media, (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.

[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, “Efficient Algorithms for Influence Maximization in Social Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, “Map-Reduce for Machine Learning on Multicore,”

Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS ’06), pp. 281-288, 2006.

[15] G. Cormode and D. Srivastava, “Anonymized Data: Generation, Models, Usage,” Proc. ACM SIGMOD Int’l Conf. Management Data,

[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, “Ricardo: Integrating R and Hadoop,” Proc. ACM SIGMOD Int’l Conf. Management Data (SIGMOD ’10), pp. 987-998. 2010.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, “The Square Kilometre Array,” Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams,"

Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[19] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

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

[21] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.

[22] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[23] M. Helft, "Google Uses Searches to Track Flu's Spread," The New York Times, http://www.nytimes.com/2008/11/12/technology/ internet/12flu.html. 2008.

[24] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.

[25] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.

[26] "IBM What Is Big Data: Bring Big Data to the Enterprise," http:// www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[27] A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44, 2009.

[28] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence,

I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002.

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

[30] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[31] W. Liu and T. Wang, "Online Active Multi-Field Learning for Efficient Email Spam Filtering," Knowledge and Information Systems, vol. 33, no. 1, pp. 117-136, Oct. 2012.

[32] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13), 2013.

[33] D. Luo, C. Ding, and H. Huang, "Parallelization with Multi-plicative Algorithms for Big Data Mining," Proc. IEEE 12th Int'l Conf. Data Mining, pp. 489-498, 2012.

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

[35] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/ 6855169886/, 2012.

[36] T. Mitchell, "Mining our Reality," Science, vol. 326, pp. 1644-1645, 2009.

[37] Nature Editorial, "Community Cleverness Required," Nature, vol. 455, no. 7209, p. 1, Sept. 2008.

[38] S. Papadimitriou and J. Sun, "Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to-End Mining," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 512-521, 2008.

[39] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems," Proc. IEEE 13th Int'l Symp. High Perfor-mance Computer Architecture (HPCA '07), pp. 13-24, 2007.

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

[41] C. Reed, D. Thompson, W. Majid, and K. Wagstaff, "Real Time Machine Learning to Find Fast Transient Radio Anomalies: A Semi-Supervised Approach Combining Detection and RFI Excision," Proc. Int'l Astronomical Union Symp. Time Domain Astronomy,

Sept. 2011.

[42] E. Schadt, "The Changing Privacy Landscape in the Era of Big Data," Molecular Systems, vol. 8, article 612, 2012.

[43] J. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," Proc. 22nd VLDB Conf., 1996.

[44] A. da Silva, R. Chiky, and G. He´brail, "A Clustering Approach for Sampling Data Streams in Sensor Networks," Knowledge and Information Systems, vol. 32, no. 1, pp. 1-23, July 2012.

[45] K. Su, H. Huang, X. Wu, and S. Zhang, "A Logical Framework for Identifying Quality Knowledge from Different Data Sources," Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[46] "Twitter Blog, Dispatch from the Denver Debate," http:// blog.twitter.com/2012/10/dispatch-from-denver-debate.html, Oct. 2012.

[47] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.

[48] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[49] X. Wu and X. Zhu, "Mining with Noise Knowledge: Error-Aware Data Mining," IEEE Trans. Systems, Man and Cybernetics, Part A, vol. 38, no. 4, pp. 917-932, July 2008.

[50] X. Wu and S. Zhang, "Synthesizing High-Frequency Rules from Different Data Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[51] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems, vol. 30, no. 1, pp. 71-88, 2005.

[52] X. Wu, "Building Intelligent Learning Database Systems," AI Magazine, vol. 21, no. 3, pp. 61-67, 2000.

[53] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online Feature Selection with Streaming Features," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 5, pp. 1178-1192, May 2013.

[54] A. Yao, "How to Generate and Exchange Secretes," Proc. 27th Ann. Symp. Foundations Computer Science (FOCS) Conf., pp. 162-167, 1986.

[55] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing Classification Data Using Rough Set Theory," Knowledge-Based Systems, vol. 43, pp. 82-94, 2013.

[56] J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, "Information Propagation in Online Social Networks: A Tie-Strength Perspective," Knowledge and Information Systems, vol. 32, no. 3, pp. 589-608, Sept. 2012.

[57] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active Learning From Stream Data Using Optimal Weight Classifier Ensemble," IEEE Trans. Systems, Man, and Cybernetics, Part B, vol. 40, no. 6, pp. 1607-1621, Dec. 2010.

[58] X. Wu, X. Zhu, G-Q. Wu, and W. Ding, "Data Mining with Big Data," IEEE Trans. On Knowledge and data engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.