A Local Pattern Analysis using Data Mining with Big Data

A. Shanmuga Velayutham
Senior Asst.Professor Information Technology
M.Kumarasamy College Of Engineering
Karur,India

P. Dhivya
Information Technology M.Kumarasamy College Of
Engineering Karur,India

L. Jeyavani
Information Technology M.Kumarasamy College Of
Engineering Karur,India

R. Lalitha
Information Technology M.Kumarasamy College Of
Engineering Karur,India

A. Musrath
Information Technology M.Kumarasamy College Of
Engineering
Karur,India

***Abstract*—Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.**

***Keywords*—big data; HACE theorem; security; privacy**

1 Introduction

DR. Yan Mo won the 2012 Nobel Prize in Literature. This is probably the most controversial Nobel prize of this category. Searching on Google with “Yan Mo Nobel Prize,” resulted in 1,050,000 web pointers on the Internet (as of 3 January 2013). “For all praises as well as criticisms,” said Mo recently, “I am grateful.” What types of praises and criticisms has Mo actually received over his 31-year writing career? As comments keep coming on the Internet and in various news media, can we summarize all types of opinions in different media in a real-time fashion, including updated, cross-referenced discussions by critics? This type of summarization program is an excellent example for Big

Along with the above example, the era of Big Data has arrived [37], [34], [29]. 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 [26]. 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 [46]. 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, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 [35]. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as 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. The above examples demonstrate the rise of Big Data applications where data collection has grown tremen- dously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time."

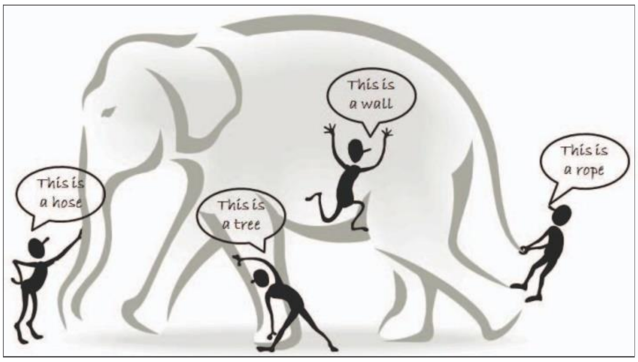

Fig. 1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

Big Data mining. Some key research initiatives and the authors' national research projects in this field are outlined in Section 4. Related work is discussed in Section 5, and we conclude the paper in Section 6.

2 Big Data Characteristics: HACE Theorem

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.

These characteristics make it an extreme challenge for

discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Fig. 1), which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the elephant according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that 1) the elephant is growing rapidly and its pose changes constantly, and 2) each blind man may have his own (possible unreliable and inaccu- rate) information sources that tell him about biased knowledge about the elephant (e.g., one blind man may exchange his feeling about the elephant with another blind man, where the exchanged knowledge is inherently biased). Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the elephant in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

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 informa- tion 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. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic-related test, micro- array expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

2.2 Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentra- lized 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 applica- tions, 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. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. More specifically, the local government regula- tions also impact on the wholesale management process and result in restructured data representations and data warehouses for local markets.

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. 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. Our friend circles may be formed based on the common hobbies or people are connected by biological relationships. Such social connections commonly exist not only in our daily activities, but also are very popular in cyberworlds. 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 the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship repre- sentation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. 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.

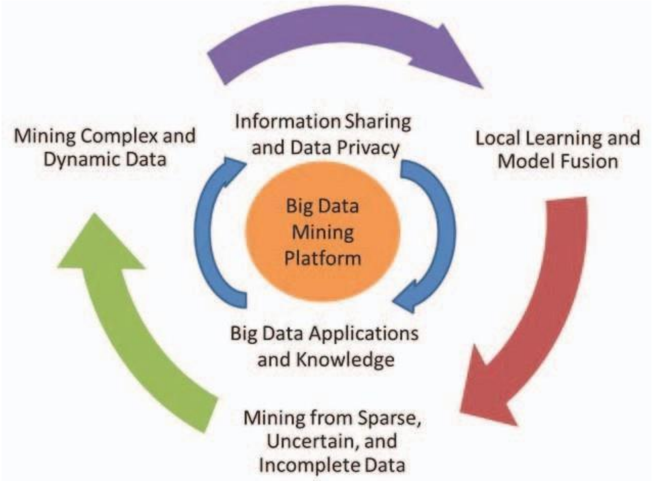
Information Sharing and Data Privacy
Mining Complex and Dynamic Data
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

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.

3 DATA MINING CHALLENGES 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. Fig. 2 shows a conceptual view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, typical data mining algorithms require all data to be loaded into the main memory, this, however, is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly differ- ent. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users' location information is clearly not acceptable for majority, if not all, applications. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. For example, in market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling, and so on. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high-level mining algorithm designs.

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. The circle at Tier III contains three stages. First, sparse, heterogeneous, uncertain, incomplete, and multisource data are prepro- cessed by data fusion techniques. Second, complex and dynamic data are mined after preprocessing. Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is fedback to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

In the following, we elaborate challenges with respect to the three tier framework in Fig. 2.

3.1 Tier I: Big Data Mining Platform

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 paralle l programming tools, such as MapR educ e 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 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 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 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 con- cerns [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 indivi- dual record [15]. For the first approach, common chal- lenges 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 anonymiza- tion 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 Tier III: Big Data Mining Algorithms

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 system- atically 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 gener- ated 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 distribu- tions. For most machine learning and data mining algorithms, high-dimensional spare data significantly de- teriorate 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 Naïve 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 estab- lished 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 diffi- culty will be even more astonishing.

Inspired by the above challenges, many data mining methods have been developed to find interesting knowl- edge from Big Data with complex relationships and dynamically changing volumes. For example, finding communities and tracing their dynamically evolving rela- tionships 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 semistruc- tured 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 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 webpages 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-interac- tion, 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, mana- ging, 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 RESEARCH INITIATIVES AND PROJECTS

To tackle the Big Data challenges and "seize the opportunities afforded by the new, data driven resolu- tion," the US National Science Foundation (NSF), under President Obama Administration's Big Data initiative, announced the BIGDATA solicitation in 2012. Such a federal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management (led by the University of Washington), analytical approaches for genomics-based massive data computation (led by Brown University), large scale machine learning techniques for high-dimensional data sets that may be as large as 500,000 dimensions (led by Carnegie Mellon University), social analytics for large- scale scientific literatures (led by Rutgers University), and several others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures that allow us to bring the massive amounts of data down to a human manageable and interpretable scale. Other coun- tries such as the National Natural Science Foundation of China (NSFC) are also catching up with national grants on Big Data research.

Meanwhile, since 2009, the authors have taken the lead in the following national projects that all involve Big Data components:

. Integrating and mining biodata from multiple sources in biological networks, sponsored by the 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 integra- tion, 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 matchin g and minin g 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 bioin- formatics.

. Key technologies for integration and mining of multiple, heterogeneous data sources, sponsored by the National High Technology Research and Devel- opment 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 informa- tion 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.

5 Related Work

5.1 Big Data Mining Platforms (Tier I)

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics 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 a 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."

Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in perfor- mance with relational databases. Improving the perfor- mance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel program- ming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algo- rithms based on the simple MapReduce programming model on multicore processors. Ten classical data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, the indepen- dent variable analysis, Gaussian discriminant analysis, expectation maximization, and back-propagation neural networks [14]. With the analysis of these classical machine learning algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could be per- formed on different subsets independently and achieve penalization executed easily on the MapReduce program- ming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then, various summation operations could be performed on the Mapper nodes to collect intermediate results. Finally, learning algorithms are executed in parallel through merging summation on Reduce nodes. Ranger et al. [39] proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multicore and multi- processor systems, and realized three data mining algo- rithms including k-Means, principal component analysis, and linear regression. Gillick et al. [22] improved the MapReduce's implementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning, iterative learning, and query-based learning in the MapReduce framework, studied data sharing between computing nodes involved in parallel learning algorithms, distributed data storage, and then showed that the MapReduce mechanisms suitable for large-scale data mining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun [38] pro- posed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques. The implementa- tion 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 capabil- ities 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 Hadoop-ML, 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.

5.2 Big Data Semantics and Application Knowledge (Tier II)

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 classifica- tion. 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 encryp- tion 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.

5.3 Big Data Mining Algorithms (Tier III)

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 algo- rithms 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 knowl- edge 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].

6 CONCLUSIONS

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 chal- lenges 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.

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. 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, pp. 1015-1018, 2009.

[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.