Issues and Challenges in Big Data Mining

Sakshi
Assistant Professor, Department of Computer Science and Applications, Guru Nanak College, Ferozepur Cantt, Punjab, India

ABSTRACT

Big Data is fast becoming a big problem since last year. Big data refers to datasets which has large size and complexity. We can't capture, store, manage and analyze with typical database software tools. Data mining is highlighted buzzword that is used to describe the range of Big data analytics, with collection, extraction, analysis and statics. Big Data mining involves to extracting useful information from these huge sets of data and streams of data, due to its volume, velocity and variety. This paper describes an overview of Big Data mining, problems related to mining and the new opportunities. During discussion we include platform and framework for managing and processing large data sets. We also discuss the knowledge discovery process, data mining, and various open source tools with current condition, issues and forecast to the future.

Keywords: *— Data mining, Big data, Big data mining, Big data management Issues and Challenges*

I. INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [1]. Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. A widely accepted formal definition of data mining is given subsequently. According to this definition, data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data [2]. Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is a new term used to identify the datasets that are of large size and have grater complexity [3]. So we cannot store, manage and analyze them with our current methodologies or data mining software tools. Big data is a heterogeneous collection of both structured and unstructured data. Businesses are mainly concerned with managing unstructured data. Big Data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity. The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Mining the information helps organizations to make knowledge driven decisions. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules [4]. It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis [4]. Enormous amount of data are generated every minute. A recent study estimated that every minute, Google receives over 4 million queries, e-mail users send over 200 million messages, YouTube users upload 72 hours of video, Facebook users share over 2 million pieces of content, and Twitter users generate 277,000 tweets [5]. With the amount of data growing exponentially, improved analysis is required to extract information that best matches user interests. Big data refers to rapidly growing datasets with sizes beyond the capability of traditional data base tools to store, manage and analyse them. Big data is a heterogeneous collection of both structured and unstructured data. Increase of storage capacities, Increase of processing power and availability of data are the main reason for the appearance and growth of big data. Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making. The data may be enterprise specific or general and private or public. Big data are characterized by 3 V's: Volume, Velocity, and Variety [6].

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

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

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

Big Data mining refers to the activity of going through big data sets to look for relevant information. Big data samples are available in astronomy, atmospheric science, social networking sites, life sciences, medical science, government data, natural disaster and resource management, web logs, mobile phones, sensor networks, scientific research, telecommunications [7]. Two main goals of high dimensional data analysis are to develop effective methods that can accurately predict the future observations and at the same time to gain insight into the relationship between the features and response for scientific purposes. Big data have applications in many fields such as Business, Technology, Health, Smart cities etc. These applications will allow people to have better services, better customer experiences, and also to prevent and detect illness much easier than before [8].

The rapid development of Internet and mobile technologies has an important role in the growth of data creation and storage. Since the amount of data is growing exponentially, improved analysis of large data sets is required to extract information that best matches user interests. New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It allows running applications on systems with thousands of nodes with thousands of terabytes of data. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure. It runs Map Reduce for distributed data processing and is works with structured and unstructured data [6].

2 Data Mining

Knowledge discovery (KDD) is a process of unveiling hidden knowledge and insights from a large volume of data [9], which involves data mining as its core and the most challenging and interesting step (while other steps are also indispensable) . Typically, data mining uncovers interesting patterns and relationships hidden in a large volume of raw data, and the results tapped out may help make valuable predictions or future observations in the real world. Data mining has been used by a wide range of applications such as business, medicine, science and engineering. It has led to numerous beneficial services to many walks of real businesses – both the providers and ultimately the consumers of services. Applying existing data mining algorithms and techniques to real-world problems has been recently running into many challenges due to the inadequate scalability (and other limitations) of these algorithms and techniques that do not match the three Vs of the emerging big data. Not only the scale of data generated today is unprecedented, the produced data is often continuously generated in the form of streams that require being processed and mined in (nearly) real time. Delayed discovery of even highly valuable knowledge invalidates the usefulness of the discovered knowledge. Big data not only brings new challenges, but also brings opportunities – the interconnected big data with complex and heterogeneous contents bear new sources of knowledge and insights. Big data would become a useless monster if we don't have the right tools to harness its "wildness". We argue to consider big data as greatly expanded assets to human. All what we need then is to develop the right tools for efficient store, access, and analytics (SA2 for short). Current data mining techniques and algorithms are not ready to meet the new challenges of big data. Mining big data demands highly scalable strategies and algorithms, more effective preprocessing steps such as data filtering and integration, advanced parallel computing environments (e.g., cloud Paas and IaaS), and intelligent and effective user interaction. Next we examine the concept and big data and related issues, including emerging challenges and the (foregoing and ongoing) attempts initiated on dealing with big data.

BIG DATA

We are awash in a flood of data today. There is variety of application areas, from where data is being collected at unmatched scale. According to McKinsey [10], Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. There is no exact definition of how Big a dataset is necessary to considered as Big Data. According to O'Reilly "Big data is data that exceeds the processing capacity of conventional database systems. The data is large in size, which moves too fast, and these data does not fit in the structures of existing database architectures. For getting value from these data, definitely there is an alternative way to process it." Big data has 3 V's characteristic which was describe by Doug Laney [11].

- **Volume**: machine-generated data is produced in much larger quantities than traditional data. For example, a single jet engine can generate 13TB of data in 25 minutes.
- **Variety**: In current day's data comes in different types of formats such as text, sensor data, audio, video, graph, and many more.
- **Velocity**: data comes as streams and we need to find interesting facts from it in the real time i.e. social media data stream.

But in current scenarios, there are two more V's:

- **Variability**: defined as the many ways in which the data may be variance in meaning, in lexicon. Differing questions which require different interpretations.
- **Value:** this is the most important feature of Big data. This feature describes for costs a lot of money to implement IT infrastructure systems to store Big data, and businesses are going to require a return on investment.

Gartner [12] in 2012 summarizes the definition of Big data as high volume, velocity and variety information assets which demand cost-effective, information processing tools for enhanced insight and decision making. There are large gap between demands of the Big data and capabilities of the current DBMSs for storage, manage, sharing, search and visualize. To overcome this large gap, Hadoop was introduced which is the core of Big data. Hadoop architecture that has a distributed file system, data storage platforms and an application layer that manages distributed processing, parallel computation, workflow and configuration management for unstructured data. There are many other non-relational databases such as NoSQL databases and MPP system that are also scalable, Networkoriented, semi-structured. With the emergence of Big Data, traditional RDBMS, MPP are transitioning into a new role of supporting Big Data management by processing structured datasets as outputs of Hadoop or MapReduce technologies.

To overcome the scalability of Big Data Google created a programming model named MapReduce [13] Which was facilitated by GFS (Google File System [14]), a distributed file system where the data can be simply partitioned over thousands of nodes in a cluster. Afterward, Yahoo and other Big companies created an Apache open-source version of Google's MapReduce framework, called Hadoop MapReduce. It uses the Hadoop Distributed File System (HDFS) an open source version of the Google's GFS. The MapReduce framework allows users to define two functions, map and reduce, which process large number data in parallel [15]. Users specify a map function a key/Value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate value associated with the same intermediate key.

4. BIG DATA MINING

In 1998, 'Big Data' term was appeared for the first time by John Mashey in his slide with title of "Big Data and the Next Wave of InfraStress" [16]. First book was published on the Big data mining in 1998 by Weiss and Indrukya [17].

However, the first academic paper with Big data was present in the 2000 by Diebold [18]. The goals of Big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous Big data has the potential to maximize our knowledge and in sights in the target domain.

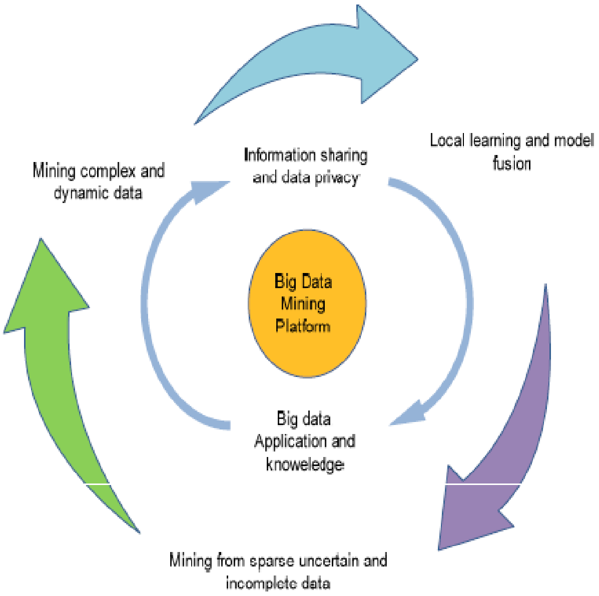
Mining complex and dynamic data
Information sharing and data privacy
Local learning and model fusion
Big Data Mining Platform
Big data Application and knoweledge
Mining from sparse uncertain and incomplete data

Figure3: A Big data mining framework

Big Data mining is necessary in many sectors:

Public sector: enables government departments and developmental organizations to analyze large amount of data across populations and to provide better governance and service.

Financial service: making better trading and risk decisions, improve product by better customer identification and marketing campaign.

Healthcare: mining DNA of each person, to discover, monitor and improve health aspects of every one.

Manufacturing: finding new opportunities to predict maintenance problems enhance manufacturing quality and reduce costs using Big Data.

Telecommunications: need of real-time data mining of data generated by mobile devices including phone calls, text messages, applications, and web browsing for better customer service and to build on retention and loyalty.

Retails: Big data mining offers numerous opportunities to retailers to improve marketing, merchandising, operations, supply chain and develop new business models

Other industries: mining can also be used in many other industries such as Oil and gas, transportation, GPS system and satellite.

5 Issues and Challenges

Our subsequent discussion centers on the following key issues and challenges: heterogeneity (or variety), scale (or volume), speed (or velocity), accuracy and trust, privacy crisis, interactiveness, and garbage mining (This section is supposedly the most interesting one of this paper).

5.1 Variety and Heterogeneity

In the past, data mining techniques have been used to discover unknown patterns and relationships of interest from structured, homogeneous, and small datasets (from today's perspective). Variety, as one of the essential characteristics of big data, is resulted from the phenomenon that there exists nearly unlimited different sources that generate or contribute to big data. This phenomenon naturally leads to the great variety or heterogeneity of big data. The data from different sources inherently possesses a great many different types and representation forms, and is greatly interconnected, interrelated, and delicately and inconsistently represented. Mining from such a gigantic and heterogeneous dataset, which is typically a tremendous network of interrelated data elements of diverse types, such as an academic social network consisting of authors, papers, conferences, universities, and companies, containing links such as work-at, write, written-by, appear-in, and present, etc.

Mining such a dataset, the great challenge is perceivable and the degree of complexity is not even imaginable before we deeply get there. Heterogeneity in big data also means that it is an obligation (rather than an option) to accept and deal with structured, semi-structured, and even entirely unstructured data simultaneously. While structured data can fit well into today's database systems, semi-structured data may partially fit in, but unstructured data definitely will not. Both semi-structured and unstructured data are typically stored in files. This is especially so in data-intensive, scientific computation areas [19]. Nevertheless, though bringing up greater technical challenges, the heterogeneity feature of big data means a new opportunity of unveiling, previously impossible, hidden patterns or knowledge dwelt at the intersections within heterogeneous big data. We shed a little more light on the implied challenge and the opportunity by looking into the examples from a familiar scenario in the following.

First, as a classic data mining example, we consider a simple grocery transaction dataset that records only one type of data, i.e., goods items. Examples insights [20] that might be mined from this dataset may include, e.g., the famous association of "beer and diapers" showing a strong linkage between the two items, and popular items like milk that are almost always purchased by customers, showing strong linkage of milk to all other items. In contrast to that, big data mining must deal with semi-structured and heterogeneous data. Now we generalize the aforementioned simple example by extending the scenario to an online market such as eBay. The dataset now is a richer network consisting of at least three different types of objects: items, buyers, and sellers (still this scenario may not be considered complex enough to demonstrate the complexity in big data mining). Interrelation may broadly exist, e.g., between commodity items in the form of "bought with", between sellers and items in the form of "sell" and "sold by", between buyers and items in the form of "buy" or "bought by", and between buyers and sellers in the form of "buy from" and "sold to". This data network has different types of objects and relationships (indicating a light shade of heterogeneity). We speculate that existing data mining techniques would not (if applicable at all) maximally uncover the hidden associations and insights in this data network.

For a heterogeneous set of big data, trying to construct a single model (if doable at all) would most likely not result in good-enough mining results; thus constructing specialized, more complex, multi-model systems is expected [21]. An interesting algorithm following this spirit is proposed in [22] that first determines whether the given dataset is truly heterogeneous, and if so, it then partitions the set into homogeneous subsets and constructs a specialized model for each homogeneous subset. Partitioning, as an intuitive approach, would speed up the process of knowledge discovery from heterogeneous big data. However, potential patterns and knowledge may miss the opportunity of being discovered after partitioning if important relationships (often implicit) crossing distinct homogeneous regions are not adequately retained.

The social community mining problem has recently received a lot attention from the researchers. This problem desires “multi-network, user-dependent, and query based analysis” [23]. It conveys that the intersections between multiple networks bear potential knowledge and insights that may not be discovered if a homogenous model is to be enforced.

Mining from heterogeneous information networks is a promising frontier of current data mining research [24]. Relational databases have been used to capture the heterogeneous information networks and new methods for in-depth network-oriented data mining and analysis have been proposed [24]. However, the degree of the heterogeneity captured does not reflect the real degree of the inherent heterogeneity existing in the big data. Mining hidden patterns from heterogeneous multimedia streams of diverse sources represents another frontier of data mining research. The output of this research has broad applicability such as detection of spreading dangerous diseases and prediction of traffic patterns and other critical social events (e.g., emerging conflicts and wars).

Like data mining, the process of big data mining shall also starts with data selection (from multiple sources). Data filtering, cleaning, reduction, and transformation then follow. There emerge new challenges with each of these preprocessing steps. With data filtering, how do we make sure that the discarded data will not severely degrade the quality of the eventually mined results under the complexity of great heterogeneity of big data? The same question could be adapted and asked to all other preprocessing steps and operations of the data mining process.

5.2 Scalability

The unprecedented volume/scale of big data requires commensurately high scalability of its data management and mining tools. Instead of being timid, we shall proclaim the extreme scale of big data because more data bears more potential insights and knowledge that we have no chance to discover from conventional data (of smaller scales). We are optimistic with the following approaches that, if exploited properly, may lead to remarkable scalability required for future data and mining systems to manage and mine the big data: (1) cloud computing that has already demonstrated admirable elasticity, which, combined with massively parallel computing architectures, bears the hope of realizing the needed scalability for dealing with the volume challenge of big data; (2) advanced user interaction support (either GUI- or language-based) that facilitates prompt and effective system-user interaction. Big data mining straightforwardly implies extremely time-consuming navigation in a gigantic search space, and prompt feedback/interference/guidance from users (ideally domain experts) must be beneficially exploited to help make early decisions, adjust search/mining strategies on the fly, and narrow down to smaller but promising subspaces.

5.3 Speed/Velocity

For big data, speed/velocity really matters. The capability of fast accessing and mining big data is not just a subjective desire, it is an obligation especially for data streams (a common format of big data) – we must finish a processing/mining task within a certain period of time, otherwise, the processing/mining results becomes less valuable or even worthless. Exemplary applications with real-time requests include earthquake prediction, stock market prediction and agent-based autonomous exchange (buying/selling) systems. Speed is also relevant to scalability – conquering or partially solving anyone helps the other one.

The speed of data mining depends on two major factors: data access time (determined mainly by the underlying data system) and, of course, the efficiency of the mining algorithms themselves. Exploitation of advanced indexing schemes is the key to the speed issue. Multidimensional index structures are especially useful for big data. For example, a combination of R- Tree and KD-tree [25] and the more recently proposed FastBit [21, 22] (developed by the data group at LBNL) shall be considered for big data. Besides, design of new and more efficient indexing schemes is much desired, but remains one of the greatest challenges to the research community. An additional approach to boost the speed of big data access and mining is through maximally identifying and exploiting the potential parallelism in the access and mining algorithms. The elasticity and parallelism support of cloud computing are the most promising facilities for boosting the performance and scalability of big data mining systems. It is interesting to note that the MapReduce parallel computing model is applicable to only a rather limited class of data-intensive computing problems.

Therefore, design of new and more efficient parallel computing models besides MapReduce is greatly desired, but calls for really creative minds.

5.4 Accuracy, Trust, and Provenance

In the past, data mining systems were typically fed with relatively accurate data from well-known and quite limited sources, so the mining results tend to be accurate, too; thus accuracy and trust have never been a serious issue for concern. With the emerging big data, the data sources are of many different origins, not all well-known, and not all verifiable. Therefore, the accuracy and trust of the source data quickly become an issue, which further propagates to the mining results as well. To (at least partially) solve this problem, data validation and provenance tracing become more than a necessary step in the whole knowledge discovery process (including data mining). History has repeatedly proven that challenges always comes hand-in-hand with opportunities (sometimes unnoticeably). In the case of big data, the copious data sources and gigantic volumes provide rich sources to extract additional evidences for verifying accuracy and building trust on the selected data and the produced mining results.

The vast volume of big data attributes additional characteristics – high dynamics and evolution. So an adequate system for big data management and analysis must allow dynamic changing and evolution of the hosted data items. This makes data provenance an integral feature in any system that deals with big data [26]. Provenance relates to the evolution history or the origin that a data item was extracted or collected from. The provenance relationships in big data often form a large collection of interrelated derivation chains, resulting in, more generally, a DAG. Trust measures are not and should not be treated static. When data evolves, trust measures shall change or be updated, too. Several unsupervised learning methods have been proposed in [27] and [28] to discover the trust measures of suspected data sources using other data sources as testimony (Here the assumed philosophy of proof is that one does not adequately prove himself innocent without having a third party's testimony). Reference [29] has shown that semi-supervised learning methods that start with ground truth data may provide higher accuracy and trust on the source data. In the context of big data, innovative methods that can run on parallel platforms (such as cloud PaaS and IaaS) dealing with scalable data with numerous sources are highly desired.

Provenance directly contributes to accuracy and trust of the source data and the derived (or mined) results. However, provenance information may not be always recorded or available. When the missing provenance of some data becomes a keen interest of the users, data mining can be reversely applied to derive and verify the provenance. Without a great many sources in the past, many provenance mining problems are unsolvable. History and archeology researches have raised a very interesting class of provenance mining problems. For example, the old question that whether Native Americans were originated from eastern Asia, after decades of debates, is still undetermined. With the advent of big data and mining tools, now we can glimpse the hope of finding the best answer to this and other questions of this type in the near future. We would rather believe the World Wide Web, as the largest data and knowledge base (indeed the Google executives firmly hold on this vision), bears sufficient information needed to derive the best answer to this and other similar questions, and yet the volume of this largest big data repository still keeps growing at an unprecedented pace. We foresee the big data mining technology will soon be able to answer many big questions like the above one though mining the whole World Wide Web as a single dataset (Digesting, consolidating, and deriving the best answer to the above question require the capacity that is way beyond the human brainpower).

5.5 Privacy Crisis

Data privacy has been always an issue even from the beginning when data mining was applied to real-world data. The concern has become extremely serious with big data mining that often requires personal information in order to produce relevant/accurate results such as location-based and personalized services, e.g., targeted and individualized advertisements. Also, with the huge volume of big data such as social media that contains tremendous amount of highly interconnected personal information, every piece of information about everybody can be mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that individual instantly disappears. You might ask, how could this be possible? Well, it is already a reality that every transaction regarding our daily life is being pushed to online and leaves a trace there: we comminute with friends via email, instant message, blog, and Facebook; we do shopping and pay our bills online too; and yet, credit card companies hold our confidential identity information; your payroll office has your personal information, too; your home phone number and address are listed in the region's directory that everyone can access; last month, you had a birthday party that disclosed your exact birthday to the circle of your friends, and some of them posted your birthday party in blogs, ... Thanks goodness, everyone so far has the righteous sense of protecting your confidential personal information, but the possibility of unintended leaking cannot be ruled out once and forever, and no leaking today does not guarantee impermeable tomorrow. As time goes, every piece of your personal information will be scattered here or there (hopefully not all available from one location). Well, we have desperately wanted and are diligently working toward powerful mining tools capable of mining a great portion or even the whole Web. So you shall not doubt such powerful mining tools or systems one day will be able to find confidential information of you (and actually of everyone else) – it's now just a matter of time. Everyone would easily gain the privilege of using such powerful tools (via SaaS on the cloud), mine your privacy, and see you entirely "naked". Without the shield of any privacy protecting you, a bad guy could open a new credit card account in your name, and transfer your hard-earned money away from your bank account... Everything seems becoming possible! Imagine how big a social disaster it would be when everyone in the US, for example, can access everyone else's social security number and other identity information, name, address, birthday, birthplace, phone numbers, etc. Even credit card companies do not ask for all this information when one requests to open a new account on the phone. So we definitely run the risk of living transparently or "naked" in an era of no privacy. Should we be proud to say that one day, we will live in a world that everyone can perfectly pretend to be any other one? Well, when anybody can "become" another body as s/he wishes, we get completely separated from our true identities. Now we need most seriously ask ourselves: would we rather to wear the "the emperor's new clothes"? The answer is certainly "no" as we all believe. Then what are the possible countermeasures? Apparently, we urgently need proper policies and approaches to manage sharing of personal data, while legitimate data mining activities shall still be granted facilitated. As said in [34], the privacy issue calls for "the development of a model where the benefits of data for businesses and researchers are balanced against individual privacy rights" [30]. The foundations of data mining need to be reformulated when dealing with big data "in such a way that privacy protection and discrimination prevention are embedded in the foundations themselves, dealing with every moment in the data-knowledge life-cycle: from (off-line and on-line) data capture, to data mining and analytics, up to the deployment of the extracted models" [31]. Measuring and prevention of privacy violation during knowledge mining are two related issues that call for serious research and innovative solutions.

5.6 Interactiveness

By interactiveness we mean the capability or feature of a data mining system that allows prompt and adequate user interaction such as feedback/ interference/ guidance from users. Interactiveness is relatively an inder emphasized issue of data mining in the past. When our society is now confronting the challenges of big data mining, interactiveness becomes a critical issue. Interactiveness relates to all the "three Vs" and can help overcome the challenges coming along with each of them. First, as we pointed out earlier, in order to conquer the volume related challenge of big data mining, prompt user feedback/guidance can help quickly narrow down into a much reduced but promising sub-space, accelerate the processing speed (or velocity) and increase system scalability. Second, the heterogeneity caused by the variety of big data straightforwardly induces accordingly high complexity in the big data itself and the mining results. Sufficient system interactiveness grants users the ability to visualize, pre evaluate, and interpret intermediate and final mining results. Such a facility might not be quite necessary for mining conventional datasets, but for big data, it is a must.

Great interactiveness boosts the acceptance of a complicated mining system and its mining results by potential users. In short, the head of the pyramid would be missing if adequate user interaction is not supported. Even though a data mining system has been very professionally designed, with perfect functional layers, without adequate interactiveness, the value of the system would be greatly discounted or simply rejected by users. Sufficient interactiveness is especially important for big data mining.

5.7 Garbage Mining

Who wants garbage when there are potentially gold? Garbage has no value. No one wants garbage. Everyone wants to get rid of garbage. In the real world, garbage collection is a business with profits. Garbage does not speak: "I am garbage, recycle me!" At home, our rooms are filled with stuff, and many items may never be needed, but we lack the wisdom to realize for sure. We easily fill up a 1000 GB disk in our desktop computers, whereas, only a small portion hoarded there are useful files (most of us would wholeheartedly agree on this!). We are not willing to spend time to clean up our disk space, more often, our memory becomes blurry as time goes and we don't remember the difference between two seemingly identical data files, and which file holds important consolidated data copied from other files that shall thus be recycled but we just did not promptly do so. Even cleaning up the disk space of desktop computer is a headache, not to mention to clean up the cyberspace! It has been a common sight that, e.g., you were searching the internet for customers' reviews and recommendations, say, for a good air-conditioning servicer in your area, and a professionally written blog caught your eyes, commending someone that you found already moved off the region after you made a couple of phone calls, and then you glimpsed the blog again, realizing the post date was in 2004. The blog space should have been cleaned; outdated and meaningless comments should have been deleted.

Unfortunately, this phenomenon does not only occur with blogs, it is common with the entire cyberspace. In the big data era, the volume of data generated and populated on the World Wide Web keeps increasing at an amazingly fast pace. In such an environment, data can (quickly) become outdated, corrupted, and useless; in addition, there is data that is created as junks (like junk emails). If the society does not pay attention and take actions now, as time goes, we will be flooded by junk data in the cyberspace. For the sake of having a relatively clean cyberspace and clean

World Wide Web, herein we call for attentions and research efforts. Cyberspace cleaning is not an easy task because of at least two foreseeable reasons: garbage is hidden, and there is an ownership issue – are you granted to collect someone else's garbage (provided you have the motivation)?

We propose applying data mining approaches to mine garbage and recycle it. We haven't yet noticed (to the best of our knowledge) the issue being realized and discussed anywhere else. But we believe garbage mining is a serious research topic, different but related to big data mining – for the sake the *sustainability* of our digital environment, "mining for garbage" (and cleaning it) is as important as "mining for knowledge" (the canonical sense of data mining). This is especially so in the new era of big data. We envision that in the future the society will develop mobile intelligent scavenger agents (with embedded garbage mining modules) and dispatch them to the cyberspace to autonomously and legitimately mine and clean up garbage in the cyberspace. Similarly, local versions of the intelligent scavenger agents shall be created and used to help clean up the disk space of desktop computers, if not entirely autonomously, at least interactively with necessary guidance and confirmation prompted from the users. "One man's trash is another's treasure". Garbage definition remains one of the greatest challenges.

6 Conclusions

We are living in the big data era where enormous amounts of heterogeneous, semi structured and unstructured data are continually generated at unprecedented scale. Big data discloses the limitations of existing data mining techniques, resulted in a series of new challenges related to big data mining. Big data mining is a promising research area, still in its infancy. In spite of the limited work done on big data mining so far, we believe that much work is required to overcome its challenges related to heterogeneity, scalability, speed, accuracy, trust, provenance, privacy, and interactiveness. This paper also provides an overview (though limited due to space limit) of state-of-the-art frameworks/platforms for processing and managing big data as well as platforms and libraries for mining big data. More specifically, we originally pointed out and analyzed the risk of privacy crisis which is deteriorated by big data and big data mining (Section 5.5) and first time proposed and formulated garbage mining – a critical issue in the big data era that has not been realized by others nor addressed anywhere else (Section 5.7). As our future work, we are at the stage of seriously planning a research project on cyberspace garbage mining to make the cyberspace a more sustainable environment. We tried to fill our discussions with sparking, constructive ideas. We hope we have (at least partially) gotten there.

II. REFERENCES

[1] Julie M. David, Kannan Balakrishnan, (2011), Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets, Int. J. of Computer and Electrical Engineering, Hong Kong, 3(1), pp163-169

[2] Julie M. David, Kannan Balakrishnan, (2011), Prediction of Learning Disabilities in School-Age Children using SVM and Decision Tree, Int. J. of Computer Science and Information Technology, ISSN 0975-9646, 2(2), pp829-835.

[3] Albert Bifet, (2013), "Mining Big data in Real time", Informatica 37, pp15-20

[4] Richa Gupta, (2014), "Journey from data mining to Web Mining to Big Data", IJCTT, 10(1),pp18-20

[5] http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/

[6] Priya P. Sharma, Chandrakant P. Navdeti, (2014), "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131

[7] Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), "Big Data:Overview", IJCTT, 9 (5) [8] Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, 14 (2), pp1-5

[9] Fayyad, U.M., Gregory, P.S., Padhraic, S.: From Data Mining to Knowledge Discovery: an Overview. In: Advances in Knowledge Discovery and Data Mining, pp. 1-36. AAAI Press, Menlo Park, CA (1996).

[10] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity.

[11] D. Laney. 3-D Data Management: Controlling Data volume, Velocity and Variety. META Group Research Note, February 6, 2001

[12] Gartner, http://www.gartner.com/it-glossary/big-data.

[13] Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: 6th Symposium on Operating System Design and Implementation (OSDI), pp. 137–150 (2004)

[14] Ghemawat, S., Gobioff, H., Leung, S.T.: The Google File System. In: 19th ACM Sympo-sium on Operating Systems Principles, Bolton Landing, New York, pp. 29–43 (2003)

[15] Dean, J., Ghemawat, S.: MapReduce: a Flexible Data Processing Tool. Communication of the ACM 53(1), 72–77 (2010)

[16] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[17] S.M.Weiss and N. Indurkhya. Predictive data mining: a practical guide. Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA, 1998.

[18] "F. Diebold. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[19] Greenwald, M., Fredian, T., Schissel, D., Stillerman, J.: A Metadata Catalog for Organization and Systemization of Fusion Simulation Data". Fusion Engineering & Design, vol. 87, no. 12, pp. 2205-2208. (2012).

[20] Shmueli, G., Patel, N.R., Bruce, P.C: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. (2nd ed) Wiley & Sons, Hoboken, New Jersey (2010).

[21] Obradovic, Z., Vucetic, S.: Challenges in Scientific Data Mining: Heterogeneous, Biased, and Large Samples. Technical Report, Center for Information Science and Technology Temple University, Chapter 1, pp.1-24 (2004).

[22] Vucetic S., Obradovic Z.: Discovering Homogeneous Regions in Spatial Data through Competition. In: 17th International Conference of Machine Learning, pp. 1095-1102. Stanford, CA (2000)

[23] Cai, D., Shao, Z., He, X., Yan, X., Han, J.: Mining Hidden Communities in Heterogeneous Social Network. In: 3rd International Workshop Link Discovery (LinkKDD), pp. 58-65 (2005).

[24] Sun, Y., Han, J., Yan, X., Yu, P.S.: Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. In: VLDB Endowment, vol. 5, no. 12, pp. 2022-2023 (2012).

[25] Zhang, X., Ai, J., Wang, Z., Lu, J., Meng, X.: An Efficient Multi-dimensional Index for Cloud Data Management. In: 1st International Workshop on Cloud Data Management, pp. 17-24. ACM Press, Hong Kong, China (2009).

[26] Agrawal, D., Bernstein, P., Bertino, E., et al: Challenges and Opportunities With big data – A Community White Paper Developed by Leading Researchers Across the United States(2012), http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.

[27] Yin, X., Han, J., Yu, P. S.: Truth Discovery with Multiple Conflicting Information Providers on the Web. In: 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1048-1052. San Jose, California (2007).

[28] Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating Conflicting Data: The Role of Source Dependence. In: VLDB Endowment, vol. 2, no. 1, pp. 550-561 (2009).

[29] Yin, X., Tan, W.: Semi-Supervised Truth Discovery. In: 20th International Conference on World Wide Web, pp. 217-226. Hyderabad, India (2011).

[30] Tene, O., Polonetsky, J.: Privacy in the Age of big data: A Time for Big Decisions. Stanford Law Review Online, vol. 64, pp. 63-69 (2012).

[31] Pedreschi, D., Calders, T., Custers, B., et al: big data Mining, Fairness and Privacy – A Vision Statement Towards an Interdisciplinary Roadmap of Research. In: Data Mining and Analytics Software, KDnuggets Review Online, vol. 11, no. 26 (2011).