From Big Data to Big Data Mining: Challenges, Issues, and Opportunities

Dunren Che[1], Mejdl Safran[1], and Zhiyong Peng[2]

[1] Department of Computer Science, Southern Illinois University
Carbondale, Illinois 62901, USA
{mejdl.safran@,dche@cs.}siu.edu
[2] Computer School, Wuhan University, Wuhan, 430072, China
peng@whu.edu.cn

Abstract. While "big data" has become a highlighted buzzword since last year, "big data mining", i.e., mining from big data, has almost immediately followed up as an emerging, interrelated research area. This paper provides an overview of big data mining and discusses the related challenges and the new opportunities. The discussion includes a review of state-of-the-art frameworks and platforms for processing and managing big data as well as the efforts expected on big data mining. We address broad issues related to big data and/or big data mining, and point out opportunities and research topics as they shall duly flesh out. We hope our effort will help reshape the subject area of today's data mining technology toward solving tomorrow's bigger challenges emerging in accordance with big data.

Keywords: data mining**,** big data, big data mining, big data management, knowledge discovery, data-intensive computation.

1 Introduction

The era of petabyte has come and almost gone, leaving us to confront the exabytes era now. Technology revolution has been facilitating millions of people by generating tremendous data via ever-increased use of a variety of digital devices and especially remote sensors that generate continuous streams of digital data, resulting in what has been called as "big data". It has been a confirmed phenomenon that enormous amounts of data have been being continually generated at unprecedented and ever increasing scales. In 2010, Google estimated that every two days at that time the world generated as much data as the sum it generated up to 2003. Regardless of the very recent "Big Data Executive Survey 2013" by NewVantage Partners [36] that states "It's about variety, not volume", many people (including the authors) would still believe the foremost issue with big data is scale or *volume*. Big data sure involves a great *variety* of data forms: text, images, videos, sounds, and whatever that may come into the play, and their arbitrary combinations (the type system shall remain constantly open). Big data frequently comes in the form of streams of a variety of

B. Hong et al. (Eds.): DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013.

types. Time is an integral dimension of data streams, which often implies that the data must be processed/mined in a timely or (nearly) real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in "a big data environment that can accelerate the time-to-answer critical business questions that demonstrate business values" [36]. The time dimension of bid data naturally leads to yet another key characteristic of big data – speed or *velocity*. We concur that big data is not *all* about size or volume, but size is the foremost characteristic of big data – size sparks off a series of interrelated vital challenges beyond size itself. Gartner analysts [29, 3] described the dominant characteristics of big data as "three Vs" – Volume, Velocity, and Variety (alternatively referred to as V3). Serious challenges are unfolded along each of the "V" axis. Bearing the brunt of criticism, the once very successful DBMSs have denounced no longer being able to meet the increasing demands of big data – it was "too big, too fast, and too hard" for existing DBMSs and tools [4] to satisfactorily handle. These challenges call for a new stack (or many alternate stacks) of highly scalable computing models, tools, frameworks and platforms, etc., being capable to tap into the most potential of today's best parallel and elastic computing facility – cloud computing.

Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed scalability with demonstrated elasticity and parallelism capacities. Numerous notable attempts have been initiated to exploit massive parallel processing architectures as reported in [5], [6], [7], [8], [9], [26], [19] and [17]. Google's novel programing model, MapReduce [5], and its distributed file system, GFS (Google File System) [6], represent the early groundbreaking efforts made in this line. We will shed more lights on these representative works later in this paper.

From the data mining perspective, mining big data has opened many new challenges and opportunities. Even though big data bears greater value (i.e., hidden knowledge and more valuable insights), it brings tremendous challenges to extract these hidden knowledge and insights from big data since the established process of knowledge discovering and data mining from conventional datasets was not designed to and will not work well with big data. The cons of current data mining techniques when applied to big data are centered on their inadequate scalability and parallelism. In general, existing data mining techniques encounter great difficulties when they are required to handle the unprecedented heterogeneity, volume, speed, privacy, accuracy, and trust coming along with big data and big data mining. Improving existing techniques by applying massive parallel processing architectures and novel distributed storage systems, and designing innovative mining techniques based on new frameworks/platforms with the potential to successfully overcome the aforementioned challenges will change and reshape the future of the data mining technology. Numerous research projects, as reported in [11], [12], [13], [14], [15] and [16], have been initiated in the last couple of years for the sake of overcoming the big data challenges. We will shed more lights on these projects later in this paper.

The theme of this paper is to provide an in-depth study on the issue of big data mining, its challenges and the perceivable opportunities. We also point to a few research topics that are either promising or much needed for solving the big data and big data mining problems. In order to make our discussion logical and smooth, we will start with a review of some essential and relevant concepts, including data mining, big data, big data mining, and the frameworks/platforms (completed or under construction) related to big data and big data mining.

The remainder of this paper is organized as follows. In Section 2, we briefly review the data mining concept and the foreseen challenges when the technology is applied to big data. In Section 3, we examine the concept of big data, comparing with conventional databases and reviewing the emerging platforms designed for big data and big data mining. In Section 4, we revisit data mining in the new context of big data, and point out emerging challenges. In Section 5, we discuss additional issues and challenges related to big data mining (in this section we especially draw attention to privacy crisis and garbage mining, which do not seem have been addressed by anyone else, to the best of our knowledge). Section 6 concludes our discussion.

2 Data Mining

Knowledge discovery (KDD) is a process of unveiling hidden knowledge and insights from a large volume of data [1], 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.

3 Big Data

We are sure living in an interesting era – the era of big data and cloud computing, full of challenges and opportunities. Organizations have already started to deal with petabyte-scale collections of data; and they are about to face the exabyte scale of big data and the accompanying benefits and challenges.

Big data is believed to play a critical role in the future in all walks of our lives and our societies. For example, governments have now started mining the contents of social media networks and blogs, and online-transactions and other sources of information to identify the need for government facilities, to recognize the suspicious organizational groups, and to predict future events (threats or promises). Additionally, service providers start to track their customers' purchases made through online, in-store, and on-phone, and customers' behaviors through recorded streams of online-clicks, as well as product reviews and ranking, for improving their marketing efforts, predicting new growth points of profits, and increasing customer satisfaction.

The mismatch between the demands of the big data management and the capabilities that current DBMSs can provide has reached the historically high peak. The three Vs (volume, variety, and velocity) of big data each implies one distinct aspect of critical deficiencies of today's DBMSs. Gigantic volume requires equally great scalability and massive parallelism that are beyond the capability of today's DBMSs; the great variety of data types of big data particularly unfits the restriction of the closed processing architecture of current database systems [4]; the speed/velocity request of big data (especially stream data) processing asks for commensurate real-time efficiency which again is far beyond where current DBMSs could reach. The limited *availability* of current DBMSs defeats the velocity request of big data from yet another angle (Current DBMSs typically require to first import/load data into their storage systems that enforces a uniform format before any access/processing is allowed. Confronted with the huge volume of big data, the importing/loading stage could take hours, days, or even months. This causes substantially delayed/reduced availability of the DBMSs).

To overcome this scalability challenge of big data, several attempts have been made on exploiting massive parallel processing architectures. The first such attempt was made by Google. Google created a programming model named MapReduce [5] that was coupled with (and facilitated by) the GFS (Google File System [6]), a distributed file system where the data can be easily partitioned over thousands of nodes in a cluster. Later, 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, to process a large number data entries in parallel [7]. More specifically, in MapReduce, the input is divided into a large set of key-value pairs first; then the map function is called and forked into many instances concurrently processing on the large key-value pairs. After all data entries are processed, a new set of key-value pairs are produced, and then the reduce function is called to group/merge the produced values based on common keys.

In order to match/support the MapReduce computing model, Google developed the BigTable – a distributed storage system designed for managing structured data. BigTable can scale well to a very large size: petabytes of data across thousands of commodity severs [8]. In the same spirit, Amazon created Dynamo [9], which is also a key-value pair storage system. The Apache open-source community acted quickly again, created HBase – an open-source version of Google's BigTable built on top of HDFS and Cassandra – an open-source version of Amazon's Dynamo. Apache Hive [25] is an open source data warehouse system built on top of Hadoop for querying and analyzing files stored in HDFS using a simple query language called HiveQL.

Hadoop is not alone; it has other competitor platforms. All these platforms lack many niceties existing in DBMSs. Some of the competitors improved on existing platforms (mostly on Hadoop), and others came up with a fresh system design. However, most of these platforms are still in their infancy. For example, BDAS, the Berkeley Data Analytics Stack [26], is an open-source data analytics stack developed at the UC Berkeley AMPLab for computing and analyzing complex data. It includes the following main components: Spark, Shark,, and Mesos. Spark is a high-speed cluster computing system that performs computations in memory and can outperform Hadoop by up to 100x. Shark is a large-scale data analysis system for Spark that provides a unified engine running SQL queries, compatible with Apache Hive. Shark can answer SQL queries up to 100x faster than Hive, and run iterative machine learning algorithms up to 100x faster than Hadoop, and can recover from failed mid-queries within seconds [27]. Mesos is a cluster manager that can run Hadoop, Spark and other frameworks on a dynamically shared pool of compute nodes. ASTERIX [19] is data-intensive storage and computing platform. As a research project, it was initiated by "the three database guys" at UC Irvine. Some notable drawbacks of Hadoop and other similar platforms, e.g., single system performance, difficulties of future maintenance, inefficiency in pulling data up to queries and the unawareness of record boundaries, are properly overcome in ASTERIX by exploring runtime models inspired by parallel database system execution engines [19]. In ASTERIX, the open software stack is layered in a different way that it sets the data records at the bottom layer, facilitating a higher-level language API at the top. While the majority of the big data management and processing platforms have been (or are being) developed to meet business needs, SciDB [17] is an open source data management and analytics (DMAS) software system for data-intensive scientific applications like radio astronomy, earth remote sensing and environment observation and modeling. The difference between SciDB and other platforms is that SciDB is designed based on the concept of array DBMS (i.e., raster data) where big data is represented as arrays of objects in unidimensional or multidimensional spaces. SciDB is designed to support integration with high-level imperative languages, algorithms, and very large scales of data [4].

In the next section, we discuss the attempts related to big data mining.

4 Big Data Mining

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 [2]. 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 insights in the target domain. However, this brings a series of new challenges to the research community. Overcoming the challenges will reshape the future of the data mining technology, resulting in a spectrum of groundbreaking data and mining techniques and algorithms. One feasible approach is to improve existing techniques and algorithms by exploiting massively parallel computing architectures (cloud platforms in our mind). Big data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of.

The need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM) has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop MapReduce. NIMBLE [11] is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel ML-DM algorithms, running on top of Hadoop. Apache's Mahout [12] is a library of machine learning and data mining implementations. The library is also implemented on top of Hadoop using the MapReduce programming model. Some important components of the library can run stand-alone. The main drawbacks of Mahout are that its learning cycle is too long and its lack of user-friendly interaction support. Besides, it does not implement all the needed data mining and machine learning algorithms. BC-PDM (Big Cloud-Parallel Data Mining) [13], as a cloud-based data mining platform, also based on Hadoop, provides access to large telecom data and business solutions for telecom operators; it supports parallel ETL process (extract, transform, and load), data mining, social network analysis, and text mining. BC-PDM tried to overcome the problem of single function of other approaches and to be more applicable for Business Intelligence. PEGASUS (Peta-scale Graph Mining System) [14] and Giraph [15] both implement graph mining algorithms using parallel computing and they both run on top of Hadoop. GraphLab [16] is a graph-based, scalable framework, on which several graph-based machine learning and data mining algorithms are implemented. The reported drawback of GraphLab is that it requires all data fitting into memory.

In the next section, we further expand our discussion along the key issues and challenges of big data mining.

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 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 [37]. 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 [10] 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" [24]. 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 [20]. 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 [20]. 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 sub-spaces.

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 [30] and the more recently proposed FastBit [18, 23] (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 [28]. 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 [31] and [32] 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 [33] 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" [34]. 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" [35]. 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 underemphasized 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 Conclusion

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.

References

1. 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 (1996)
2. Berkovich, S., Liao, D.: On Clusterization of big data Streams. In: 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York (2012)
3. Beyer, M.A., Laney, D.: The Importance of 'Big Data': A Definition. Gartner (2012)
4. Madden, S.: From Databases to big data. IEEE Internet Computing 16(3), 4–6 (2012)
5. 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)
6. Ghemawat, S., Gobioff, H., Leung, S.T.: The Google File System. In: 19th ACM Symposium on Operating Systems Principles, Bolton Landing, New York, pp. 29–43 (2003)
7. Dean, J., Ghemawat, S.: MapReduce: a Flexible Data Processing Tool. Communication of the ACM 53(1), 72–77 (2010)
8. Chang, F., Dean, J., Ghemawat, S., et al.: Bigtable: A Distributed Storage System for Structured Data. In: 7th Symposium on Operating Systems Design and Implementation, vol. 7, pp. 205–218. USENIX Association Berkeley, CA (2006)
9. DeCandia, G., Hastorun, D.: Jampani, et al: Dynamo: Amazon's Highly Available Key-Value Store. In: 21st ACM SIGOPS Symposium on Operating Systems Principles, pp. 14–17. Stevenson, Washington (2007)
10. Shmueli, G., Patel, N.R., Bruce, P.C.: Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, 2nd edn. Wiley & Sons, Hoboken (2010)
11. Ghoting, A., Kambadur, P., Pednault, E., Kannan, R.: NIMBLE: a Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on MapReduce. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, pp. 334–342 (2011)
12. Mahout, `http://lucene.apache.org/mahout/`
13. Yu, L., Zheng, J., Shen, W.C., et al.: BC-PDM: Data Mining, Social Network Analysis and Text Mining System Based on Cloud Computing. In: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1496–1499 (2012)
14. Kang, U., Tsourakakis, C.E., Faloutsos, C.: PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations. In: 9th IEEE International Conference on Data Mining, pp. 229–238 (2009)
15. Apache Giraph Project, `http://giraph.apache.org/`
16. Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A., Hellerstein, J.M.: Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. VLDB Endowment 5(8), 71–727 (2012)
17. Brown, P.G.: Overview of SciDB: Large Scale Array Storage, Processing and Analysis. In: ACM SIGMOD International Conference on Management of Data, pp. 963–968 (2010)
18. Wu, K.: FastBit: An Efficient Indexing Technology for Accelerating Data-intensive Science. Journal of Physics, Conference Series 16, 550–560 (2005)
19. Borkar, V.R., Carey, M.J., Li, C.: big data Platforms: What's Next? ACM Crossroads 19(1), 44–49 (2012)

20. Sun, Y., Han, J., Yan, X., Yu, P.S.: Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. VLDB Endowment 5(12), 2022–2023 (2012)
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, ch. 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, Stanford, CA, pp. 1095–1102 (2000)
23. Wu, K., Ahern, S.: Bethel, et al: FastBit: Interactively Searching Massive Data. SciDAC 180 (2009)
24. 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)
25. Apache Hive, `http://hive.apache.org/`
26. Berkeley Data Analytics Stack (BDAS),
`https://amplab.cs.berkeley.edu/bdas/`
27. Xin, R.S., Rosen, J., Zaharia, M., Franklin, M., Shenker, S., Stoica, I.: Shark: SQL and Rich Analytics at Scale. In: ACM SIGMOD Conference (accepted, 2013)
28. 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`
29. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner (2001)
30. 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 (2009)
31. 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, San Jose, California, pp. 1048–1052 (2007)
32. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating Conflicting Data: The Role of Source Dependence. VLDB Endowment 2(1), 550–561 (2009)
33. Yin, X., Tan, W.: Semi-Supervised Truth Discovery. In: 20th International Conference on World Wide Web, Hyderabad, India, pp. 217–226 (2011)
34. Tene, O., Polonetsky, J.: Privacy in the Age of big data: A Time for Big Decisions. Stanford Law Review Online 64, 63–69 (2012)
35. Pedreschi, D., Calders, T., Custers, B., et al.: big data Mining, Fairness and Privacy - A Vision Statement Towards an Interdisciplinary Roadmap of Research. Data Mining and Analytics Software, KDnuggets Review Online 11(26) (2011)
36. NewVantage Partners: Big Data Executive Survey (2013),
`http://newvantage.com/wp-content/uploads/2013/02/`
`NVP-Big-Data-Survey-2013-Summary-Report.pdf`
37. Greenwald, M., Fredian, T., Schissel, D., Stillerman, J.: A Metadata Catalog for Organization and Systemization of Fusion Simulation Data. Fusion Engineering & Design 87(12), 2205–2208 (2012)