International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 5 Ver VII || May 2018 || PP 100-109

Study of Big data Using Data Mining

Archana Panda[1], Mohini Prasad Mishra[2], Priyabrata Nayak[3]
[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Abstract: *Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and softwaretechniques. In most enterprise scenarios the volume of dataistoobigoritmovestoofastoritexceedscurrentprocessing capacity.Despitetheseproblems,bigdatahasthepotentialtohelp companies improve operations and make faster, moreintelligent decisions.ThispaperpresentsaHACEtheoremthatcharacterizes the features of the Big Data revolution, and proposes a BigData processing model, from the data mining perspective. This data-drivenmodelinvolvesdemanddrivenaggregationofinformationsources,miningandanalysis,userinterestmodeling,andsecurity and privacy considerations.*
Keywords: *Big Data, Data Mining, Heterogeneity, Autonomous Sources, Complex and Evolving Associations*

I. Introduction

DrYanMowonthe2012NobelPrizeinLiterature.ThisisprobablythemostcontroversialNobelprizeofthiscategory.Searchingon Google with “Yan Mo Nobel Prize,” resulted in 1,050,000 webpointerson theInternet(asof3January2013).“Forallpraisesaswellascriticisms,”saidMorecently,“Iamgrateful.”WhattypesofpraisesandcriticismshasMoactuallyreceivedoverhis31yearwritingcareer?AscommentskeepcomingontheInternetandin various news media, can we summarize all types of opinions indifferentmediainareal-timefashion,includingupdated,cross- referenced discussions by critics? This type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing.

Along with the above example, the era of Big Data has arrived [29,34,37].Everyday,2.5quintillionbytesofdataarecreatedand 90 percent of the data in the world today were produced within the past two years [26]. Our capability for data generation has neverbeensopowerfulandenormouseversincetheinventionoftheinformationtechnologyintheearly19thcentury.Asanother example,on4October2012,thefirstpresidentialdebatebetween President Barack Obama and Governor Mitt Romneytriggeredmorethan10milliontweetswithin2hours[46].Amongallthesetweets,thespecificmomentsthatgeneratedthemostdiscussions actuallyrevealedthepublicinterests,suchasthediscussionsabout medicare and vouchers. Such online discussions provide a new meanstosensethepublicinterestsandgeneratefeedbackinreal- time, and are mostly appealing compared to generic media,such asradioorTVbroadcasting.AnotherexampleisFlickr,apublic picturesharingsite,whichreceived1.8millionphotosperday,on average, from February to March 2012 [35]. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB)storageeverysingleday.Indeed,asanoldsayingstates:“a picture is worth a thousand words,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, socialevents,publicaffairs,disasters,andsoon,onlyifwehavethe power to harness the enormous amount of data.The above examplesdemonstratetheriseofBigDataapplicationswheredata collection has grown tremen-dously and is beyond the ability of commonly used software tools to capture, manage, and process withina“tolerableelapsedtime.”Themostfundamentalchallenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [40].Inmanysituations,theknowledgeextractionprocesshastobeveryefficientandclosetorealtimebecausestoringallobserved dataisnearlyinfeasible.Forexample,thesquarekilometerarray (SKA) [17] in radio astronomy consists of 1,000 to 1,500 15- meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes(GB)/seconddatavolume,thedatageneratedfromtheSKAareexceptionallylarge.Althoughresearchershaveconfirmed that interesting patterns, such as transient radio anomalies [41] canbediscoveredfromtheSKAdata,existingmethodscanonly work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classifica-tion for such Big Data.The remainder of the paper is structured as follows:InSectionII,weproposeaHACEtheoremtomodelBig Data characteristics. Section III summarizes the key challenges for Big Datamining.

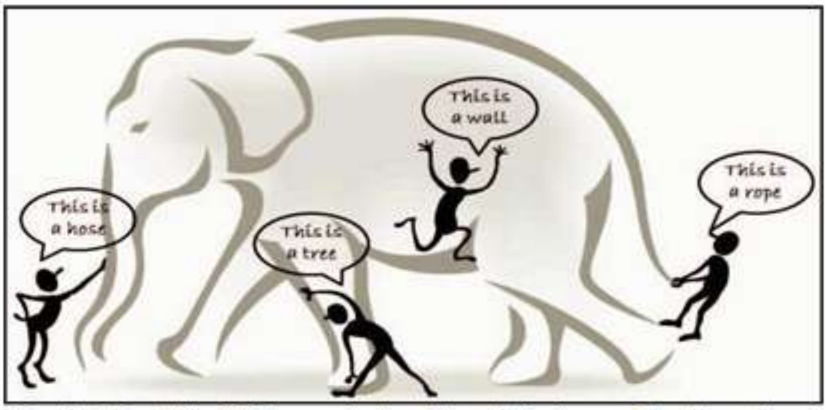

Fig. 1: The Blind Men and the Giant Elephant: The Localized (Limited) View of Each Blind Man Leads to a Biased Conclusion

Some key research initiatives and the authors' national research projects in this field are outlined in Section IV. Related work is discussed in Section V, and we conclude the paper in Section VI.

II. Big Data Characteristics: HACE Theorem

HACETheorem.BigDatastartswithlarge-volume,heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationshipsamong data.

Thesecharacteristicsmakeitanextremechallengefordiscovering useful knowledge from the Big Data. In a naı¨ve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Fig. 1), which will be the Big Data in this context. Thegoalofeachblindmanistodrawapicture(orconclusion)of theelephantaccordingtothepartofinformationhecollectsduring the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To maketheproblemevenmorecomplicated,letusassumethat(1) the elephant is growing rapidly and its pose changes constantly, and (2) each blind man may have his own (possible unreliable and inaccu-rate) information sources that tell him about biased knowledgeabouttheelephant(e.g.,oneblindmanmayexchange hisfeelingabouttheelephantwithanotherblindman,wherethe exchanged knowledge is inherently biased). Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous informationfromdifferentsources(blindmen)tohelpdrawabestpossiblepicturetorevealthegenuinegestureoftheeleph antina real-timefashion.Indeed,thistaskisnotassimpleasaskingeach blind man to describe his feelings about the elephant and then gettinganexperttodrawonesinglepicturewithacombinedview, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

A. Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities.Thisisbecausedifferentinforma-tioncollectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, each single human being in a biomedicalworldcanberepresentedbyusingsimpledemographic informationsuchasgender,age,familydiseasehistory,andsoon. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visualinformationfordoctorstocarrydetailedexaminations.For a DNA or genomic-related test, micro-array expressionimagesandsequencesareusedtorepresentthegeneticcodeinformationbecausethisisthewaythatourcurrentt echniquesacquirethedata. Undersuchcircumstances,theheterogeneousfeaturesrefertothe different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each singleobservation.

B. Autonomous Sources with Distributed and DecentralizedControl

Autonomous data sources with distributed and decentra-lized controlsareamaincharacteristicofBigDataapplications.Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting whereeachwebserverprovidesacertainamountofinformation andeachserverisabletofullyfunctionwithoutnecessarilyrelying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data-related applica-tions, such as Google,Flicker,Facebook,andWalmart,alargenumberofserverfarmsaredeployedallovertheworldtoensurenonstopservices andquickresponsesforlocalmarkets.Suchautonomoussources are not only the solutions of the technical designs, but also the resultsofthelegislationandtheregulationrulesindifferentcountries/ regions. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. Morespecifically,thelocalgovernmentregula-tionsalsoimpact on the wholesale management process and result in restructured data representations and data warehouses for localmarkets.

C. Complex and EvolvingRelationships

WhilethevolumeoftheBigDataincreases,sodothecomplexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This issimilar to using a number of data fields, such as age, gender, income, educationbackground,andsoon,tocharacterizeeachindividual.Thistypeofsamplefeaturerepresentationinherentlytreatseach individualasanindependententitywithoutconsideringtheirsocial connections, which is one of the most important factors of Wuetal.:Data Mining WithBigData

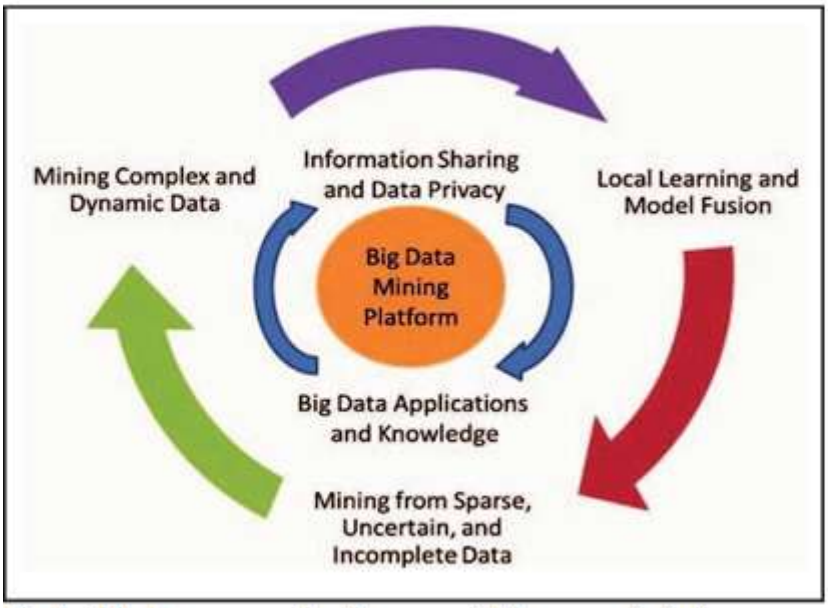

Fig.2:ABigDataprocessingframework:

Theresearchchallenges formathreetierstructureandcenteraroundthe“BigDatamining platform” (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, applicationdomain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual miningalgorithms.

The human society. Our friend circles may be formed based on the common hobbies or people are connectedbybiologicalrelationships.Suchsocialconnectionscommonlyexistnotonlyinourdailyactivities,butalsoare verypopularincyberworlds.For example,majorsocialnetworksites,suchasFacebookorTwitter, are mainly characterized by social functions such as friend- connectionsandfollowers(inTwitter).Thecorrelationsbetween individuals inherently complicate the whole data representation and any reasoning process on the data. In the sample-feature representation, individuals are regarded similar if they share similarfeaturevalues,whereasinthesample-feature-relationship repre-sentation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the indivi-duals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications,wherethekeyistotakethecomplex(nonlinear,manytomany)datarelationships,alongwiththeevolvingchanges,intoconsideration, to discover useful patterns from Big Datacollections.

III. Data Mining Challenges With Big Data

Foranintelligentlearningdatabasesystem[52]tohandleBigData, the essential key is to scale up to the exceptionally largevolume ofdataandprovidetreatmentsforthecharacteristicsfeaturedby the aforementioned HACE theorem. Fig. 2 shows a conceptual viewoftheBigDataprocessingframework,whichincludesthree tiers from inside out with considerations on data accessing and computing(TierI),dataprivacyanddomain knowledge(TierII), and Big Data mining algorithms (TierIII).

The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow,an effective computing platform will have to take distributedlarge- scaledatastorageintoconsiderationforcomputing.Forexample, typical dataminingalgorithmsrequirealldatatobeloadedintothemainmemory,this,however,isbecomingacleartechnicalbarrier for Big Data because moving data across different locations is expensive (e.g.,subjecttointensive networkcommunicationand other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center around semantics and domain knowledgefordifferentBigDataapplications.Suchinformation can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) andmining algorithms (TierIII).Forexample,dependingondifferentdomainapplications,thedataprivacyandinformationsharingmechanisms between data producers and data consumers can besignificantlydifferent.Sharingsensornetworkdataforapplicationslikewaterqualitymonitoringmaynotbediscouraged,whereasreleasingand sharingmobileusers'locationinformationisclearlynotacceptable for majority, if not all, applications. In addition to the above privacyissues,theapplicationdomainscanalsoprovideadditional informationtobenefitorguideBigDataminingalgorithmdesigns.
Forexample,inmarketbaskettransactionsdata,eachtransaction is considered independent and the discovered knowledge is typicallyrepresentedbyfindinghighlycorrelateditems,possibly with respect to different temporal and/or spatial restrictions. In a social network, on the other hand, users are linked and share dependencystructures.Theknowledgeisthenrepresentedbyuser communities,leadersineachgroup,andsocialinfluencemodeling, and so on. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high-level mining algorithm designs.

At Tier III, the data mining challenges concentrate onalgorithm designsintackling thedifficultiesraisedbytheBigDatavolumes, distributed data distributions, and by complex and dynamic data characteristics. The circle at Tier III contains three stages. First, sparse, heterogeneous, uncertain, incomplete, andmultisourcedataarepreprocessedbydatafusiontechniques.Second,complexanddynamicdataareminedafterpreprocessing.Third,theglobal knowledgeobtainedbylocallearningandmodelfusionistested and relevant information is fedback to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback.Inthewholeprocess,informationsharingisnotonlyapromiseofsmoothdevelopmentofeachstage,butalsoapurpose of Big Dataprocessing.Inthefollowing,weelaboratechallengeswithrespecttothethree tier framework in Fig.2.

A.Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to haveefficientaccessto,atleast,twotypesofresources:dataandcomputingprocessors.Forsmallscaledataminingtasks,asingle desktopcomputer,whichcontainsharddiskandCPUprocessors, is sufficient to fulfill the data mining goals. Indeed, many data mining algorithm are designed for this type of problem settings. Formediumscaledataminingtasks,dataaretypicallylarge(and possibly distributed) and cannot be fit into the main memory. Common solutions are to rely on parallel computing [33,43], or collectivemining[12]tosampleandaggregatedatafromdifferentsourcesandthenuseparallelcomputingprogramming(suchasthe Message Passing Interface) to carry out the miningprocess.
ForBigDatamining,becausedatascaleisfarbeyondthecapacity that a single personal computer (PC) can handle, a typical Big Data processing framework will rely on cluster computers with ahigh-performancecomputingplatform,withadataminingtask being deployed by running some parallel programming tools, such as Map Reduce or Enterprise Control Language (ECL), on alargenumberofcomputingnodes(i.e.,clusters).Theroleofthe softwarecomponentistomakesurethatasingledataminingtask, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes. For example, as of this writing, the world most powerful super computer Titan, whichisdeployedatOakRidgeNationalLaboratoryinTennessee, contains 18,688 nodes each with a 16-coreCPU.

Such a Big Data system, which blends both hardware and software components, is hardly available without key industrial stockholders'support.Infact,fordecades,companieshavebeen making business decisions based on transactional data stored in relational databases. Big Data mining offers opportunities to go beyond traditional relational databases to rely on less structured data:weblogs,socialmedia,e-mail,sensors,andphotographsthatcanbeminedforusefulinformation.Majorbusinessintelligencecompanies,suchIBM ,Oracle,Teradata,andsoon,haveallfeatured their own products to help customers acquire and organizethese diversedatasourcesandcoordinatewithcustomers'existingdata to find new insights and capitalize on hiddenrelationships.

B. Tier II: Big Data Semantics and Application Knowledge

SemanticsandapplicationknowledgeinBigDatarefertonumerous aspects related to the regulations, policies, user knowledge, and domain information. The two most important issues at this tier include(1)datasharingandprivacy;and(2)domainandapplicationknowledge.Theformerprovidesanswerstoresolvecon cerns on howdataaremaintained,accessed,andshared;whereasthelatter focuses on answering questions like "what are the under-lying applications ?" and "what are the knowledge or patterns users intend to discover from the data ?"

1. Information Sharing and DataPrivacy

Informationsharingisanultimategoalforallsystemsinvolving multiple parties [24]. While the motivation for sharing is clear, a real-world concern is that Big Data applications are related to sensitive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not resolve privacy con-cerns [19,25,42]. For example, knowing people's locations and theirp references, one can enable avarietyofuseful location-based services, but public disclosure of an individual's locations/movements over time can have serious consequences forprivacy.Toprotectprivacy,twocommonapproachesareto(1) restrict access to the data, such as adding certification or access controltothedataentries,sosensitiveinformationisaccessibleby alimitedgroupofusersonly,and(2)anonymizedatafieldssuchthatsensitiveinformationcannotbepinpointedtoanindivi -dual record [15]. For the first approach, common chal-lenges are to design secured certification or access control mechanisms, such thatnosensitiveinformationcanbemisconductedbyunauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indistinguishable from k 1 others. Common anonymiza-tion approaches are to use suppression, generalization, perturbation, andpermutationtogenerateanalteredversionofthedata,which is, in fact, some uncertain data.One of the major benefits of the data ammonization-based information sharing approaches is that,onceanonymzed,datacanbefreelysharedacrossdifferentpartieswithoutinvolvingrestrictiveaccesscontrols.This naturally leads to another research area namely privacy preserving data mining[30],wheremultipleparties,eachholdingsomesensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the data. This privacy preserving mining goal, in practice, can be solved through two types of approaches including (1) using special communication protocols,suchasYao'sprotocol[54],torequestthedistributions of the whole data set, rather than requesting the actual values of eachrecord,or(2)designingspecialdataminingmethodstoderive knowledge from anonymized data (this is inherently similar to the uncertain data mining methods).

2. Domain and ApplicationKnowledge

Domain and application knowledge [28] provides essential information for designing Big Data mining algorithms and systems. In a simple case, domain knowledge can help identify rightfeaturesformodelingtheunderlyingdata(e.g.,bloodglucose level is clearly a better feature than body mass in diagnosing Type II diabetes). The domain and application knowledge can alsohelpdesignachievablebusinessobjectivesbyusingBigDataanalyticaltechniques.Forexample,stockmarketdataar eatypical domainthatconstantlygeneratesalargequantityofinformation, such as bids, buys, and puts, in every singlesecond.Themarketcontinuouslyevolvesandisimpactedbydifferentfactors,suchasdomesticandinternationalne ws,governmentreports,andnatural disasters, and so on. An appealing Big Data mining task is to designaBigDataminingsystemtopredictthemovementofthemarketinthenextoneortwominutes.Suchsystems,evenift he predictionaccuracyisjustslightlybetterthanrandomguess,will bring significant business values to the developers[9].Withoutcorrectdomainknowledge,itisaclearchallengetofindeffectivematrices/measurestocharacteri zethemarketmovement,andsuchknowledgeisoftenbeyondthemindofthedataminers,althoughsomerecentresearchha sshownthatusingsocialnetworks,suchas Twitter,itispossibletopredictthestockmarketupward/downward trends [7] with goodaccuracies.

C. Tier III: Big Data Mining Algorithms

1. Local Learning and Model Fusion for Multiple InformationSources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources toacentralizedsiteforminingissystem-aticallyprohibitivedue to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models, just like the elephant and blind men case. Under such a circumstance, a Big Data mining system has to enable an informationexchangeandfusionmechanismtoensurethatalldistributedsites(orinformationsources)canworktogetherto achieveaglobaloptimizationgoal. Model mining and correlations are the key steps to ensure that modelsorpatternsdiscoveredfrommultipleinformationsources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (localminingandglobalcorrelation)process,atdata,model,and atknowledgelevels.Atthedatalevel,eachlocalsitecancalculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover localpatterns.Byexchangingpatternsbetweenmultiplesources, new global patterns can be synthetized by aggregating patterns across all sites [50]. At the knowledge level, model correlation analysis investigates the relevance between models gener-ated from different data sources to determine how relevant the data sourcesarecorrelatedwitheachother,andhowtoformaccurate decisions based on models built from autonomoussources.

2. Mining from Sparse, Uncertain, and Incomplete Data

Spare, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high-dimensional space (such as more than 1,000 dimensions) do not show clear trends or distribu-tions. For most machine learning and data mining algorithms, high-dimensional spare datasignificantlyde-terioratethereliabilityofthemodelsderived from the data. Common approaches are to employ dimension reductionorfeatureselection[48]toreducethedatadimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

Uncertain data are a special type of data reality where each data fieldisnolongerdeterministicbutissubjecttosomerandom/errordistributions.Thisismainlylinkedtodomainspecificapplications with inaccurate data readings and collections. For example,data produced from GPS equipment are inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recordinglocationisrepresentedbyameanvalueplusavariance to indicate expected errors. For data privacy-relatedapplications [36], users may intentionally inject randomness/errors into the data to remain anonymous. This is similar to the situation that an individual may not feel comfortable to let you know his/her exactincome,butwillbefinetoprovidearoughrangelike[120k, 160k].Foruncertaindata,themajorchallengeisthateachdataitem is represented as sample distributions but not as a singlevalue,somostexistingdataminingalgorithmscannotbedirectlyapplied.Commonsolutionsaretotakethedatadistributionsinto consideration to estimate model parameters. For example, error awaredatamining[49]utilizesthemeanandthevariancevalueswithrespecttoeachsingledataitemtobuildaNaı¨veBayes model for classification. Similar approaches have also been appliedfor decision trees or database queries. Incomplete data refer to the missingofdatafieldvaluesforsomesamples.Themissingvalues can be caused by different realities, such as the malfunction ofa sensornode,orsomesystematicpoliciestointentionallyskipsome values (e.g., dropping some sensor node readings to save power for transmission). While most modern data miningalgorithmshaveinbuiltsolutionstohandlemissingvalues(suchasignoringdatafieldswithmissingvalues),dataimputationisanestab-lished research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data). Many imputation methods [20] exist for this purpose, and themajorapproachesaretofillmostfrequentlyobservedvaluesortobuildlearningmodelstopredictpossiblevaluesforeachdata field, based on the observed values of a giveninstance.

3. Mining Complex and DynamicData

TheriseofBigDataisdrivenbytherapidincreasingofcomplex data and their changes in volumes and in nature [6]. Documents posted on WWW servers, Internet back-bones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving. For example, researchers have successfully usedTwitter,awell-known social networking site, to detect events such as earthquakes and major social activities, with nearly real-time speed and very high accuracy. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks [23]. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each otherwithasocialconnection.Suchaconnectionisquadraticwithrespecttothenumberofnodesinthenetwork,soamillion node network may be subject to one trillion connections. For a large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily useractions/interactionsintoconsideration,thescaleofdifficulty will be even moreastonishing.

Inspiredbytheabovechallenges,manydataminingmethodshave been developed to find interesting knowl-edge from Big Data with complex relationships and dynamically changing volumes. Forexample,findingcommunitiesandtracingtheirdynamically evolving rela-tionships are essential for understanding and managing complex systems [3], [10]. Discovering outliers in a socialnetwork[8]isthefirststeptoidentifyspammersandprovide safe networking environments to oursociety.

If only facing with huge amounts of structured data, users can solvethe problem simplybypurchasingmorestorageorimproving storageefficiency.However,BigDatacomplexityisrepresented in many aspects, including complex heterogeneous data types, complex intrinsic semantic associations in data, and complex relationshipnetworksamongdata.Thatistosay,thevalueofBigData is in its complexity.

Complexheterogeneousdatatypes.InBigData,datatypesincludestructureddata,unstructureddata,andsemistruc-tureddata,andso on.Specifically,therearetabulardata(relationaldatabases),text, hyper-text, image, audio and video data, and so on. Theexisting datamodelsincludekey-valuestores,bigtableclones,document databases,andgraphdatabases,whicharelistedinanascending order of the complexity of these data models. Traditional data models are incapable of handling complex data in the context of Big Data. Currently, there is no acknowledged effective and efficient data model to handle BigData.

Complex intrinsic semantic associations in data. News on the web,commentsonTwitter,picturesonFlicker,andclipsofvideo on YouTube may discuss about an academic award-winning event at the same time. There is no doubt that there are strong semantic associations in these data. Mining complex semantic associationsfrom“text-image-video”datawillsignificantlyhelp improve application system performance such as searchengines orrecommendationsystems.However,inthecontextofBigData, itisagreatchallengetoefficientlydescribesemanticfeaturesand to build semantic association models to bridge the semantic gap of various heterogeneous datasources.

Complex relationship networks in data. In the context of Big Data, there exist relationships between individuals. On the Internet, individuals are webpages and the pages linking toeach other via hyperlinks form a complex network. There also exist socialrelationshipsbetweenindividualsformingcomplexsocial networks, such as big relationship data from Facebook, Twitter, LinkedIn, and other social media [5], [13], [56], including call detail records (CDR), devices and sensors information [1], [44], GPSandgeocodedmapdata,massiveimagefilestransferredby theManageFileTransferprotocol,webtextandclick-streamdata [2], scientific information, e-mail [31], and so on. To deal with complex relationship networks, emerging research efforts have begun to address the issues of structure-and-evolution, crowds- and-interac-tion, and information-and-communication.

The emergence of Big Data has also spawned new computer architectures for real-time data-intensive proces-sing, suchas the open source Apache Hadoop project that runs on high- performance clusters. The size or complexity of the Big Data, includingtransactionandinteractiondatasets,exceedsaregular technicalcapabilityincapturing,managing,andprocessingthesedatawithinreasonablecostandtimelimits.InthecontextofBig Data,real-timeprocessingforcomplexdataisaverychallenging task.

IV. Research Initiatives and Projects

To tackle the Big Data challenges and “seize the opportunities afforded by the new, data driven resolu-tion,” the USNationalScienceFoundation(NSF),underPresidentObamaAdministration’sBigDatainitiative,announcedtheBIGDATAsolicitationin2012. Such a federal initiative has resulted in a number of winning projectstoinvestigatethefoundationsforBigDatamanagement (led by the University of Washington), analytical approaches for genomics-based massive data computation (led by Brown University), large scale machine learning techniques for high- dimensionaldatasetsthatmaybeaslargeas500,000dimensions (led by Carnegie Mellon University), social analytics for large- scalescientificliteratures(ledbyRutgersUniversity),andseveral others. These projects seek to develop methods, algorithms, frameworks, and research infrastructures that allow ustobringthemassiveamountsofdatadowntoahumanmanageableand interpretablescale.Othercoun-triessuchastheNationalNatural Science Foundation of China (NSFC) are also catching up with national grants on Big Dataresearch.

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

Integratingandminingbiodatafrommultiplesourcesinbiological networks, sponsored by the US National Science Foundation, MediumGrantNo.CCF-0905337,1October2009-30September 2013.
Issues and significance. We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and miningbiologicalnetworks.Wehaveexpandedandintegratedthe techniquesandmethodsininformationacquisition,transmission, and processing for information networks. We have developed methodsforsemantic-baseddataintegra-tion,automatedhypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refinemodels.
Big Data Fast Response. Real-time classification of Big Data Stream, sponsored by the Australian Research Council (ARC), Grant No. DP130102748, 1 January 2013 - 31 Dec. 2015.
Issuesandsignificance.Weproposetobuildastream-basedBig Dataanalyticframeworkforfastresponseandreal-timedecision making. The key challenges and research issuesinclude:

- DesigningBigDatasamplingmechanismstoreduceBigData volumes to a manageable size forprocessing;
- Building prediction models from Big Data streams. Such modelscanadaptivelyadjusttothedynamicchangingofthe data, as well as accurately predict the trend of the data in the future; and
- A knowledge indexing framework toensure real-time datamonitoring and classification for Big Data applications.

V. RelatedWork

A. Big Data Mining Platforms (TierI)

Due to the multisource, massive, dynamic characteristics of application data involved in a distributed environment, one characteristicsofBigDataistocarryoutcomputingonthepetabyte (PB),eventheexabyte(EB)-leveldatawithacomplexcomputing process. Therefore, utilizing a parallel computinginfrastructure, its corresponding programming language support, and analyze and mine the distributed data are the critical goals for Big Data processing“quality.”

B. BigDataSemanticsandApplicationKnowledge(Tier II)

In privacy protection of massive data, Ye et al. [55] proposed a multilayer rough set model, which can accurately describe the granularitychangeproducedbydifferentlevelsofgeneralization and provide a theoretical foundation for measuring the data effectivenesscriteriaintheanonymizationprocess,anddesigned a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization/refinement order for classifica- tion.ArecentpaperonconfidentialityprotectioninBigData[4] summarizes a number of methods for protecting public release data, including aggregation (such as k-anonymity, I-diversity, etc.),suppression(i.e.,deletingsensitivevalues),dataswapping (i.e., switching values of sensitive data records to preventusers

from matching), adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions. For applications involving Big Data and tremendous data volumes, itisoftenthecasethatdataarephysicallydistributedatdifferent locations, which means that users no longer physically possess thestorageoftheirdata.TocarryoutBigDatamining,havingan efficientandeffectivedataaccessmechanismisvital,especially for users who intend to hire a third party (such as data minersor data auditors) to process their data. Under such a circumstance, users’ privacy restrictions may include 1) no local data copies or downloading, 2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings,andmanyothers.InWangetal.[48],aprivacy-preserving public auditing mechanism for large scale data storage (such as cloud computing systems) has been proposed. The public key- based mechanism is used to enable third-party auditing (TPA), so users can safely allow a third party to analyze their data withoutbreachingthesecuritysettingsorcompromisingthedata privacy.

For most Big Data applications, privacy concerns focus on excluding the third party (such as data miners) from directly accessing the original data. Common solutions are to rely on someprivacy-preservingapproachesorencryp-tionmechanisms to protect the data. A recent effort by Lorch et al. [32] indicates thatusers’“dataaccesspatterns”canalsohaveseveredataprivacy issues and lead to disclosures of geographically co-locatedusersoruserswithcommoninterests(e.g.,twouserssearchingforthesamemaplocationsarelikelytobegeographi callycolocated).In their system, namely Shround, users’ data access patterns from the servers are hidden by using virtual disks. As a result, it can support a variety of Big Data applications, such as microblog search and social network queries, without compromising the userprivacy.

C. Big Data Mining Algorithms (TierIII)

Toadapttothemultisource,massive,dynamicBigData,researchers have expanded existing data mining methods in manyways,includingtheefficiencyimprovementofsinglesourceknowledgediscoverymethods[11],designingadatam iningmechanismfromamultisourceperspective[5051],aswellasthestudyofdynamicdataminingmethodsandtheanaly sisofstreamdata[12],[18].Themainmotivationfordiscoveringknowledgefrommassivedataisimprovingtheefficiency ofsingle-sourceminingmethods.Onthe basis of gradual improvement of computer hardware functions, researcherscontinuetoexplorewaystoimprovetheefficiencyofknowledgediscoveryalgorithmstomakethembetterfor massive data.Becausemassivedataaretypicallycollectedfromdifferent data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-worlddataoftencomeasadatastreamoracharacteristicflow,awellestablishedmechanismisneededtodiscoverknowled geand master the evolution of knowl-edge in the dynamic data source. Therefore,themassive,heterogeneousandrealtimecharacteristicsofmultisourcedataprovideessentialdifferencesbetwe ensingle- source knowledge discovery and multisource datamining.

Datastreamsarewidelyusedinfinancialanalysis,onlinetrading, medical testing, and so on. Static knowledge discoverymethodscannotadapttothecharacteristicsofdynamicdatastreams,suchascontinuity,variability,rapidity,an dinfinity,andcaneasilyleadto thelossofusefulinformation.Therefore,effectivetheoreticaland technical frameworks areneededtosupportdatastreammining18], [57].

Knowledge evolution is a common phenomenon in real-world systems. For example, the clinician’s treatment programs will constantly adjust with the conditions of the patient, such as familyeconomicstatus,healthinsurance,thecourseoftreatment, treatment effects, anddistribution.

VI. Conclusions

Drivenbyrealworldapplicationsandkeyindustrialstakeholdersandinitializedbynationalfundingagencies,m anagingandmining BigDatahaveshowntobeachallengingyetverycompellingtask. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the BigDataare1)hugewithheterogeneousanddiversedatasources,2)autonomouswithdistributedanddecentralizedcontr ol,and3)complexandevolvingindataandknowledgeassociations.Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values [27].ToexploreBigData,wehaveanalyzedseveralchallengesatthedata,model,andsystemlevels.TosupportBigDatami ning,high- performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At thedatalevel,theautonomousinformationsourcesandthevariety of the data collection environments, often result in data with complicatedconditions,suchasmissing/uncertainvalues.Inother situations, privacy concerns, noise, and errors can beintroduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge.At the model level, the key challenge is to generate global models bycombininglocallydiscoveredpatternstoformaunifyingview. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiplesourcestogainabestmodeloutoftheBigData.Atthe system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineeringdomains.WithBigDatatechnologies,wewillhopefullybeabletoprovidemostrelevantandmostaccurateso cialsensingfeedbacktobetter understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

VII. Acknowledgments

This work was supported by the National 863 Program of China (2012AA011005), the National 973 Program of China (2013CB329604), the National Natural Science Foundation of China (NSFC 61229301, 61273297 and 61273292),theUS National Science Foundation (NSF CCF-0905337), and the Australian Research Council (ARC) Future Fellowship (FT100100971) and Discovery Project (DP130102748). The authors would like to thank the anonymous reviewers for their valuable and constructive comments on improving thepaper.

References

[1]. R.Ahmed,G.Karypis,“AlgorithmsforMiningtheEvolution of Conserved Relational States in Dynamic Networks,” Knowledge and Information Systems, Vol. 33, No. 3, pp. 603-630, Dec.2012.
[2]. M.H.Alam,J.W.Ha, S.K.Lee,“Novel Approachesto Crawling Important Pages Early, ”Knowledge and Information Systems, Vol. 33, No. 3, pp 707-734, Dec.2012.
[3]. S.Aral,D.Walker,“IdentifyingInfluentialandSusceptible Members of Social Networks,” Science, Vol. 337, pp. 337- 341, 2012.

[4]. A. Machanavajjhala, J.P. Reiter,“Big Privacy: Protecting Confidentialityin Big Data,” ACM Cross roads, Vol.19,No. 1, pp. 20-23, 2012.
[5]. S. Banerjee, N. Agarwal,“Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, Vol. 33, No. 3, pp. 523-547, Dec. 2012.
[6]. E.Birney,“TheMakingofENCODE:LessonsforBig-Data Projects,” Nature, Vol. 489, pp. 49-51,2012.
[7]. J.Bollen,H.Mao,X.Zeng,“TwitterMoodPredictstheStock Market,” J. Computational Science, Vol. 2, No. 1, pp. 1-8, 2011.
[8]. S. Borgatti, A. Mehra, D. Brass, G. Labianca,“Network Analysis in the Social Sciences,” Science, Vol. 323, pp. 892-895, 2009.
[9]. J. Bughin, M. Chui, J. Manyika,"Clouds, Big Data, and SmartAssets:TenTech-Enabled Business Trendsto Watch", McKin Sey Quarterly,2010.
[10]. D. Centola,“The Spread of Behavior in an Online Social Network Experiment,” Science, Vol. 329, pp. 1194-1197, 2010.
[11]. E.Y.Chang,H.Bai,K.Zhu,“Parallel Algorithms for Mining Large-ScaleRich-Media Data, ”Proc. 17th ACM Int’l Conf. Multi-media, (MM ’09,) pp. 917-918,2009.
[12]. R.Chen,K.Sivakumar, H.Kargupta, “Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,” Knowledge and Information Systems,Vol.6,No.2,pp.164- 187, 2004.
[13]. Y.-C.Chen,W.-C.Peng,S.-Y.Lee,“EfficientAlgorithmsfor Influence Maximization in Social Networks,” Knowledge andInformationSystems,Vol. 33,No.3,pp.577-601,Dec. 2012.
[14]. C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, K. Olukotun,“Map-Reduce for Machine Learning on Multicore,” Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS ’06), pp. 281-288,2006.
[15]. G.Cormode,D.Srivastava,“AnonymizedData:Generation, Models, Usage,” Proc. ACM SIGMOD Int’l Conf. Management Data, pp. 1015-1018,2009.
[16]. S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, J. McPherson,“Ricardo: Integrating R and Hadoop,” Proc. ACM SIGMOD Int’l Conf. Management Data (SIGMOD ’10), pp. 987-998. 2010.
[17]. P. Dewdney, P. Hall, R. Schilizzi, J. Lazio,“The Square Kilometre Array,” Proc. IEEE, Vol. 97, No. 8, pp. 1482- 1496, Aug.2009.
[18]. P.Domingos, G.Hulten,“Mining High-Speed Data Streams,” Proc. Sixth ACMSIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’00), pp. 71-80,2000.
[19]. G.Duncan,“PrivacybyDesign,”Science,Vol.317,pp.1178- 1179,2007.
[20]. B.Efron,“Missing Data, Imputation,and the Bootstrap,”J. Am. Statistical Assoc., Vol. 89, No. 426, pp. 463-475, 1994.
[21]. A.GhotingandE.Pednault,“Hadoop-ML:AnInfrastructure for the Rapid Implementation of Parallel Reusable Analytics,” Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS ’09),2009.
[22]. D. Gillick, A. Faria, J. DeNero,"MapReduce: Distributed Computing for Machine Learning", Berkley, Dec.2006.
[23]. M.Helft,“GoogleUsesSearchestoTrackFlu’sSpread,”The New York Times, [Online] Available: http://www.nytimes. com/2008/11/12/technology/ internet/12flu.html.2008.