Big data analytics: six techniques

Hong Shua,b

aThe State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China; bCollaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China

ABSTRACT

Big data have 4V characteristics of volume, variety, velocity, and veracity, which authentically calls for big data analytics. However, what are the dominant characteristics of big data analysis? Here, the analytics is related to the entire methodology rather than the individual specific analysis. In this paper, six techniques concerning big data analytics are proposed, which include: (1) Ensemble analysis related to a large volume of data, (2) Association analysis related to unknown data sampling, (3) High-dimensional analysis related to a variety of data, (4) Deep analysis related to the veracity of data, (5) Precision analysis related to the veracity of data, and (6) Divide-and-conquer analysis related to the velocity of data. The essential of big data analytics is the structural analysis of big data in an optimal criterion of physics, computation, and human cognition. Fundamentally, two theoretical challenges, ie the violation of independent and identical distribution, and the extension of general set-theory, are posed. In particular, we have illustrated three kinds of association in geographical big data, ie geometrical associations in space and time, spatiotemporal correlations in statistics, and space-time relations in semantics. Furthermore, we have illustrated three kinds of spatiotemporal data analysis, ie measurement (observation) adjustment of geometrical quantities, human spatial behavior analysis with trajectories, data assimilation of physical models and various observations, from which spatiotemporal big data analysis may be largely derived.

1. Big data and its 4V characteristics

There exist two common sources of big data, ie collective gathering and individual generation. Examples of collective gathering big data are smart city data, national geographic conditions monitoring data, and earth observation data (Li, Yao, and Shao 2014). In general, collective gathering data are obtained with sampling strategies, and its data quality is high. Examples of individual generation data are electronic business data, social media data on the Internet, and crowd sourcing data (Shaw and Fang 2014). Individual generation data is obtained in more freedom, and its usability is relatively low.

The big data technology arises from the rapid development of the computer and communication technology, which evenly causes the paradigm transition of human cognition about our living world. On the one hand, the computer technology, particularly the digital and intelligent technology, is much advanced at the stages of data sampling, storage management, data computation, and data communication. Meanwhile, the cost of data collection has been reduced greatly. Since computing is ubiquitous, anything can be digitalized and defined with software. On the other hand, human being is highly intended to understand and change the geographical world around us. Over centuries, human knowledge about the geographical reality is mostly obtained in rationalism rather than in empiricism. However, so far empiricism in inductive logic or empirical knowledge learning is getting increasingly important for human knowledge construction. In Rao’s saying, all judgments must be statistics on the basis of hypothesis (Rao 1997). In the traditional scientific paradigm, the theory is proved with the experiments. In the current scientific paradigm, the scientific finding is often obtained by computer simulation, and is mainly explored from multi-source observations.

Big data is emerging in the lifetime of data collection, management, analysis and applications. In data collection, space-, air- and ground-based sensor networks, eg global earth observation system of systems (GEOSS), are built. Even each person is considered as a sensor on the Internet, eg volunteered geographic information (VGI) or crowd sourcing data. In data management, distributed file system and cloud computing are developed for today’s tremendous unstructured data. In data analysis, statistics and machine learning are widely employed for big data analysis. In data applications, online and offline business are smoothly combined, and precision...
industries have appeared, e.g. precision agriculture, precision medicine, and precision marketing.

In the scientific community, the term “big data” is put forward first by computer scientist. Generally, computer software development is concerned with three aspects of computation, i.e. problems computability, algorithm complexity, and distributed intelligence. For problem computability, Kurt Gödel’s incompleteness theorem and Alan Turing’s halting problem are fundamental. Kurt Gödel stated that no consistent system of axioms, whose theorems can be listed by an “effective procedure” (i.e. an algorithm), is capable of proving all truths about the relations of the natural numbers (arithmetic). For any such system, there will always be statements about the natural numbers that are true, but that are not proved within the system, and such a system cannot demonstrate its own consistency. Alternatively, the halting problem is the problem of determining, from a description of an arbitrary computer program and an input, whether the program will finish running or continue to run forever. In 1936, Alan Turing proved that a general algorithm to solve the halting problem because all possible program-input pairs cannot exist. The halting problem is undecidable over Turing machines. Computing is ubiquitous nowadays, but there still exist a large number of problems which cannot be solved by the computer. For those computable problems, further there exist the issue of algorithm complexity of problem solving, including time complexity and space complexity. Time complexity refers to the increasing speed of the number of problem solving steps against the scale of problems. Space complexity refers to the increasing speed of the size of problem solving storage against the scale of problems. It is admitted that problems solving in polynomial time complexity are easy, and problems solving in exponential time complexity are difficult. In a computing environment of multi-cores of CPU, GPU, or FPGA (field programmable gate array), the time complexity of algorithms can be reduced a lot. In a computing environment of the networked and high-density storage, the space complexity of algorithms can be greatly reduced. Following the issues of problem computability and algorithm complexity, artificial intelligence has become the major challenge of software. In sum, the computer software technology is explored from the computable problems, feasible algorithms, and trusted software.

By computer technology driving, it is summarized that big data have 4V characteristics of volume, variety, velocity, and veracity (Barwick 2012; Hilbert 2015). They can be explained as follows:

(1) The characteristic of volume refers to the large quantity of generated and stored data.
(2) The characteristic of variety refers to the multi-type and multi-source of data.
(3) The characteristic of velocity refers to the high speed at which the data are generated and processed.
(4) The characteristic of veracity refers to the high quality and value of captured and analyzed data. Data quality is measured comprehensively with inherent information content and its user demands satisfaction. Potentially, big data own large gross value but sparsely dense value. The veracity, roughly termed data value or data usability, seems especially important in practice.

2. Revisiting two mathematical theories for big data analysis

Originally, data are a set of values of qualitative or quantitative variables. The independent and identical distribution theory (IID) and set theory are common mathematics used by data analysis of statistics and optimization. Under violation of IID, spatiotemporal autocorrelation and generalized distribution are often assumed in geographical big data statistics. Extending set theory, probability-measured set, metric space, and topological space are often used in big data analysis.

2.1. Independent and identical distribution

Statistics is the science of data sampling and inference. Straightly, big data science is considered as the extension of statistics, termed big data statistics. Through market investigation, big data analysis focuses on statistics and machine learning. Statistics is considered as the optimal decision about sample estimation of population in the asymptotic theory. Machine learning may be statistical learning or function approximation in specific domain criterion.

In probability theory and statistics, a sequence or collection of random variables are independent and identically distributed only if each random variable has the same probability distribution as the others and all are mutually independent. It is noteworthy that an IID sequence does not imply the probabilities for all elements of the sample space or event space must be the same. The IID assumption of random variables, assigned observations, tends to simplify the underlying mathematics of many statistical inferences. In particular, the IID assumption is important in the central limit theorem, which states that the probability distribution of the sum (average) of IID variables with finite variance approaches a normal distribution. Normal distribution, also termed Gaussian distribution, is an elementary probability distribution which can be extended into the mixed Gaussian distribution and generalized distribution. Under the assumption of Gaussian distribution and linear models, least square estimate is equal to maximum
likelihood estimate. Gaussian distribution function in mathematical analysis is correspondingly related to linear model in algebra. To solve complex problems, the mixed Gaussian distribution and generalized distribution in mathematical analysis are correspondingly related to non-linear models in algebra. For example, power law distribution, one of exponential distributions, is widely used for modeling heavy-tailed distributed human behaviors from social media.

However, the assumption of completely independent and identical distribution may not be realistic in big data analysis and geographical data analysis. The law of universal gravitation, given by Newton in 1687, states that any two bodies in the universe attract each other with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Accordingly, the first law of geography, given by Tobler (1970), states that everything is related to everything else, but near things are more related than distant things. The first law of geography implies spatial dependence and spatial autocorrelation, which are utilized specifically by the spatial interpolation of inverse distance weighting and Kriging interpolation in the regionalized variable theory. In nature, spatial statistics, including geostatistics, spatial regression, and random point processes, is the probabilistic realization of the first law of geography. To test the violation of independent assumption, the empirical autocorrelation in a given data-set is evaluated and plotted. To test spatial autocorrelation, Moran’s I is computed and Moran scatter plot is drawn.

2.2. Set theory

It is argued that big data just need to be precisely retrieved rather than trivially analyzed. In relational databases or very large databases, a relation is defined as a set of tuples that have the same attributes or fields. A tuple stands for an object and its information. A domain is a set of possible values for a given attribute, and can be also considered as a constraint on the values of the attribute. Relational algebra or relational calculus is the theory of relational database. Relational algebra is constructed with a set of tuples and five set operations, i.e. [union, intersection, join, projection, selection]. The union operator combines the tuples of two relations and removes all duplicate tuples from the result. The intersection operator produces the set of tuples that two relations share in common. The join operator is the Cartesian product of two relations, which is restricted by some join criteria. In particular, spatial join operators are executed by spatially conditioning on spatial databases (spatial data tables). The projection operation extracts only the specified attributes from tuples. The selection operator retrieves some tuples from a relation or a table, limiting the results to only those that meet a specific criterion, for forming a subset of the set.

Statistics may be the most important mathematics for big data analysis. A probability measure is a real-valued function defined on a set of events in a probability space, and satisfies measure properties such as countable additivity and the value 1 assigning to the entire probability space. A probability triple, $(\Omega, F, P)$, is a mathematical constructor that models a real-world process of randomly occurred states. A probability space is a probability-measured set as follows:

1. A sample space, $\Omega$, which is the set of all possible outcomes.
2. A set of events $F$, where each event is a set containing zero or more outcomes.
3. A probability-measured function $P$ from events to probabilities.

For complex system modeling and big data analysis, a probability space is often characterized with the mixed Gaussian distribution and generalized distribution. In geographical data analysis, the probability distribution function parameterized with spatial covariance is particularly emphasized in spatial statistics.

Machine learning is another common used computing tool for big data analysis. A mathematical space is defined as a set assigned with an added structure. Besides probability space, topological space, Euclidean metric space, norm space, inner space, and vector space are interchangeably used by machine learning. For example, topological space and Euclidean space are used by manifold learning. L0 and L1 norm spaces are used by compressed sensing. Inner space is used by kernel learning. Vector space is used in principle component analysis.

3. Six techniques of big data analysis

Raw data or unprocessed data is a set of numbers or characters. Data is sampled and analyzed for extracting decision-supported information, while knowledge is derived from a large amount of experience on a subject. Physical experience occurs whenever an object or environment changes. Social experience, forming norms, customs, values, traditions, social roles, symbols and languages, provides individuals with the skills and habits necessary for participating within their own societies. In philosophy, phenomena come to mean that what is experienced is the basis of reality. Phenomena are often understood as things that appear or experience for human being. In physics, a phenomenon may be an observed feature and event of matter, energy, and the motion of objects. Data can be considered as observation or experience about real phenomena. Accordingly, geographical data are considered as observation about phenomena of physical and human geography. Basically, geographical big data analysis is aimed at exploring the complexity of geographical reality.

In the sense of data structural storage and structural analysis, the characteristics of big data analysis is derived
from the characteristics of big data. Thus, six techniques of big data analytics are proposed in Figure 1.

3.1. Ensemble analysis

Ensemble data analysis, roughly termed multi-data-set analysis or multi-algorithm analysis, is made for the whole data-set or a large volume of data. Big data are argued to be the whole data-set without any sampling purpose. What is the whole set? Approximately, it may be resampling data, labeled data and unlabeled data, prior data and posterior data. It is known that the term “ensemble” appears at least in the context of ensemble learning in machine learning, the ensemble system in statistics mechanics, and ensemble Kalman filtering in data assimilation.

In supervised learning and image classification, the whole data are often divided into three parts, ie training data, test data, and validation data. Image classifier is trained by training data, and is applied to test data. Finally, image classification results are validated by ground truth. Generally, the classifier achieves a good performance on training data rather than test data. Toward a tradeoff between model’s complexity and classification performance on training data, ensemble analysis is made to achieve the global optimal performance on the whole set of training data, test data, and validation data. In support vector machine (SVM) learning, structural risk minimization is used to overcome the issue of over fitting, ie the learning model strongly tailored to the particularities of training data but poorly applied to test data.

To extract information from a data subset, an algorithm is designed. To extract information from the whole data-set, many algorithms are integrated. Ensemble analysis is concerned with bootstrapping, boosting, bagging, stacking, and random forest learning. In most cases, integrated multiple learning algorithms can achieve better estimation performance when compared to any of the constituent learning algorithms. However, it is not easy to clarify the consistency among constituent algorithms and determine the weight of each constituent algorithm. Inspired by Markov chain Monte Carlo (MCMC) statistics computing, Efron’s bootstrap resampling (Efron 1979) and Valiant’s probably approximately correct (PAC) learning (Valiant 1984) may become two potential theories of ensemble analysis. PAC is the formalism for deciding how much data you need to collect in order for a given classifier to achieve a given probability of correct predictions on a given fraction of future test data. In short, PAC is a general framework of building a hypothesis (a generalization function) with high probability that was approximately correct or the selected function will have low generalization error. PAC is part of a field called computational learning theory constituted of Bayesian learning, minimum description length learning, the Vapnik-Chervonenkis theory of learning (Vapnik 2000), and classical frequentist statistics. Here, machine learning can be considered as statistics estimation and function approximation. More basically, the Weierstrass approximation theorem states that every continuous function defined on a closed interval \([a, b]\) can be uniformly approximated as closely as desired by a polynomial function.

3.2. Association analysis

Usually, big data are collected without special sampling strategies. Normally, data producers are quite different from data users, so that the cause-effect relation hidden in observation data is not clear to specific data users. Set theory, ie the theory about members and their relations in a set, is general enough to deal with data analysis and problem solving. In a sense, the relation among set members corresponds to the association in big data. Association analysis is critical to multi-sourcing, multi-type, and multi-domain data analysis. Typically, association analysis is exemplified with association rule algorithms in data mining (Agrawal and Srikant 1994), data association in target tracking, and links analysis in networks.

The Apriori algorithm of association rules mining was first proposed by Agarwal and Srikant. An association rule has two parts, an antecedent (if) and a consequent (then). Association rules are created by analyzing data to identify frequent if/then patterns and using the criteria
support and confidence to identify the significance of relationships. Support is an indicator of how frequently the items appear in the transaction database. Confidence is an indicator of how frequently the if/then statements have been found to be true. Association rules mining is an instance of estimating conditional probability \( P(\text{consequent} | \text{antecedent}) \) with transaction data. However, unlike the asymptotic analysis of statistics estimators, the effectiveness of mined association rules just depends on the thresholds of support and confidence, which are manually given by users.

In target tracking, usually multiple sensors are used and multiple targets are moving in the complex environment. To identify and track targets, it is necessary to analyze three associations of measurement (observation) data association, target state association, and target-to-measurement association. Enabling the tracking system to perform properly, the most likely potential target location measurement is used to update the target's state estimator. The probability of the given measurement being correct is a distance function from the predicted state to the measured state. Kalman filtering is a time-varying recursive estimator or a time sequence filtering, which is suitable for real-time data stream processing. Kalman filtering, also known as linear quadratic estimation or trajectory optimization, is an algorithm of using a sequence of noise-disturbed measurements (observations) over time to produce the estimates of unknown states (variables) that tend to be more precise than those based on a single measurement. State error covariance and state forecast operator are considered as state associations stochastically and deterministically. Likewise, measurement error covariance and observation operator are considered as measurement associations stochastically and deterministically. In a sense, Kalman filtering is characterized with data association and the velocity of big data.

In social media, social persons and social relations constitute a social network. In natural language or mental models, concepts and conceptual relations constitute a semantic network. Social relations and conceptual relations are usually formalized by the links of networks, also termed linked data. In nature, the network is a topological model. The link of networks describes the neighborhood or adjacency, which could be also termed association. As stated in Section 2.2, big data is modeled at a few levels of mathematical spaces, typically, metric space and topological space. Network is one form of topological space, and network link is the qualitative representation of nearness in topological spaces.

Fundamentally, the association in big data is considered as the relation in set theory. When metrics assigned, associations are formalized as advanced relations. For example, statistical correlation and geometric relation are associations assigned subsequently with probability metric and geometric metric.

### 3.3. High-dimensional analysis

Big data implies a high variety of data. In mathematics, the dimension of a mathematical space (object) is informally defined as the minimum number of coordinates needed to specify any point within it, and the dimension of a vector space is the number of vectors in any basis for the space or the number of coordinates necessary to specify any vector. The dimension of an object is an intrinsic property independent of the space in which the object is embedded. In reality, the dimension is the number of perspectives from which the real world is recognized. Classical physics describe the real world in three dimensions. Starting from a particular point in space, the basic directions in which we can move are up/down, left/right, and forward/backward. The equations of classical mechanics are symmetric with respect to time. In thermodynamics, time is directional with a reference to the second law of thermodynamics, which states that an isolated system's entropy never decreases. Such a system spontaneously evolves toward thermodynamic equilibrium of the states with maximum entropy. In statistics and econometrics, there exists the multi-dimensional analysis for a sample (realization) of multiple random variables. High-dimensional statistics studies data whose dimension is larger than those dimensions in classical multi-variate analysis. In many applications, even the dimension of data vectors may be larger than sample size.

Usually, the curse of dimensionality emerges in big data analysis. Once the dimensionality increases, the volume of the space increases so fast that available data become sparse. While handling big data traditional algorithms (Marimont and Shapiro 1979) this sparsity will lead to the statistical significance error, and the dissimilarity between objects in high-dimensional space. This phenomenon happens only in high-dimensional space rather than in low-dimensional space, and this is the so-called curse of dimensionality.

To keep the variability of original variables as much as possible, using the metrics of distance or variance, dimension reduction is the process of reducing the number of random variables and finding a subset of the original variables (termed subspace). There exist linear and non-linear transforms of dimension reduction. For example, principal component analysis is a linear transform of dimension reduction. In manifold learning, Isomap is a non-linear transform of dimension reduction preserving the global distance of Geodesic, and locally linear embedding is a non-linear transform of dimension reduction preserving local Euclidean distance. Apart from standard Euclidean distance, there exist distance metrics of Mahalanobis distance, Kullback-Leibler divergence, and Wasserstein distance in statistics. Dimension reduction can be used for simplifying entity features and graph structures.
In compressive sensing, the sparsity is formalized with a regular term adding to the cost function. Usually, this regular term is the count of non-zero values, denoted zero norm or L0 function. Furthermore, to facilitate the numeric computation of cost function maximization, L0 function is approximated by L1 function for convex optimization.

3.4. Deep analysis

Nowadays, the volume of emerging data is large enough for complex artificial neural networks training. In the meantime, the high-performance computing technologies of multi-cores of CPU, GPU, and FPGA greatly reduce the training time of complex artificial neural networks. Under such circumstances, traditional artificial neural networks are functionally enhanced with hidden layers of latent variables, and so-called deep learning is developed as compared to shallow learning. It is also believed that human cognition about the real world is getting deeper as our experience is getting rich. Deep analysis is potential in exploring complex structure properties of big data. Unobservable variables, hidden parameters, hidden hierarchies, local correlations, and the complex distribution of random variables can be found through deep analysis. Typically, there exist the models of latent Dirichlet allocation, deep Boltzmann machine, deep belief net, hidden Markov model, and Gaussian mixture model.

Deep learning architectures are built from multi-layered artificial neural networks (LeCun, Bengio, and Hinton 2015). The general artificial neural network has numerous non-linear processing units of neurons (activation units), such as the sigmoid function. Theoretically, a number of neurons of each layer are considered as multiple variables. Deep analysis is considered as a deterministic (or stochastic) transform function from input to output. Numerical approximation models, data-fitting functions, and maximum likelihood estimators may become the potential theories of building multi-layered artificial neural networks. Equipped with multiple layers, multiple variables, and non-linear activation units, the multi-layered artificial neural network is powerful in complex data modeling. However, until now, no universal theories of approximation are available for deep learning of any complex function. In practice, it is not possible to compare the performance of multiple architectures all together which are not all evaluated on the same data-sets.

In multi-variate multi-layered neural networks fully connected or sparsely connected, there exist a maximum solution to the data fitting function or statistical estimator, this will raise the issue of ill-posed problems and poor computation performance. Usually, the complex model leads to over fitting, and to spend expensive computation time. Thus, regularization methods are developed to overcome over fitting issues. And some numerical algorithms, such as pre-training, concave-to-convex approximation, computing the gradient on several training data-sets at once, are developed to increase the efficiency.

3.5. Precision analysis

In numerical analysis, accuracy is the nearness of a calculation to the true value, while precision is the resolution of the representation, often defined by the number of decimal or binary digits. Statistics prefers to use the bias and variability instead of the accuracy and precision. Bias is the amount of inaccuracy, and variability is the amount of imprecision. In practice, trueness is the closeness of the mean of a set of measurement results to the actual (true) value, and precision is the closeness of agreement among a set of results. Ideally, a measurement device is both accurate and precise, with measurements all close to and tightly clustered around the true value. A measurement system can be accurate but not precise, precise but not accurate, neither, or both. When an experiment contains a systematic error, increasing the sample size generally increases precision but does not improve accuracy. Eliminating the systematic error improves accuracy but does not change precision.

Precision analysis is used to evaluate the veracity of data from the perspective of data utility and data quality. To make an analogy between big data and linguistics, the veracity is analogous to the semantics, and the utility is analogous to the pragmatics. In a broad sense, the uncertainty in the human-computing machine-earth system is studied in Figure 2 (Shu et al. 2003). Big data are somewhat collected from the system of human-machine (computer)-environment (things). It is realized that user preferences also are one special kind of data. Big data analysis may need user data for constructing the integral model and optimization criteria. Precision analysis frequently appears in marketing, medicine, and agriculture, short for precision marketing, precision medicine, and precision agriculture. To build the models of user preferences or user requirements, users are distinguished with each other, and even users are somehow linked to be a social network.

Marketers solicit personal preferences directly from recipients. Personalized marketing, termed personalization or one-to-one marketing, refers to marketing strategies applied directly to a specific consumer. Personalization tries to make a unique product offering for each customer. In 2015, US President Barack Obama stated his intention to fund a United States national “precision medicine initiative” (Lauran 2015). To change the situation of disease treatment and prevention strategies developed for the average person, precision medicine refers to the tailoring of medical treatment to the
3.6. Divide-and-conquer analysis

Divide-and-conquer is a general computational strategy for improving the efficiency of problem-solving and the velocity of big data computation. Through divide-and-conquer analysis, a problem is recursively broken down into two or more sub-problems in the stage of dividing, until these become simple enough to be solved directly in the stage of conquering. Upon completion, the solutions to sub-problems are combined into a solution to the original problem. In a sense, distributed computing, such as cloud computing and distributed intelligence computing, may be considered as the computation of divide-and-conquer in space, and parallel computing (multi-core computing and cluster computing) may be considered as the computation of divide-and-conquer in time. Steam processing and real-time computing somewhat are the computation of divide-and-conquer from the perspectives of unstructured data and time-constraints. Cloud computing and distributed intelligence computing somewhat are the computation of divide-and-conquer from the perspectives of virtualized resources and the society of mind.

High-performance parallel computation in a shared-memory multi-processor uses parallel algorithms while the coordination of a large-scale distributed system uses distributed algorithms. Divide-and-conquer analysis is easily executed in multi-processor machines, and distinct sub-problems can be executed on different processors. Information is exchanged by passing messages between the processors. Parallel computing uses multiple processing elements simultaneously to solve a problem. This is accomplished by breaking the problem into independent parts so that each processing element can execute its part of the algorithm simultaneously with the others. The processing elements can be diverse and include resources such as a single computer with multiple processors and several networked computers. Since frequency scaling hit a wall, multi-core systems have become more widespread, and parallel algorithms are popularly used. A multi-core processor is a processor that includes multiple processing units (cores) on the same chip. Main memory in a parallel computer is either shared memory, that is, memory shared between all processing elements in a single address space, or distributed memory in which each processing element has its own local address space.

Parallel computing is a type of computation in which many calculations are carried out simultaneously, in the principle that large problems can often be divided into smaller ones, which are then solved at the same time. Parallel computing can be divided into a few levels of bit-, instruction-, data-, and task-parallelism. Task parallelism involves the decomposition of a task into sub-tasks and then allocating each sub-task to a processor for execution. The processors would then execute these sub-tasks simultaneously and often cooperatively. In programming languages, OpenMP are two of most widely used shared memory APIs (Application Programming Interfaces), whereas message passing interface is the
most widely used message-passing system API. The MapReduce in the computers cluster is a parallel computing programming paradigm. The Map and Reduce functions of the MapReduce are both defined with respect to data structured in (key, value) pairs. Provided that each mapping operation is independent of the others, all maps can be performed in parallel. All outputs of the map operation that share the same key are presented to the same reducer at the same time. Then, the Reduce function is applied in parallel to each group. However, as compared to relational databases, the MapReduce is lack of schema of structural data and quick assess techniques of B-trees and hash partitioning, and it is mostly applied to simple or one-time processing tasks.

Real-time computing puts emphasis on the deadline, time constraints, and time validity of data analysis. A real-time database is a database system which uses real-time processing to handle workloads whose state is constantly changing. This differs from traditional databases of persistent data, mostly unaffected by time. A stream can be a sequence of data or a series of operations. Streaming data is often considered as non-structural data, and continuous data of infinite size. In contrast to archived data, stream data is usually real-time read-only sequence. The time window of data analysis for streaming data segment, incremental update, and Kalman filtering has implications for the data sampling rate and inherent temporal correlation of windowed-data.

In cluster computing, all nodes are on the same local network, and use similar hardware. A computing cluster is a group of loosely coupled computers that work together closely, so that in some respects they can be considered as a single computer. Clusters are composed of multiple standalone machines connected by a network. In grid computing, nodes are shared across geographically and administratively distributed systems, and use more heterogeneous hardware. Grid computing is the most distributed form of parallel computing, and it makes use of computers communicating over the Internet to work on a given problem.

In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other by message passing, and each processor has its own private memory (distributed memory).

The computational entities are called computers or nodes. A distributed database system consists of loosely coupled sites that share no physical components. The distribution is transparent, that is, users are able to interact with the system as if it were one logical system. Besides, transactions are transparent, that is, each transaction must maintain database integrity across multiple databases.

For enabling ubiquitous, on-demand access to a shared pool of configurable computing resources, cloud computing, also known as on-demand computing, is a kind of Internet-based computing that provides shared processing resources and data to computers and other devices on-demand. Cloud computing provides the tools and technologies to build data/computation intensive parallel applications with much more affordable prices compared to traditional parallel computing techniques. The main enabling technology for cloud computing is virtualization. Virtualization software separates a physical computing device into one or more “virtual” devices, each of which can be easily used and managed to perform computing tasks. Cloud computing adopts concepts from service-oriented architecture (SOA) that can help the user break these problems into services that can be integrated to provide a solution. Cloud-computing providers offer their “services” according to different models, which happen to form a stack, ie infrastructure-, platform- and software-as-a-service. In its own right, there exist private, community, and public cloud services.

In distributed intelligence computing, the human-machine-environment interaction system can be roughly conceptualized to be a system of agents. In Simon’s theory of bounded rationality, when individuals make decisions (Simon 1982), their rationality is limited by the available information, the tractability of the decision problem, the cognitive limitations of their minds, and the time available to make the decision. A rational agent is an agent that has clear preferences, models uncertainty via expected values of variables, or functions of variables, and always chooses to perform the action with the optimal expected outcome for itself from among all feasible actions. A rational agent can be anything that makes decisions, typically a person, machine, or software. An intelligent agent is an autonomous entity which observes through sensors and acts upon an environment using actuators and directs its activity toward achieving goals. Intelligent agents may also learn or use knowledge to achieve their goals. Furthermore, Marvin Minsky constructs a model of human intelligence from the interactions of agents (Minsky 1986), termed the society of mind.

4. Geographical big data analysis

For geographical big data, six techniques in big data analytics are still usable. Compared to general big data, the special thing of geographical big data is spatiotemporal association, eg geometrical relations, statistical correlations, and semantics relations. Spatiotemporal data analysis can be roughly categorized into geometrical measurement analysis, human geographical data analysis, and physical geographical analysis. Three kinds of analysis partly imply the analytics of the digital earth, smart earth, and earth simulator.

With the development and application of computer and communication technologies, the human-machine-environment system is increasingly observed in digital form. A large volume of geographical data,
such as earth observation satellite images and mobile Internet data, are collected collectively or individually. In most cases, geographical big data are nearly the whole data-set regardless of application domains and sampling strategies. Ensemble data analysis is widely applied to geographical big data.

Basically, spatiotemporal relations are categorized into geometrical relations, statistical correlations, and semantic relations. In mathematics, space and time can be formalized into geometric quantities. For example, measurement adjustment in surveying is least square estimation of geometrical feature (parameters) of angles, edges, and elevations. Time is scaled by the motion frequency of objects, like timing objects of celestial bodies and atom clocks. Nowadays, using various surveying instruments, phenomena of physical and human geography are observed for a large amount of georeferenced data. In the digital earth, topographic observation and geographical phenomena sensing are digitally recorded in the computer, and geometrical relation analysis is emphasized in geographical big data analytics.

With the advent of mobile devices and the Internet, the trajectories of mobile objects are easily captured. The trajectory is digitally represented by a sequence of space-time coordinate points, \((x, y, t)\). For human beings, spatial behaviors, individual behaviors or group activities, can be somewhat explored by trajectories. Normally, time geography and spatial cognition are considered as the scientific foundation of human geographical data analysis. In time geography, the elements of time geography are geometrically figured by the prism (Hägerstrand 1970). Extending the elementary prism, probabilistic and multi-scale trajectories are modeled (Miller 1991). In spatial cognition, a mental map is constituted of landmarks (nodes), paths, and blocks, and the salience of cognitive objects is emphasized (Golledge and Stimson 1997). In artificial intelligence, general problem solving can be considered as wayfinding, in which states (original state, middle states, final state) are metaphorized as nodes (starting node, middle nodes, target node) and state links metaphorized as paths (Newell, Shaw, and Simon 1959). As stated in Section 2.1, spatial or spatiotemporal relations exploring is central to geographical big data analysis. A remotely sensed image is considered as a sample of the vector random field, in which each pixel and each band are considered as random variables. This will give arise to the high-dimensional analysis of multi-spectral and high-resolution images. Again, typically crowd-sourcing data is a kind of unknown sampling data. In the smart earth, spatiotemporal statistics and spatial cognition become the potential theories of spatiotemporal data analysis.

On the Earth’s surface, physical entities are varying over space and time. The mass and energy of physical entities are governed by the law of conservation and conversion, which is usually modeled by partial differential equations (PDE). In the earth system or land surface system, the trajectory can be modeled by a non-linear function of space and time. Solving PDE of space-time functions numerically, a sequence of space-time varying states is obtained. Particularly, the partial differential equation for the dynamical geographical system can be numerically solved by the two dimensional cellular automaton (CA) of large neighborhoods and the large number of states. A CA consists of a regular grid of cells, and each cell has one of a finite number of states. The grid can be in any finite number of dimensions. A set of cells around each cell, called its neighborhood, is defined with regard to the specified cell. An initial state is selected by assigning a state to each cell. A new generation is created advancing \( t \) by 1, according to some fixed rules (generally, a mathematical function) that determine the new state of each cell in terms of the current state of the cell and the states of the cells in its neighborhood.

A large volume of observations are collected by the sensors on the platforms of satellite, weather stations, ships, buoys, and others. In the meantime, the mathematical model for the land surface system normally consists of numerous PDE and physical variables. To solve these physical equations efficiently, the model reduction and dimension reduction are needed. The initial perturbation is easy to make butter-fly effects in non-linear PDE forecasting. Due to multi-scale observations and physical processes, the multi-resolution analysis is often needed. In numerical computation of physical equations, sub-grid computing of sub-processes is needed. Data assimilation is the process by which observations of the physical system are subsequently incorporated into the forecast states of a numerical model. In Bayesian statistics, the analysis is an application of Bayes’ theorem, and the assimilation is recursive Bayesian estimation. A typical cost function, to be minimized for creating the analysis, would be the sum of the squared deviations of the analysis values from the observations weighted by the accuracy of the observations, plus the sum of the squared deviations of the forecast states from the analysis values weighted by the accuracy of the forecast. Data assimilation of combining various earth observations and the earth system models is aimed at simulation results at high-precision and high-resolution of space and time. In the earth simulator, the earth system of atmosphere, ocean, land surface, geology, and socioeconomics are simulated with supercomputers.

5. Conclusions

With the rapid development of computer and communication technologies, the human-machine-environment system is increasingly observed by the space-, air- and ground-based sensor digital networks. Starting from the 4V characteristics of big data (volume, variety, velocity, and veracity), six techniques in big data analytics are proposed here. Roughly, ensemble analysis and
association analysis are applied to the volume of big data. High-dimensional analysis is applied to the variety of big data. Deep analysis and precision analysis are applied to the veracity of big data. Divide-and-conquer analysis is applied to the velocity of data. Fundamentally, two theoretical challenges of the violation of independent and identical distribution and the extension of general set-theory are addressed.

For geographical big data, we have illustrated the associations of geometrical associations in space and time, spatiotemporal correlations in statistics, and space-time relations in semantics. Furthermore, we have illustrated spatiotemporal data analysis of measurement (observation) adjustment of geometrical quantities, human spatial behavior analysis with trajectories, and data assimilation of physical models and observation data.

Six techniques in big data analytics are general for any data analysis, but spatiotemporal associations analysis is special for geographical data. Just like the four characteristics of big data given by the industrial community, six techniques in big data analytics are given in experience, and six techniques in big data analytics are not mutually exclusive and collectively exhaustive in theory.

Funding

This study is supported jointly by the Fundamental Research Funds for the Central Universities, the Key Project of National Natural Science Foundation of China (grant number 41331175), and the LIESMARS Special Research Funding.

Notes on contributor

Hong Shu is Professor of Geoinformatics in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) at Wuhan University, China. Since 1994, he has been working on geographical spatiotemporal issues in a comprehensive way of algebra, geometry, deterministic and stochastic analysis, mathematical logic, particularly including spatiotemporal statistics, computational geometry, scale analysis in space and time, data assimilation in statistics computing of non-linear PDE, space-time common-sense reasoning in logic. He is the principal investigator of several major research projects funded by the National High-Tech Research and Development Program of China and National Natural Science Foundation of China, and the author of approximately 100 scientific papers on spatiotemporal information and analysis.

References

Agrawal, R., and R. Srikant. 1994. “Fast Algorithms for Mining Association Rules in Large Databases.” In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 487–499. San Francisco, CA: Morgan Kaufmann Publisher.

Barwick, H. 2012. “The ‘Four Vs’ of Big Data.” Implementing Information Infrastructure Symposium, Computerworld Australia. http://www.computerworld.com.au/article/396198/iis_four_vs_big_data/. Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife.” The Annals of Statistics 7 (1): 1–26. http://projecteuclid.org/download/pdf_1/euclid.aos/1176344552. Grolle, R. G., and R. Stimson. 1997. Spatial Behavior: A Geographic Perspective. New York: Guilford Press. Hägerstrand, T. 1970. “What about People in Regional Science?” Papers of the Regional Science Association 24 (1): 7–24. doi:http://dx.doi.org/10.1111/j.1435-5597.1970.tb01464.x. Hilbert, M. 2015. “Big Data for Development: A Review of Promises and Challenges.” Development Policy Review 34 (1): 135–174. Laurant, N. 2015. Obama Proposes ‘Precision Medicine’ to End One-Size-Fits-All. New York: Associated Press. http://www.dailynews.com/general-news/20150130/obama-proposes-precision-medicine-to-end-one-size-fits-all. LeCun, Y., Y. Bengio, and G. Hinton. 2015. “Deep Learning.” Nature 521 (7553): 436–444. doi:http://dx.doi.org/10.1038/nature14539. Li, D. R., Y. Yao, and Z. F. Shao. 2014. “Big Data in the Smart City.” Geomatics and Information Science of Wuhan University 39 (6): 630–640. doi://dx.doi.org/10.13203/j.whugs20140135. Marimont, R. B., and M. B. Shapiro. 1979. “Nearest Neighbour Searches and the Curse of Dimensionality.” IMA Journal of Applied Mathematics 24 (1): 59–70. doi://dx.doi.org/10.1093/imamat/24.1.59. Miller, H. J. 1991. “Modelling Accessibility Using Space-Time Prism Concepts within Geographical Information Systems.” International Journal of Geographical Information Systems 5 (3): 287–301. Minsky, M. 1986. The Society of Mind. New York: Simon & Schuster. Newell, A., J. C. Shaw, and H. A. Simon. 1959. “Report on a General Problem-Solving Program.” Communications of the ACM 2 (7): 256–264. Rao, C. R. 1997. Statistics and Truth: Putting Chance to Work. 2nd ed. Singapore: World Scientific. Shaw, S. L., and Z. X. Fang. 2014. “Rethinking Human Behavior Research from the Perspective of Space-Time GIS. (in Chinese).” Geomatics and Information Science of Wuhan University 39 (6): 667–670. doi://dx.doi.org/10.13203/j.whugs20140127. Shu, H., S. Spaccapietra, C. Parent, and D. Q. Sedas. 2003. “Uncertainty of Geographic Information and Its Support in MADS.” In Proceedings of the 2nd International Symposium on Spatial Data Quality, 1–13. Hong Kong: Hong Kong Polytechnic University, March 19–20. Simon, H. A. 1982. Models of Bounded Rationality. Cambridge: MIT Press. Tobler, W. R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” Economic Geography 46 (2): 234–240. doi://dx.doi.org/10.2307/1433141. Valiant, L. 1984. “A Theory of the Learnable.” Communications of the ACM 27 (11): 1134–1142. doi://dx.doi.org/10.1145/1968.1972. Vapnik, V. 2000. The Nature of Statistical Learning Theory. Berlin: Springer.