Int. J. Mach. Learn. & Cyber. (2018) 9:133–144
https://doi.org/10.1007/s13042-015-0376-z

ORIGINAL ARTICLE

Universal knowledge discovery from big data using combined dual-cycle

Bin Shen[1]

Received: 27 November 2014 / Accepted: 9 May 2015 / Published online: 23 May 2015

Abstract Many people hold a vision that big data will provide big insights and have a big impact in the future. However, how to turn big data into deep insights with tremendous value still remains obscure. Here I highlight universal knowledge discovery from big data. The new concept focuses on discovering universal knowledge, which exists in the statistical analyses of big data and provides valuable insights into big data. Universal knowledge comes in different forms, e.g., universal patterns, rules, correlations, models and mechanisms. To accelerate big data assisted scientific discovery, a unified research paradigm should be built based on techniques and paradigms from related research domains, especially big data mining and complex systems science. Therefore, I propose a dual-cycle methodology with three types of cycle-driven UKD process, i.e., big-data-cycle-driven, mechanism-cycle-driven and combined-dual-cycle-driven mining. A case study is also given to illustrate the effectiveness of the proposed processes. This paper lays a foundation for the future development of universal knowledge discovery, and offers a pathway to the discovery of "treasure-trove" hidden in big data.

Keywords Universal knowledge discovery · Big data · Combined-dual-cycle-driven mining

✉ Bin Shen
tsingbin@zju.edu.cn

[1] Ningbo Institute of Technology, Zhejiang University, 315100 Ningbo, China

1 Introduction

In the era of big data, data mining (DM), which is the analysis step of knowledge discovery in databases (KDD), has become a key means of knowledge acquisition from data. Following the classical knowledge discovery process (which sequentially involves the steps of data selection, preprocessing, subsampling, transformations, data analysis, post-processing and knowledge utilization) [17], data mining technology has achieved great success in the past ten years [27] in many areas, such as business, military, bioinformatics, medicine and education. Using advanced data mining methods and algorithms (e.g., pattern mining, clustering and classification methods), many previously unknown but interesting knowledge is discovered from various types of data (e.g., texts, web data, graphs and data streams) [2, 10, 47].

However, the following deficiencies arise when applying traditional data mining methods. Firstly, a large amount of knowledge can be easily produced by traditional data mining approaches, but few insights are obtained. The reason for this embarrassing situation is as below. Although traditional data mining technology specializes in discovering interesting patterns from data, it is not good at discovering internal mechanisms behind the data. Secondly, the knowledge discovered by traditional data mining methods is temporary and dynamic in nature, and its validity always changes over time. Due to its temporary nature, dynamic knowledge has to been mined and updated continuously [55]. Thus, the value of dynamic knowledge is limited, because it may be not capable for further use. Universal knowledge (defined below), which has a certain degree of universality, has not yet attracted enough attention in the field of data mining. However, universal knowledge is valuable, because it is capable for providing

insights for people. Thirdly, the discovered knowledge cannot sufficiently support meaningful decision-making actions, and there is a significant gap between mining results and real-world application requirements [10].

To overcome these limitations, I advocate that in the era of big data, the next generation of data mining technology should focus more on mining universal knowledge, which exists in the statistical analyses of big data. Generally speaking, in the history of our information age, humans have experienced two stages of data based information acquisition: the first one is using data collection and retrieval techniques to find desired information, and the second is using data mining technology to produce dynamic knowledge. I argue that the next stage will be "wisdom discovery", which adopts universal knowledge discovery technology to mine global and valuable universal knowledge (which can be easily transformed into "wisdom" in the sense of human cognition and decision making) from big data.

So, towards the vision of "wisdom discovery", a new concept called Universal Knowledge Discovery from big data (UKD for short) is proposed. In order to build a solid theory foundation for the future development of UKD, the concept of UKD is defined, and various categories of universal knowledge are described. To accelerate the development of UKD, a unified research paradigm should be built based on techniques and paradigms from related research domains, which include not only data mining, but also machine discovery [28, 29, 38, 52, 60, 61], big data analytics [16, 31, 43, 56, 58, 59] and complex systems science [8, 22, 41, 42, 51, 53]. To meet this need, a novel research paradigm called dual-cycle methodology is proposed. In this methodology, three types of discovery process are also proposed: (1) big-data-cycle-driven mining, (2) mechanism-cycle-driven mining and (3) combined-dual-cycle-driven mining.

Based on the new concept (i.e., UKD) and its methodology, this paper also addresses how to gain insights into big-data, which is a fundamental issue and urgent problem for all scientific disciplines [25, 30, 44]. Overall, this work serves three purposes. First, UKD is highlighted as a pathway to the discovery of "treasure-trove" hidden in big data. Second, I advocate that an interdisciplinary and unified research paradigm for UKD should be built. Therefore, a methodology with combined-dual-cycle-driven UKD process is proposed. Last, this research lays a solid foundation for the future development of UKD technology.

2 Technical origins

To integrate methods in related disciplines to build advanced UKD technology, we need to sort out paradigms and techniques adopted in related fields. Therefore, in this section, I review technical origins of universal knowledge mining, which include machine discovery, big data analytics, complex systems science, etc.

2.1 From machine discovery to data mining, and then big data mining

As early as the 1970s, Langley [28, 29] had developed six versions of the BACON system one after another, which are able to automatically rediscover some physical and chemical laws using heuristic searching methods. Soon afterwards, based on inductive learning and logical reasoning, a number of systems, such as STAHL [60], FAHRENHEIT [61] and IDS [38], were developed in this field. In China, Chinese scholars, such as Wu [52], have made fruitful achievements in mechanical theorem proving in geometries. Thus, in the 1980s, the research direction of machine discovery found focus. However, these studies have some limitations. Firstly, the algorithms can only deal with pure and small data as the input and need to be accompanied by a certain searching direction; otherwise the algorithms are infeasible due to huge searching space. Secondly, the systems only rediscovered some laws already known, and few important and new scientific laws have been discovered by such systems.

Since the mid-1990s, data mining research has avoided the limitations of machine discovery by discovering meaningful patterns from the statistical perspective instead of scientific laws. Thus, a large number of effective data mining algorithms have been developed, and data mining which integrates related disciplines (e.g., statistics and machine learning) has achieved great success in practice.

The advent of the era of big data presents a new opportunity for the future development of data mining technology. Currently, research in big data mining (or big data analytics) commonly focuses on developing advanced technology to deal with the technical aspects of big data, e.g., large volume [31, 43], high velocity, and/or high variety [26]. These studies mainly are concerned with the performance, the scalability and the scope of proposed methods, but neglect fundamental changes in thinking brought about by the shift from processing small data to tackling big data [36]. Most data mining researchers have not been aware of the potential great changes in research paradigm and mining methods as they shift from "knowledge discovery from small data" to "universal knowledge discovery from big data".

For instance, McKinsey Global Institute [34] pointed out that big data is the next frontier for innovation, competition and productivity. However, the methods listed in the report are still traditional ones. Cohen et al. [15] highlighted Magnetic, Agile and Deep (MAD) skills for big data analysis, which depart from traditional enterprise data

warehouses. Zikopoulos et al. [59] introduced popular Hadoop distributed processing platform for big data analytics. Besides, almost all of the ten articles [12, 32] presented at the sessions of "Scaling-up methods for big data" and "Statistical techniques for big data" of SIGKDD 2014 focus on high performance processing of big data. For example, Chen et al. [12] proposed a fast flux discriminant method for large-scale nonlinear classification. In the survey paper of big data mining [16], Fan et al. mainly discussed the computability in big data processing and how to process various types of big data.

Although the treatability of big data is an important aspect of big data management and mining, it is not enough to discover "treasure-troves" hidden in big data. Obviously, compared with "small" data mining, big data mining does not merely mean that the amount of data that can be processed increases, but also, more importantly, means a general shift in thinking (e.g., focusing on data externality and highlighting correlations [36]) as well as a new data mining research paradigm. In this way, we can not only process big data as easily as small data, but also gain insights cleverly based on externalities of big data [57]. I consider that the following works are representative of big data analytics.

For instance, during the 2014 Chinese Spring Festival, based on big data of more than 200 million smartphones, China Central Television exhibited the macroscopic statistical regularities of population movements within China during this period. Based on such knowledge, related public agencies are able to efficiently schedule resources of passenger transportation, and individuals can better arrange their travel plans [48]. This work takes full advantage of the externalities of smartphone trajectory data. In addition, Zhu et al. [58] analyzed the big data of Chinese food recipes, and found that the geographical proximity rather than climate similarity is a crucial factor determining the similarity of regional cuisines. Here, the data analyzed is the entire dataset of Chinese recipes, rather than sampling data. Furthermore, if big data of recipes from all countries around the world were analyzed, the conclusion would be more persuasive. Zhang et al. [56] used trajectories from a fleet of GPS-equipped taxicabs to detect gas station visits and measure the time spent in real time, and then estimate overall gas usage of the city. In this case, the system has the capability of real-time scheduling and decision-making based on the comprehensive dynamic perception data of an entire smart city.

The above practical applications of big data are quite good. However, overall, applications of big data analytics are still in the stage of trial and error. An ideal application paradigm of big data has not been formed yet.

2.2 Scientific progress and research paradigm of big data based complex systems science

In China, early in the 1980s, local pioneer scientists Qian [41, 42] and Xu et al. [53] began to advocate the necessity of system science and proposed some important thoughts and ideas. However, without the relevant data, it was difficult to prove these ideas at that time. With the rise of the big data technologies in recent years, especially the techniques of big data collection, storage and analysis, it has been feasible to explore hidden mechanisms and laws behind big data and verify some thoughts based on large datasets, which has led to a new surge of complex systems research.

For example, the well-known small world model [51] and the scale-free network theory [5] are both based on large-scale empirical datasets, e.g., the actor cooperation network and the World Wide Web network datasets. Besides, the research on the inverse problem of complex network (i.e., uncovering the topology of unknown networks according to observation data which reflects network dynamics), such as link prediction [14], are all required to be verified by big data of large scale networks. For another example, by analyzing the flow trajectories of 464,670 dollar bills, Brockmann [8] uncovered the knowledge of dynamic and statistical properties of human travel on a large scale, and discovered the dispersal of bank notes matches a continuous time random walk process incorporating scale free jumps as well as long waiting times between displacements. Likewise, Gonzalez [22], Yan [54] and Peng [39] reached some valuable conclusions by analysing the trajectories of over 100 thousand mobile users, 230 volunteers' 6-week travel diaries and the movement paths of 15.8 thousand Shanghai taxies respectively.

Based on the above cases, we may conclude that complex systems science adopts a different technical paradigm, compared with the data mining research field. Its characteristics are summarized below [45].

1. *A systems theory perspective*
   System science has abandoned reductionism and embraces holism. Complex network, as a rapidly growing field of system science, offers a fresh perspective on the use of networks for system modeling and analytics. On this basis, core issues, such as system dynamics and function emergence, are being researched. Macroscopic phenomena can be reproduced via setting appropriate micro-level simple rules for system dynamics.
2. *The viewpoints of evolution and dynamics*
   System science takes the perspective of evolution and dynamics on systems. It not only discusses the

dynamics of system topology, but also cares about various dynamic phenomena on systems. Taking complex network research as an example, it explores a variety of dynamic mechanisms based on certain network topologies, such as the spread of infectious diseases, cascading failures, network congestion and crowd-powered socially embedded search engine.

3. *Extensive using statistical physics methods*
Statistical physics has played an important role in complex systems research and provided a variety of effective approaches, e.g., the mean-field theory, the percolation theory, the master equation and the Fokker–Planck equation. At the same time, the complexity research also concerns physical processes taking place in systems and the emergence of physical properties of systems, such as Bose–Einstein condensation and random walks on complex networks.
4. *Combination of simulations and statistics on big data*
System science tries to find and validate statistical laws based on big data, and discover the internal mechanisms of systems. Computer simulation is also a frequently used tool in the research of system science. In summary, the research paradigm of complex systems science has apparent advantages in system modeling, dynamics analysis, mechanism exploring, etc. However, there is a bottleneck in big data processing and analytics for system science. Wang et al. [50] considered that systematically using empirical data to explore systems is still in its preliminary stage, and the key issue is the lack of effective theoretical framework for well-targeted data accumulation and analytics. But in fact, there are many advanced methods and valuable experiences in the data mining field for data acquisition, modeling and processing [45]. Therefore, it is necessary to integrate the methods in complex systems science and data mining for solving big-data-assisted discovery problems. A detailed discussion on the integration will be given in Sect. 4.

3 Concepts and categories

3.1 What are universal knowledge and universal knowledge discovery from big data

Universal KNowledge (UKN) in big data is defined as knowledge and laws with a certain degree of universality and immutability in the statistical analyses of big data. In this definition, universality implies that the phenomenon highlighted by the knowledge exists extensively in nature and society, and immutability implies that the phenomenon remains unchanged under certain conditions with the evolution of the universe. With its repeatability and predictability, UKN provides insights into big data. Based on the definition of UKN, the concept of universal knowledge discovery from big data (UKD) is quite straightforward. It is defined as the process of discovering universal knowledge from big data. To understand the concept of UKN and its applications easily, in the following paragraphs, I give three examples to illustrate the current progress towards this direction.

Example 1 Universal knowledge about viral marketing

A marketing manager wants to set a number of individual nodes (i.e., monitoring staff) in a social network to monitor the status of a viral marketing campaign, and forecast whether or not a phenomenon designed by the campaign will "go viral". Chesney [13] discovered some UKN in multiple evolutionary simulations. For example, the number of individual nodes adopting a viral behavior follows an s-curve, and whether the adoption proportion exceeds a critical value or not will determine its prevalence. In addition, a node's predictive accuracy is negatively related to its geodesic distance to the nearest member of the critical mass (a small group of early adopters), and positively related to its centrality. Based on these findings, marketers can take appropriate actions, e.g., selecting individuals with high network centrality which are scattered in different clusters as observation nodes.

Example 2 Universal metaphors and its applications

After analyzing web based big data in various languages, we may find many frequent metaphors, such as the comparison of "father" to "sun". Because such metaphors are frequently used in various languages, they can be called universal metaphors. According to Cancho's work [9], a "small world" network of human language can be built, if any two frequent co-occurring words, which include pairs of tenors and vehicles of universal metaphors, are connected with each other. We also know that human brains use networks of neurons connected by synapses to store the "small world" of human language [40]. Therefore, we can speculate that the neural connections of concepts "father" to "sun" are universal in human brains. Exploring the origin of such universal neural connections is helpful for the in-depth study of human languages and brain science.

Example 3 Universal small world phenomenon and power law distribution

Milgram [37] found the phenomenon of small world from his famous small-world experiment, and Watts [51] discovered the same phenomenon from empirical data of actor collaboration networks, power grids, etc. Now, the

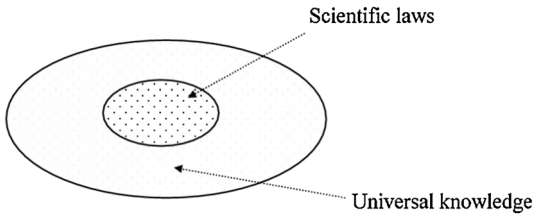

Fig. 1 The difference between scientific laws and universal knowledge

existence of such phenomenon has been confirmed in more and more fields, such as Internet and the spread of diseases. Barabási [5] found that the distribution of node degree of a scale-free network follows a power law. We now know that such power law distribution exists extensively in nature and society. Examples include the Zip's-law-like word frequency distribution, the Pareto distribution of social wealth, the Gutenberg–Richter power law distribution of earthquake sizes, and the intensity distribution of solar flares. It shows that this law is in line with the evolution of the universe. From the above examples, we can find that the discovery of such UKN is meaningful, and exploring the origins of universal phenomenon highlighted by the UKN leads to profound insights.

It is necessary to point out that universal knowledge includes but is not limited to scientific laws (as shown in Fig. 1). That is to say, scientific laws can be regarded as a special kind of universal knowledge. Compared with scientific laws, universal knowledge stands in the statistical analyses of big data and encompasses a much broader scope. Universal knowledge is also very meaningful. For example, using large-scale mobile phone data, Gao et al. [19] suggests that communication traffic surges after disasters and the dominant component of this traffic is between eye-witnesses and their significant friends and relatives. The discovery of such universal knowledge is quite useful for emergency management.

3.2 Categories of universal knowledge

Universal knowledge can be classified into the following non-exhaustive representative categories, which can be extended in accordance with human cognition expansion in the future.

- *Universal knowledge in data distributions* The distribution of big data helps us to characterize big data. Common data distributions include normal/Gaussian distribution, log-normal distribution, exponential distribution, Poisson distribution, power law, power-law with exponential cutoff, shifted power-law, stretched exponential distribution, Weibull distribution, fat-tailed distribution, and so on. For example, Lévy flights are expected in places where prey is scarce, whereas a Brownian strategy is more likely to be adopted where prey is abundant [21].
- *Universal knowledge towards characteristics of systems* Common characteristics of systems include the following: self-similarity and fractal nature, small world, scale free, bursts and so on. For example, Barabási [4] shows that the pattern of bursts appears in a wide range of human daily activities, from web browsing to letter writing, from Wikipedia editing to online trading. For another example, financial markets normally have long-term memory.
- *Universal knowledge about indexes* People may find that there are certain correlations between phenomena and their indicators. These correlations include but are not limited to causal relationships, as discussed in the part of universal correlations below. In order to indicate the fluctuation of a certain phenomenon, we may combine correlated indicators to build indexes. For example, China's "Li Keqiang index" combines electricity consumption, railway freight and lending to measure the country's economic growth.
- *Universal patterns* Many kinds of patterns have been developed in the data mining field, such as frequent co-occurrence patterns and behavior patterns [11]. If these patterns exist in the big data of a system, we can call them universal patterns. For further example, since words "I" and "am" frequently appear in sequence in English, "I am" can be called a universal sequential pattern. For another example, because the comparison of "father" to "sun" is frequently used in different languages, it can be seen as a universal metaphor.
- *Universal rules* Based on big data, we can obtain universal rules, e.g., universal association rules, universal IF-THEN rules and universal intervention rules. For example, by analyzing the birth defects monitoring data of a certain region (e.g., region "A"), we find that the proportion of birth defects of newborns decreases obviously if the intervention of early pregnancy folic acid supplements is adopted. Thus, an intervention rule is obtained as below [49]. "Folic acid (deficiency $\rightarrow$ sufficiency) $\Rightarrow$ Birth defects (Yes $\rightarrow$ No), support $= 0.33$". Then, if the validity of this rule is confirmed by testing in different regions around the world, this rule can be called a universal intervention rule.
- *Universal correlations* Universal correlations exist in the statistical sense of big data. Universal correlations are not restricted to universal causality, but also include much broader universal positive/negative correlations and universal co-occurrence relationships, etc. Two phenomena, which are positive/negative correlated or co-occurrence, may be caused by the same reason or

merely co-occur. However, finding such correlations is also meaningful. For instance, based on the relevance of "flu" related online search queries and the breakout of seasonal pandemic influenza, Google Flu Trends can predict the occurrence of influenza in a more timely and accurate fashion [20].

- *Universal models* If we use machine learning models to learn universal phenomena inductively, or adopt mathematical formulas or physical models, etc. to characterize universal phenomena, universal machine learning models, universal mathematical models, universal physical models, etc. can be obtained correspondingly. For instance, based on the unsupervised deep learning method [7], Google's artificial brain learns to identify a cat. This artificial neural network is universal, because it can recognize any cat.
- *Universal mechanisms* Discovering universal mechanisms, which are behind universal phenomena in big data, has significant implications for gaining insights. For example, Barabási [5] explained that growth and preferential attachment are mechanisms common to a number of self-organizing complex networks, and they make vertex connectivities in large networks following power-law distributions at the system level.

3.3 Comparison of universal knowledge and knowledge generated by traditional data mining methods

Based on the above discussion about various categories of universal knowledge, we make a comparison between UKN and knowledge discovered by the traditional data mining techniques (traditional knowledge for short) as shown in Table 1. From the table, we can find that UKN has significantly different characteristics compared with traditional knowledge. (1) The scope is broader and its categories are more diverse. (2) UKN is extracted from the perspective of the whole system and is normally described at the macro level. (3) Both precise and general descriptions are acceptable. (4) UKN is mined from and tested in big data. (5) UKN affords easy comprehension and insights, and has a strong actionability.

4 Methodology

4.1 Comparison of research paradigms of big data mining and system science

In the era of big data, facing a novel class of data mining tasks, i.e., universal knowledge mining, it is urgent to do an interdisciplinary integration of research paradigms and techniques of related fields. For this purpose, in Table 2, a comparison is made between big data mining and system science, which are two main technical origins of UKD. The comparison shows that the approaches in these two fields have high mutual complementarity when treating big data. That is, complex systems science research is good at discovering macroscopic laws and internal mechanisms from big data; by contrast, data-driven big data mining specializes in handling various types of big data. So it is necessary to integrate the related fields, especial big data mining and complex systems science, to build a unified and interdisciplinary research paradigm for big-data-assisted universal knowledge discovery.

4.2 Dual-cycle methodology: a unified and interdisciplinary paradigm for UKD

To meet the above need of integrating related fields, in this subsection, I propose dual-cycle methodology and its discovery processes, which provide a unified and interdisciplinary research paradigm for universal knowledge discovery.

Firstly, we use "big data mining" (containing the techniques of data-driven analytics and statistics) and "system science" (including complexity thinking, system modelling and dynamics, etc.) to generalize the paradigms of these two fields respectively. In addition, we need to adopt the method of "simulation" in system science, and draw on some classical methods (or paradigms) in other disciplines or fields, as described below.

- *Simulation* This means the use of computer algorithms or other simulation systems to simulate real systems. This approach is especially fit for multi-agent simulation, system evolution simulation, numerical simulation, etc., can get simulation results quickly and has the advantage of low cost.
- *Controlled experiment* This means that an observer tests hypotheses by looking for changes brought on by manipulating one or more independent variables while all other variables are held constant [1]. It includes the steps of putting forward premises (including hypotheses), experimental operation, result verification, etc.
- *Theoretical analysis* This means the use of field theories accompanied by mathematical, physical and other tools in modelling, theoretical analysis, deduction, proof, etc.
- *Intelligence* Intelligence comes from machine intelligence (which comes from expert systems, machine learning, web intelligence, etc.), human intelligence (which can be individual and also collective intelligence), intelligence generated from human-computer interaction, etc.

Table 1 A comparison between characteristics of two types of knowledge

Aspects	Traditional knowledge	Universal knowledge
Types	Typically patterns, rules and machine learning models	Not only includes patterns, rules and machine learning models, but also contains distributions, characteristics, indexes, cor-relations, mathematical models, etc
Description level	Generally at the micro level or meso-level	Using a systematic perspective and describing at the macro level
Degree of precision	Most knowledge is precise and is described in detail	Both precise and general descriptions are acceptable
Data	Typically local and small data	Global big data
Evaluation	Some samples	Big data
Universality	Without further universal testing	A certain degree of universality
Comprehensibility	Commonly hard to understand	Affords easy comprehension and insights
Actionability	Typically weak	Strong
Calculation	Tend to be accurately calculated	Estimates are also effective

Table 2 Comparison of big data mining and system science

Aspects	Big data mining	System science
Focus	Treatability of big data, performance of platforms and algorithms, interestingness of discovered knowledge, etc	Three elements of systems, i.e., system structure, dynamics and functions
Perspectives	Big data based induction, statistics and machine learning	Systematic, evolutionary and network perspectives
Thinking	Collecting and using a complete data set, utilizing externalities of big data, effective estimates, looking for correlations, etc	Complexity thinking, e.g., self-organization, emergence, self-similarity and chaos
Models	Typically machine models	Mathematical and also physical models
Methods	Statistical analysis, data mining and machine learning algorithms	Statistical physics, computer simulations, algorithms, etc
Knowledge categories	Machine learning models, patterns and rules	Mathematical and also physical models, universal laws and internal mechanisms
Virtues	Specializing in handling various types of big data efficiently and discovering hidden statistical patterns	Having advantages in modeling systems, and discovering universal laws and internal mechanisms

- *Empirical research* Big data based UKD is a typical empirical problem. Therefore, we need to use an empirical cycle [23] for hypothesis testing, which includes observation (i.e., collecting and organizing empirical facts), induction (i.e., formulating hypothesis), modelling, deduction (i.e., deducting consequences of hypothesis as testable predictions), testing (i.e., testing the hypothesis with new empirical material) and evaluation. In each cycle, hypotheses and models are adjusted gradually to better fit the empirical facts.

On the basis of the above discussion, I propose a novel research paradigm for UKD called dual-cycle methodology. The methodology uses a combined dual-cycle to systematically organize various methods (e.g., machine intelligence, big data mining, experiment, simulation and theoretical analysis) by adopting the extended empirical research framework from the system science perspective. Human intelligence plays a core role in this creative discovery process.

The traditional knowledge discovery process [17] is unable to fully satisfy the requirements of UKD. Therefore, in the methodology, three types of cycle-driven UKD processes are proposed as below.

1. *Big-data-cycle-driven mining*
   Big-data-cycle based mining may be applied directly, under the condition of appropriate datasets and certain mining target of universal knowledge. Big-data-cycle-driven mining is shown in the left cycle of Figure 2, which consists of five steps: (1) big data accumulation and acquisition, (2) big data preprocessing, (3) big-data-driven analytics, (4) universal phenomenon discovery, and (5) putting forward universal phenomenon and post-processing. These five steps constitute a spiraling process, in order to discover universal phenomena from big data. The process of big-data-cycle-

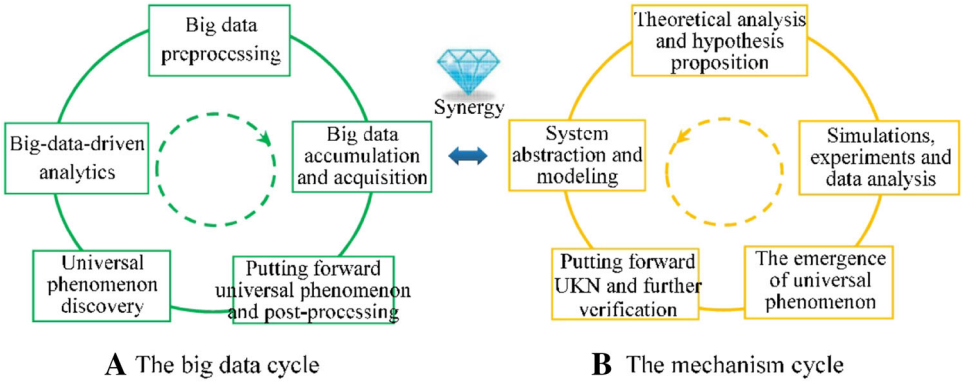

Fig. 2 Combined-dual-cycle-driven analytics. As two cycles rotate clockwise and counterclockwise respectively, execute the corresponding steps in each cycle synergistically, as they come face-to-face in the centre of the combined process

driven mining is suitable for searching, mining and verifying UKN with a clear target, such as verifying whether the data follows a certain distribution, mining frequent universal patterns from the data set, building a universal neural network model for the given data set, etc.

2. *Mechanism-cycle-driven mining*
   If empirical big data is missing, we may adopt "mechanism-cycle-driven mining" to explore UKN by theoretical analysis, experiments and simulations. It contains five steps as shown in the right cycle of Fig. 2: (1) system abstraction and modeling, (2) theoretical analysis and hypothesis proposition, (3) simulations, experiments and data analysis, (4) the emergence of universal phenomenon and (5) putting forward UKN and further verification. These five steps form a spiraling cycle called mechanism cycle, which adopts the empirical research framework. In the fourth step, if the deduction of hypotheses (which is proposed in the second step) well explains the observed experimental phenomena, the hypotheses are acceptable; otherwise, the hypotheses have to be revised gradually until the phenomena can be well-explained. Thus, accepted hypotheses and discovered universal phenomenon can be put forward, and need to be further verified when big data is available in the future.
   This process is designed to explore various classifications of UKN, especially internal mechanisms, mathematical and physical models, etc.
3. *Combined-dual-cycle-driven mining*
   Big data cycle specializes in discovering statistical universal phenomena, but is hard to induce internal mechanisms; by contrast, mechanism cycle is good at constructing internal mechanisms, but needs the support from big data cycle. So, it is necessary to combine them together. Actually, big-data-cycle-driven and mechanism-cycle-driven mining can be regarded as two special cases of combined-dual-cycle-driven mining. Therefore, in most cases, combined-dual-cycle-driven mining may be adopted if empirical big data is available.
   Combined-dual-cycle-driven mining combines both mechanisms in two cycles as shown in Fig. 2. It requires rotating two cycles clockwise and counterclockwise respectively, where corresponding steps execute in synergy as they come face-to-face in the centre of the combined process: (1) "big data accumulation and acquisition" in the big data cycle and "system abstraction and modeling" in the mechanism cycle are coordinated. The latter supervises the former, and the former reflects the latter. (2) Execute "big data preprocessing" in the left cycle and "theoretical analysis and hypothesis proposition" in the right cycle. They provide the basis for the subsequent steps. (3) Coordinate "big-data-driven analytics" in the left cycle and "simulations, experiments and data analysis" in the right cycle. They inspire each other. Besides, the former provides data analysis tools for the latter. (4) Integrate "universal phenomenon discovery" in the big data cycle and "the emergence of universal phenomenon" in the mechanism cycle. The universal phenomena discovered in two cycles should be consistent with each other. Besides, they should be well-explained by the hypotheses proposed in step 2. Otherwise, amend these hypotheses until the phenomena can be well-explained. (5) Integrate "putting forward UKN and post-processing" in two cycles. Here, the proposed UKN includes universal phenomena discovered in two cycles and accepted hypotheses. Thus, we have completed a combined dual-cycle. Universal phenomena and mechanisms discovered in this dual-cycle can be used as the basis for a new round of combined-dual-cycle-driven mining.

This process benefits from the capabilities of two cycles, and is capable for progressively discovering both statistical universal phenomena and in-depth internal mechanisms in big data.

5 Case study: mining universal knowledge from international trade flow data

In this section, I use our recent work [46], which mines UKN from international trade flow data, as an example to illustrate three types of cycle-driven mining processes proposed in the above section, especially combined-dual-cycle-driven mining.

The background is as below. The international trade system is a global system which connects countries around the world. From the perspective of data completeness of a system [36], a complete (or approximately complete) international trade dataset depicting the whole international trade system can be regarded as big data. Here, several illustrations are adopted to show the effectiveness of the proposed UKD processes. In these illustrations, the mining processes are focused. More details can be referred to [46].

5.1 Illustration of big-data-cycle-driven mining

If suitable big data are available, and the target is to discover universal phenomena from data, big-data-cycle-driven mining may be adopted directly. The process is quite straightforward. For example, suppose the target is to verify whether the trading value of countries always follows a certain type of distribution for a large variety of commodities. The steps are as follows.

Step 1. Big data accumulation and acquisition

A trade dataset (i.e., NBER-United Nations trade data [18]) is available, which describes the details of global bilateral trading from 1962 to 2000. There are totally 190 countries and 1288 types of commodities in the dataset. Therefore, it is quite suitable for the verification.

Step 2 and *3.* Big data preprocessing and data-driven analytics

Towards the target, the dataset is preprocessed. Find an acceptable model which can well fit the distribution data.

Step 4 and *5.* UKN discovery, presentation and post-processing

Based on the analysis, it is find that the data always follows a power-law form with an exponential cutoff. Then, putting forward the UKN for further utilization.

More work can be done using this process, such as verifying whether countries prefer to trade with other countries with close geographical distances and whether countries' trade value is significantly correlated with their GDP. The discovered universal phenomena can well support mechanism-cycle-driven analytics, e.g., building a universal growth model for trade flow networks [46].

5.2 Illustration of mechanism-cycle-driven mining

Suppose we do not have the empirical big data. Since the international trade system is a typical flow system with different types of commodity flows, we want to use a flow network model to model the system, and then explore internal mechanisms of flow systems. The steps of mechanism-cycle-driven mining are as follows.

Step 1. System abstraction and modeling

We construct a multi-layer open flow network (MFN) [46] to model the global trade system, where each layer contains a specific type of commodity flow, and all flows start from the source node and end at the sink node. This model can well describe multi-types of product flows in the international trade system.

Step 2. Theoretical analysis and hypothesis proposition

In a MFN, a fundamental problem is to calculate the flow distance between two nodes according to the mechanism of flow systems. Therefore, a thought experiment of particles' random walks (which belongs to theoretical analysis) is designed and several types of flow distances (such as symmetric minimum flow distances) are deduced [24, 46]. This step builds a foundation for the subsequent steps.

Step 3. Simulations, experiments and data analysis

According to the proposed thought experiment, simulations can be performed to simulate particles' random walks in a MFN. Besides, based on the deduced formulas of flow distances, given a MFN, the distance matrices, which record different types of flow distances between nodes, can be obtained.

Step 4. The emergence of universal phenomenon

By analyzing the generated distance matrices, several universal phenomena are discovered. For example, diagonal elements of the first-passage flow distance matrix are all 0; elements in the total flow distance matrix are always no smaller than the corresponding elements in the first-passage flow distance matrix.

Step 5. Putting forward UKN and further verification

The UKN discovered in this mechanism cycle (such as MFN model, thought experiment of particles' random walks in the MFN, formulas of flow distances and zero diagonal elements in the first-passage flow distance matrix) can be put forward, and be verified further.

Thus, the above five steps form a spiraling process for UKD. Based on the achievements of this cycle, a new round of mechanism-cycle-driven mining can be triggered and more in-depth universal knowledge can be produced. For instance, we may define node centrality and realize

community detection in the MFN based on the formulas of flow distances.

5.3 Illustrations of combined-dual-cycle-driven mining

Since big data cycle and mechanism cycle are actually two special cases of dual-cycle process, in most cases, combined-dual-cycle-driven mining is adopted. In real applications, it is a multi-round mining process, which contains multiple rounds of dual-cycle. Here is an example.

5.3.1 The first round of mining

Step 1. The coordination of "big data accumulation and acquisition" and "system abstraction and modeling"

The available big data contains the bilateral trade information of year, exporter, importer, product category, quantity and value. Therefore, we adopt multi-layer open flow network [46] for system modeling, where the weight of edge records the value of the trading, and the direction of edge reflects the direction of product flow. According to the system modeling, some work (such as accumulating trading data at a finer time granularity) can be carried out towards big data accumulation. Thus, the system modeling and in-depth analytics can be further supported.

Step 2. "Big data preprocessing" and "theoretical analysis and hypothesis proposition"

To meet the need of system modeling, the big data is preprocessed. Then, based on the MFN model, theoretical analysis and hypothesis proposition are performed. Here, the formulas of different types of flow distances are deduced. This work builds a basis for the subsequent steps.

Step 3. The coordination of "big-data-driven analytics" and "simulations, experiments and data analysis"

We applied the calculation methods of flow distances to the empirical big data. For each layer of MFN, we simulate the thought experiment of particles' random walk, and compute different types of flow distances. Then, mine these data and compare patterns among different layers using data-driven analytics techniques [46]. Here, "big-data-driven analytics" provides tools for deeply analyzing the data generated by "simulations and experiments".

Step 4. The integration of "universal phenomenon discovery" and "the emergence of universal phenomenon"

Universal phenomena emerge based on analyzing flow distance data in Step 3. For example, the distribution of flow distances from the source to each country node emerges regular patterns for different types of products; the flow distances from the source to the sink for different types of products are correlated with the categories of products [46].

Step 5. Putting forward UKN produced in two cycles

The UKN discovered in the dual-cycle is put forward for further utilization.

5.3.2 The second round of mining

Based on the UKN obtained in the first round of dual-cycle, a new round of combined-dual-cycle-driven mining can be triggered.

Step 1. "Big data accumulation" and "system modeling"

This step is skipped due to the same with that in the first round.

Step 2. "Big data preprocessing" and "theoretical analysis and hypothesis proposition"

According to the formulas of flow distances obtained from the previous round, we may further define node centrality. And then propose the hypothesis that competitive countries occupy the central positions in the trading.

Step 3. "Big-data-driven analytics" and "simulations, experiments and data analysis"

Focusing on the hypothesis, countries' centralities are computed and ranked for different types of products, and then displayed using visualization techniques.

Step 4 and *5.* The integration of "universal phenomenon discovery" and "the emergence of universal phenomenon", and putting forward UKN produced

Similar to social stratification, it is interesting to find the phenomenon of centrality stratification in global trading. That is to say, the competitive countries tend to be in the center positions in the trading of a large variety of products, while underdeveloped countries likely rank low in their limited varieties of trading products [46]. Thus, the hypothesis is confirmed and put forward.

On the basis of UKN discovered in these two rounds, more rounds of dual-cycle mining can be carried out. For instance, some abnormal countries are detected, which import many types of commodities that the vast majority of countries do not need to import. It may indicate that these countries are at high risk. Thus, a new round of mining is needed to verify whether this universal indicator exists or not.

6 Discussions and conclusions

The flood of data in scientific field has been and is still being generated constantly. In result, the deluge of big data has attracted increasing attention in recent years [3, 6, 33, 35]. Data-intensive scientific discovery has become the fourth paradigm for scientific exploration, after experimental research, theoretical research and computer simulations, according to the pioneer scientist Jim Gray

[25]. Big data is regarded as a big deal since it promises to change the world [44]. However, how to deal with big data to provide big insights and make a big impact is still an open question [30].

To meet the big-data challenges for offering insights, this paper highlights a framework of universal knowledge discovery from big data (UKD). The new concept is defined as knowledge and laws with a certain degree of universality and immutability in the statistical analyses of big data. Since big data means a "complete" or "approximately complete" data set [36], this new concept implies that if knowledge stands in the statistical analyses of big data, this knowledge is universal. This kind of valuable universal knowledge encompasses a much broader scope than scientific laws, and is clearly suggested as the mining target.

UKD has three meaningful implications. (1) UKD offers a pathway to the discovery of "treasure-trove" hidden in big data. It is quite hopeful that more and more universal knowledge with a big impact will be discovered from big data in the future. (2) If a novel universal phenomenon discovered from big data, it is necessary to explore the origins of this universal phenomenon. (3) Traditional knowledge obtained based on small-scale experimental data, simulations, or incomplete large-scale real data in a spectrum of scientific disciplines should be re-examined by big data. If the knowledge still stands in big data, it becomes universal knowledge.

To accelerate big data discovery, an interdisciplinary research paradigm is proposed, which integrates related disciplines, especially big data mining and complex systems science. The work tries to answer the question that how to integrate the fourth paradigm (i.e., data driven scientific discovery) with the previous three paradigms (i.e., experimental research, theoretical research and computer simulations) for scientific discovery, and builds a foundation for the future development of UKD technology.

I advocate that that it is necessary to bring scientists from related disciplines together under one roof to mine valuable universal knowledge from big data. I envision that the universal knowledge discovery paradigm will become a key basis to address the spectrum of big data challenges. In the light of UKD, it is quite hopeful that more and more universal knowledge with a big impact will be discovered from big data in the future. Further work towards UKD may include developing advanced techniques to mine various types of UKN, and applying combined-dual-cycle-driven mining to various applications.

Acknowledgments The author would like to thank A/Prof. Jiang Zhang for the collaboration on the case study. Besides, thanks are given to Prof. Rao Kotagiri, Prof. Tao Zhou, as well as the editors and the anonymous reviewers, and others for their valuable suggestions. This work is supported by Zhejiang Provincial Philosophy and Social Science Foundation of China (No. 15NDJC145YB), National Nature Science Foundation of China (No. 71271191), the National Science and Technology Pillar Program during the 12th Five-year Plan Period of China (No. 2012BAF12B11), Zhejiang Provincial Natural Science Foundation of China (No. LY15F020036) and Scientific Research Foundation for the Returned Overseas Chinese Scholars (Ministry of Human Resources and Social Security of China, 2013).

References

1. What is a controlled experiment? http://www.innovateus.net/innopedia/what-controlled-experiment. Accessed 15 May 2014
2. Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, vol 22, pp 207–216. ACM
3. Akil H, Martone ME, Van Essen DC (2011) Challenges and opportunities in mining neuroscience data. Science (New York, NY) 331(6018):708
4. Barabási AL (2010) Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades. Penguin Group, New York
5. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
6. Bell G, Hey T, Szalay A (2009) Beyond the data deluge. Science 323(5919):1297–1298
7. Bengio Y (2009) Learning deep architectures for ai. Foundations Trends Mach Learn 2(1):1–127
8. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439(7075):462–465
9. i Cancho RF, Solé RV (2001) The small world of human language. Proc R Soc Lond Ser B Biol Sci 268(1482):2261–2265
10. Cao L (2012) Actionable knowledge discovery and delivery. Wiley Interdiscip Rev Data Mining Knowl Discov 2(2):149–163
11. Cao L, Yu S (2012) Behavior computing. Springer, Berlin
12. Chen W, Chen Y, Weinberger KQ (2014) Fast flux discriminant for large-scale sparse nonlinear classification. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 621–630. ACM
13. Chesney T (2014) Networked individuals predict a community wide outcome from their local information. Decis Support Syst 57:11–21
14. Clauset A, Moore C, Newman ME (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101
15. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C (2009) Mad skills: new analysis practices for big data. Proc VLDB Endow 2(2):1481–1492
16. Fan W, Bifet A (2013) Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newsl 14(2):1–5
17. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37
18. Feenstra RC, Lipsey RE, Deng H, Ma AC, Mo H (2005) World trade flows: 1962–2000. Tech. rep, National Bureau of Economic Research
19. Gao L, Song C, Gao Z, Barabási AL, Bagrow JP, Wang D (2014) Quantifying information flow during emergencies. Scientific reports 4
20. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
21. Viswanathan GM (2010) Fish in lévy-flight foraging. Nature 465:1018–1019
22. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782
23. Groot AD, Spiekerman JA (1969) Methodology: foundations of inference and research in the behavioral sciences. Mouton, The Hague

24. Guo L, Lou X, Shi P, Wang J, Huang X, Zhang J (2015) Flow distances on open flow networks. arXiv preprint arXiv:1501.06058
25. Hey AJ, Tansley S, Tolle KM et al (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Washington
26. IBM (2012) What is big data? http://www-01.ibm.com/software/data/bigdata/. Accessed 10 Dec 2012
27. Jiawei H, Kamber M (2001) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco
28. Langley P (1978) Bacon. 1: a general discovery system. In: Proceedings 2nd Biennial conference of the Canadian society for computational studies of intelligence, pp 173–180
29. Langley P (1979) Rediscovering physics with bacon. 3. In: IJCAI, pp 505–507
30. Lazer DM, Kennedy R, King G, Vespignani A (2014) The parable of google flu: traps in big data analysis. Science 343(6176):1203–1205
31. Lin J, Ryaboy D (2013) Scaling big data mining infrastructure: the twitter experience. ACM SIGKDD Explor Newsl 14(2):6–19
32. Liu CL, Tsai TH, Lee CH (2014) Online chinese restaurant process. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 591–600. ACM
33. Lynch C (2008) Big data: how do your data grow? Nature 455(7209):28–29
34. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed 20 May 2011
35. Marx V (2013) Biology: the big challenges of big data. Nature 498(7453):255–260
36. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, Boston
37. Milgram S (1967) The small world problem. Psychol Today 2(1):60–67
38. Nordhausen B, Langley P (1993) An integrated framework for empirical discovery. Mach Learn 12(1–3):17–47
39. Peng C, Jin X, Wong KC, Shi M, Liò P (2012) Collective human mobility pattern from taxi trips in urban area. PloS one 7(4):e34–487
40. Pulvermüller F (2002) The neuroscience of language: on brain circuits of words and serial order. Cambridge University Press, Cambridge
41. Qian X (1981) A rediscussion on the system of system science. Syst Eng Theory Prac 1:1–3
42. Qian X, Yu J, Dai R (1990) A new field of science: open complex giant system and its methodology. Nature Mag China 13:3–10
43. Rajaraman A, Ullman JD (2011) Mining of massive datasets. Cambridge University Press, Cambridge
44. Shawn J (2014) Why "big data" is a big deal: information science promises to change the world. Harv Mag. Harvardmagazine.com
45. Shen B (2014) A comparative study and an integration of research paradigms of complex networks and data mining. Complex Syst Complex Sci 11(1):48–52
46. Shen B, Jiang Z, Qiuhua Z (2015) Exploring multi-layer flow network of international trade based on flow distances. arXiv preprint. http://arxiv.org/abs/1504.02361v1
47. Shen B, Yao M, Wu Z, Gao Y (2010) Mining dynamic association rules with comments. Knowl Info Syst 23(1):73–98
48. Tang C (2014) http://blog.sciencenet.cn/home.php?mod=space&uid=287179&do=blog&id=765603. Accessed 8 Feb 2014
49. Tang C, Zhang Y, Tang L, Li C, Chen Y (2008) A survey on mining kinetic intervention rule from sub-complex systems. J Comput Appl 28(11):2732–2736
50. Wang B, Zhou T, Zhou C (2012) Statistical physics research for human behaviors, complex networks and information mining. J Univ Shanghai Sci Technol 34(2):103–117
51. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
52. Wu WT (1994) Mechanical theorem proving in geometries: basic principles. Springer, Berlin
53. Xu G, Gu J, Che H (2000) System science. Shanghai scientific and technological education press, Shanghai
54. Yan XY, Han XP, Wang BH, Zhou T (2013) Diversity of individual mobility patterns and emergence of aggregated scaling laws. Scientific reports 3
55. Yang Q, Wu X (2006) 10 challenging problems in data mining research. Int J Info Tech Decis Mak 5(04):597–604
56. Zhang F, Wilkie D, Zheng Y, Xie X (2013) Sensing the pulse of urban refueling behavior. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp 13–22. ACM
57. Zhou T (2013) In the big data era, china has not lagged behind. http://blog.sciencenet.cn/blog-3075-657481.html. Accessed 29 Jan 2013
58. Zhu YX, Huang J, Zhang ZK, Zhang QM, Zhou T, Ahn YY (2013) Geography and similarity of regional cuisines in china. PloS One 8(11):e79–161
59. Zikopoulos P, Eaton C et al. (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, New York
60. Zytkow JM, Simon HA (1986) A theory of historical discovery: the construction of componential models. Mach Learn 1(1):107–137
61. Zytkow JM, Zhu J, Hussam A (1990) Automated discovery in a chemistry laboratory. In: AAAI, pp 889–894