Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (22): 4936-4942, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.4936.4942

Study on the Big Data Method for Low Carbon Campus Governance

[1]Liping Fu, [2]Yifang Liu and [1]Yong Liu
[1]Research Center of Public Resource of Management, Tianjin University, China
[2]College of Management and Economics, Tianjin University, China

Abstract: The ability to manage big data and turn it into actionable information that can improve the research methodologies of low carbon campus management is a significant success factor in achieving validity and reliability of research. Big data refers not only to data itself but also to the methods employed to mine and analyze large collections of information to solve complex problems. The above mentioned characteristics pose challenges for the low carbon campus management researchers who want to take advantage of big data and employ it to better research method. Traditional methods and data management technologies were not designed to accommodate high volumes of data that are dynamic, furthermore, they were also not designed to collect and provide access to real-time data, as well as managing heterogeneous data from multiple sources that include both structured and unstructured data. Therefore, in order to take advantage of big data, researchers should corporate with technology partners that understand big data, employing interdisciplinary methods to manage, normalize and analyze it. Employing psychological measurement, system dynamics and complex system methods, as well as agent-based simulation platform, the present paper put forward the conceptual protocol for describing the methodologies based on big data, which can be adopted in the research of low carbon campus management and helps to understand the complexity and betters the management of low carbon campus and establishes a solid foundation for further research.

Key words: Low carbon campus management, research methodologies, big data

INTRODUCTION

Information generation is outstripping growth in storage capacity and the gap continues to grow. There are almost 2 billion regular Internet users globally and total internet data traffic will top 667 exabytes by 2013. Governments, citizen and communities are creating and consuming vast amounts of data. According to the research of Gartner that enterprise data in all forms will grow 650% over the next five years. Similarly, according to the findings of IDC, the world's volume of data doubles every 18 months. On the other hand, data marketplaces, where you go to get the data you need, are growing. Third party data availability is on the rise, with the estimated worldwide market valued at $100B (Hall and Karmasphere, 2011). The flood of data, often named as "information overload," or "big data," creates a challenge for management, especially low carbon campus management. The ability to manage big data and turn it into actionable information that can improve the research methodologies of low carbon campus management is a significant success factor in achieving validity and reliability of research.

CHARACTERISTICS OF BIG DATA

The definition of "Big Data" varies greatly and is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time, meanwhile, big data refers not only to data itself but also to the methods employed to mine and analyze large collections of information to solve complex problems. The Characteristics of big data include high data volume, velocity and variety (Laney, 2001).

Volume and velocity: Volume refers to unprecedented quantities of information generated by today's fast-moving businesses and technology. Data collected over the course of days or weeks exceeds the entire corpus of legacy data in a given domain. Velocity refers to the speed at which data is being collected and processed. This involves streams of data, structured record creation and availability for access and delivery. Velocity means both how fast data is being produced and how fast the data must be processed to meet demand.

Variety: A distinguishing feature of Big Data is a mixture of traditional structured data together with massive

Corresponding Author: Yong Liu, Research Center of Public Resource Management, Tianjin University, Tianjin of China

amounts of unstructured information. There are more types of information to analyze, including tabular data, hierarchical data, government documents, e-mail, metering data, video, still images, audio, stock ticker data etc., which can be organized into two broad categories:

Structured data: Structured Data by definition already resides in formal data stores and it is often categorized as legacy data. It typically refers to formal data groupings into database records with named fields or row and column organization.

Unstructured data: Unstructured Data, by contrast, comprises data collected during other activities and stored in amorphous logs or other files in a file system. Unstructured data can include raw text or binary and contain a rich mix of lexical information and/or numerical values, with or without delimitation, punctuation or metadata.

CONCEPTUAL PROTOCOL OF THE METHODOLOGIES

The above mentioned characteristics pose challenges for the low carbon campus management researchers who want to take advantage of big data and employ it to better research method. Traditional methods and data management technologies were not designed to accommodate high volumes of data that are dynamic, furthermore, they were also not designed to collect and provide access to real-time data, as well as managing heterogeneous data from multiple sources that include both structured and unstructured data. Therefore, in order to take advantage of big data, researchers should employ interdisciplinary approaches to data management and focus on the following stages in the data lifecycle:

Identify and filter: The first step to manage big data is to prepare the researcher to be able to quickly accommodate data sources, then to find where the data is coming from, who is creating it and where the content lives. In order to filter these data effectively, it's necessary to provide tools, such as data storage infrastructure and filter skills, to determine what is important and what does not matter. Interdisciplinary approaches can be employed to explore big data, especially for unstructured data, which are often multi-dimensional constructs, for example, behavior of government and residents. The process of identifying the unstructured data involved: specifying the domain of data; designing the items of different facets and purifying them; assessing reliability and validity of the process. The paradigm has worked well in many studies; it can reduce the tendency to apply extremely sophisticated analysis to faulty data and thereby execute a GIGO (garbage in, garbage out) routine.

Specifying domain and generating items: To specify the domains of each construct, an initial list of facet of unstructured data was developed, based mainly on the literature. These facets should be then reviewed by experts. The ability of these experts to comment on the facets was established according to their educational level and specialties. A five-point Likert scale (one = least important; five = most important) was used as the measurement. According to the marks given by the experts and the frequency in the literature, facets of unstructured data may be deleted due to low marks and frequency. Next, according to the literature, items were designed for the remaining facets and a five-point Likert scales was used to measure them. Consequently, a prototype questionnaire with facets and items was developed.

Collecting data and purifying measures: The prototype questionnaire with facets and items should be tested empirically and data can be collected. By analyzing the data of the questionnaires with an exploratory factor-analysis model (using the principal components method and oblique rotation), some factors of unstructured data with eigenvalues greater than 1.00 (explaining some percent of the variance) were identified. Examination of the facets revealed that some of them (a) Had low factor loadings (not greater than 0.50) and square multiple correlations, (b) Were highly correlated and Ross-loaded and (c) Presented elevated modification indices. Thus, the most problematic facets should be removed. A closer look at the remaining facets that loaded on factors indicated the "name" of factors. Finally, a questionnaire of factors and facets was developed.

Assessing reliability and validity: Reliability is defined as the extent to which a questionnaire, test or any measurement procedure produces the same results on repeated trials. it is the stability or consistency of scores over time. There are three aspects of reliability including equivalence, stability and internal consistency. It is important to understand the distinction between these three as it will guide one in the proper assessment of reliability given the research protocol. Equivalence refers to the amount of agreement between two or more instruments that are administered at nearly the same point in time. Equivalence is measured through a parallel forms procedure in which one administers alternative forms of the same measure to either the same group or different

group of respondents. This administration of the various forms occurs at the same time or following some time delay. The higher the degree of correlation between the two forms, the more equivalent they are. In practice the parallel forms procedure is seldom implemented, as it is difficult, if not impossible, to verify that two tests are indeed parallel (i.e., have equal means, variances and correlations with other measures). Inter rater reliability refers to the consistency with which observers or raters make judgments. The validity of a questionnaire relies first and foremost on reliability. If the questionnaire cannot be shown to be reliable, there is no discussion of validity. Demonstrating validity is easy, compared to reliability. If we have reached this point and have a reliable instrument for measuring the unstructured data, validity will not be difficult.

Validity refers to whether the questionnaire measures what it intends to measure. While there are very detailed and technical ways of proving validity that are beyond the level of this discussion, there are some concepts that are useful to keep in mind. The overriding principle of validity is that it focuses on how a questionnaire or assessment process is used. Reliability is a characteristic of the instrument itself, but validity comes from the way the instrument is employed. There are a number of types of validity including:

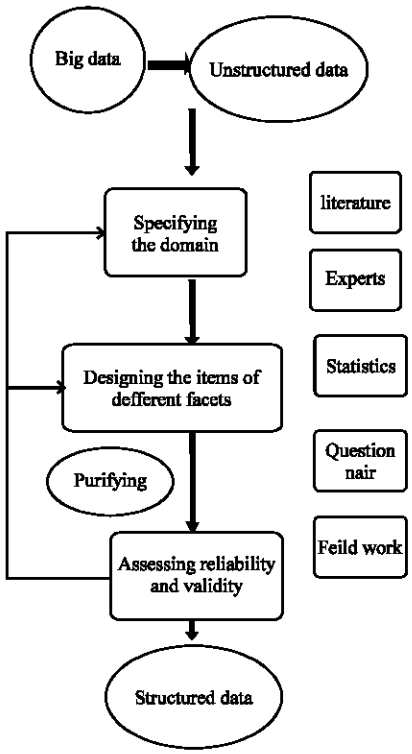

Fig. 1: Process of identifying unstructured data

Face validity: The questions appear to be measuring the construct. This is largely a common-sense assessment, but also relies on knowledge of the way people respond to survey questions and common pitfalls in questionnaire design.

Content validity: Whether all important aspects of the construct are covered. Clear definitions of the construct and its components come in useful.

Criterion validity: Whether scores on the questionnaire successfully predict a specific criterion. It is a measure of how well one variable or set of variables predicts an outcome based on information from other variables and will be achieved if a set of measures from a personality test relate to a behavioral criterion on which psychologists agree. A typical way to achieve this is in relation to the extent to which a score on a personality test can predict future performance or behavior. Another way involves correlating test scores with another established test that also measures the same personality characteristic.

Concurrent validity is a parameter used in sociology, psychology and other psychometric or behavioral sciences. Concurrent validity is demonstrated where a test correlates well with a measure that has previously been validated. The two measures may be for the same construct, or for different, but presumably related, constructs. The two measures are taken at the same time. This is in contrast to predictive validity, where one measure occurs earlier and is meant to predict some later measure. Concurrent validity and predictive validity are two types of criterion-related validity. The difference between concurrent validity and predictive validity rests solely on the time at which the two measures are administered. Concurrent validity applies to validation studies in which the two measures are administered at approximately the same time. For example, an employment test may be administered to a group of workers and then the test scores can be correlated with the ratings of the workers' supervisors taken on the same day or in the same week. The resulting correlation would be a concurrent validity coefficient. Finally, the unstructured data can be transformed to structured data, which shows in Fig. 1.

Analysis and simulation: In the Complex Adaptive System (CAS), nested hierarchies, a multiplicity of cross-scale interactions and feedback loops imply a high degree of complexity and non-linear behavior which predictive equilibrium models fail to calculate. Establishment of a CAS model means to incorporate

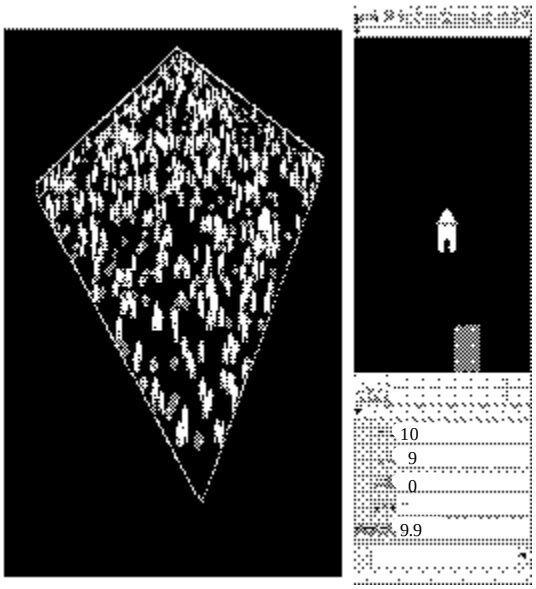

Fig. 2: NetLogo simulation platform of low carbon campus management

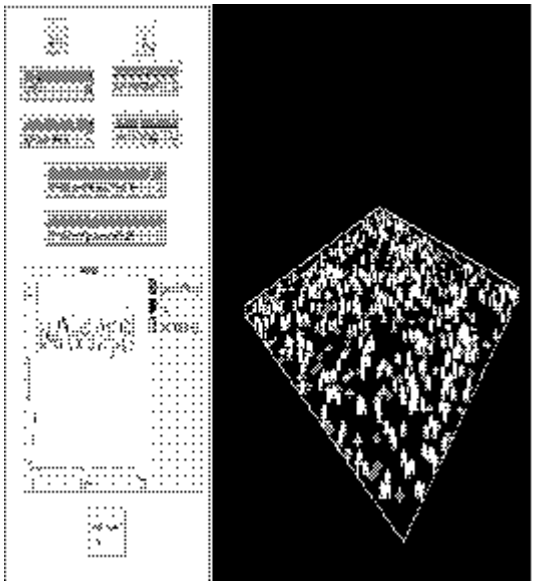

Fig. 3: NetLogo's user interface, with the present model

variability, adaptations, uncertainty and nonlinearity while aiming for improved understanding of how co-evolutionary processes and dynamic patterns emerge and interact across different agents. Path dependence and system memory also offer a conceptual framework for applying the insights and data from the CAS model. An agent-based model has been successfully applied in many researches (Guerci *et al*., 2005; Veit *et al*., 2009; Genoesea *et al*., 2007; Liu and Hong, 2012); it provides a powerful analytical method that enables the modeling of many heterogeneous real-world agents as individual software programs.

The complex, dynamic evolvement and behavior are captured within each software agent, for example, NetLogo platform (Fig. 2 and 3), thereby creating a virtual simulation of the real world (Peters and Brassel, 2000; Tesfatsion, 2006). NetLogo is a multi-agent programming language and modeling environment for simulating complex phenomena. It is used across a wide range of fields and allows experiments to be carried out by easily changing parameter values and viewing an updated image of the system over time, together with the trends of key variables. NetLogo is written in Java language and can be run on all major platforms, which uses three types of agents: turtles, patches and observer. Turtles are agents that are moving inside the world. The world is composed by patches and the observer does not have a specific location, just like an entity that observes the world. NetLogo uses four types of variables: global variables, turtle's variables, patches variables and system variables. Modelers can give instructions to independent agents all operating concurrently, which makes it possible to explore connections between micro-level behaviors of individuals and macro-level patterns that emerge from their interactions.

This method for exploring the unstructured data (behavior) of low carbon campus management is considered useful because of its ability to capture the complex of agents, dynamic evolvement of behavior and their interactions in a way that is beyond the scope of traditional methods.

An agent is characterized by state variables: The variables selected are substantiated by the literature or a field survey. The platform needs design concepts, as mentioned below, which provide a common framework for designing and communicating the agent-based model.

Emergence: System-level phenomena emerge from individual traits and phenomena that are imposed on agents.

Adaptation: Agents have adaptive traits that directly or indirectly affect their potential fitness in response to changes in themselves or the social environment. For example, companies plan optimal strategies to maximize

their profits and change their behavior according to influencing factors. They also learn from these factors and adjust their strategies accordingly.

Fitness: In agent-based models that do not address animals or plants, various agents can be considered in terms of "fitness." Such as firms are profit-maximizing agents. Fitness is calculated according to the amount of profits.

Prediction: Agents predict future consequences of their decisions and the future environment they will experience.

Sensing: With regard to their adaptive decisions, agents are assumed to understand their own internal conditions, such as income and age. They also understand some external environmental variables, such as job-searching market and governmental regulations.

Interaction: Agents may compete in social environment and interact with other agents directly or indirectly, for example, community, doctor and government.

Stochasticity: Stochasticity is part of the model and includes such factors as the location of a agent. Stochasticity is used to control these variables because the focus of the model is on system-level environmental behavior phenomena, not individual behavior.

Collectives: Agents with a different category of behavior are grouped under collective behavior.

Predictive and optimization: It's necessary to predict future trends and behavior patterns or to optimize the performance of a process in the research of low carbon campus management, which means researchers have to evolve from data analysis to insight to prediction. Applying the right methods in the right case is crucial to this evolution, for example system dynamics. System dynamics is an approach to understanding the behavior of complex systems over time. It deals with internal feedback loops and time delays that affect the behavior of the entire system. What makes using system dynamics different from other approaches to studying complex systems is the use of feedback loops and stocks and flows. These elements help describe how even seemingly simple systems display baffling nonlinearity.

System dynamics is a computer-aided approach to policy analysis, predictive and optimization. It applies to dynamic problems arising in complex social and economic systems, which characterized by interdependence, mutual interaction, information feedback and circular causality. The system dynamics approach involves: (1) Defining the research problems dynamically, including an endogenous, behavioral view of the significant dynamics of a system, focus on the characteristics of a system generated or exacerbated the perceived research problems. (2) Checking on concepts in the real system as continuous quantities interconnected in loops of information feedback and circular causality. (3) Identifying different categories of variables, including independent stocks or accumulations (levels) and their inflows and outflows (rates). (4) Formulating a behavioral model capable of reproducing, this is usually a computer simulation model. (5) Deriving understandings and applicable policy insights from the results of operating the model (Fig. 4).

CONCLUSION

In information technology, big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime and determine real-time roadway traffic conditions.

Although the definition of "Big Data" varies greatly, it's generally believed the characteristics of big data include volume, velocity and variety. Volume refers to the unprecedented quantities of information. Velocity refers to the speed at which data is being collected and processed. Variety refers to the types of information, which can be organized into two broad categories: structured data and unstructured data. The above mentioned characteristics pose challenges for the low carbon campus management researchers who want to take advantage of big data and employ it to better research method. Employing interdisciplinary approaches, such as system dynamic and multi-agent based simulation, conceptual protocol of the methodologies to big data management has been put forward, which include identify and filter, analysis and simulation, as well as predictive and optimization. The protocol can be adopted in the research of low carbon campus management and helps to understand the complexity and betters the management of low carbon campus management (Fig. 5). As can be seen from Fig. 4, in order to take advantage of big data, interdisciplinary methods were employed to manage and

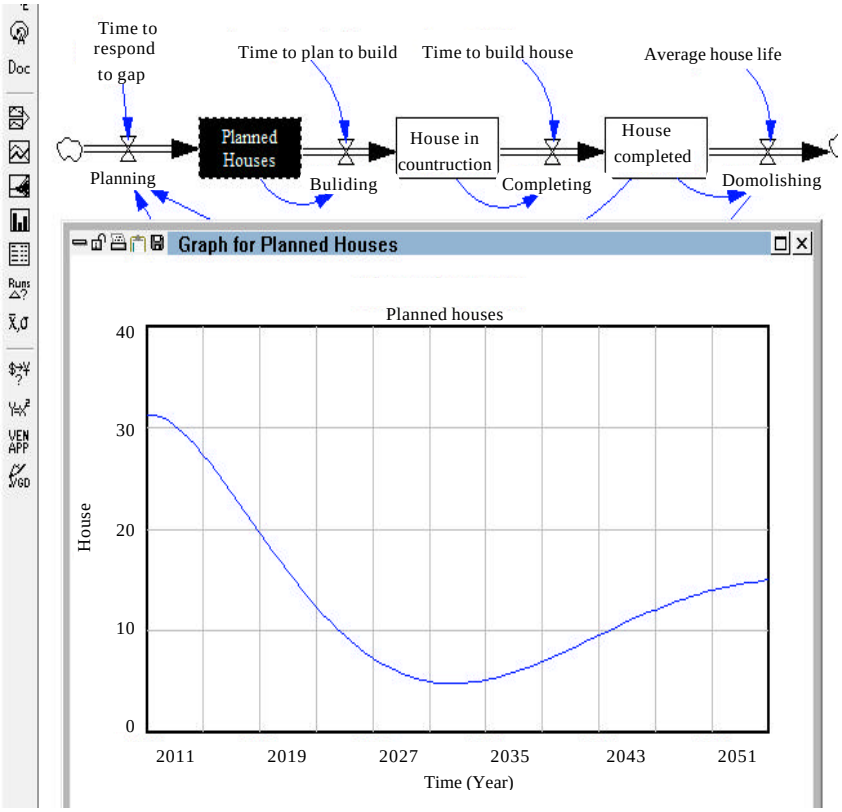

Fig. 4: Prediction platform of system dynamics

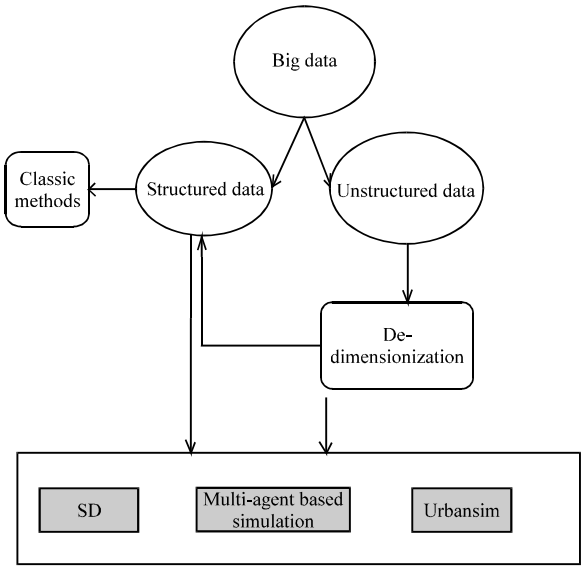

Fig. 5: Conceptual protocol of the methodologies

analyze it, therefore, researchers should corporate with different researchers who focus on cloud computing technology, complex system and psychological measurement etc. (Hamid, *et al*., 2010; Bo and Yang, 2012).

Finally, some limitations are worth mentioning. This study did not include all of the methods that may be employed to explore the big data. Moreover, the methodologies are just conceptual models. Nevertheless, the explorative research provides a context and starting point for further investigations with expanded methods and will help further studies into low carbon campus management in China.

REFERENCES

Bo, J.F. and Z.X. Yang, 2012. The constructing of mineral exploration data management system based on spatial database. AISS, 4: 234-240.

Genoesea, M., F. Sensfu, D. Mosta and O. Rentza, 2007. Agent-based analysis of the impact of CO_2 emission trading on spot market prices for electricity in Germany. Pac. J. Optim., 3: 401-423.

Guerci, E., S. Ivaldi, S. Pastore, S. Cincotti, 2005. Modeling and implementation of an artificial electricity market using agent-based technology. Phys. A Stat. Mech. Appl., 355: 69-76.

Hall, M. and F. Karmasphere, 2011. Understanding the elements of big data: More than a hadoop distribution. White Paper, Martin Hall, Founder, Karmasphere, May 2011.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. Application Delivery Strategies, META Group Inc., Stamford, CT., USA., February 6, 2001.

Liu, Y. and H. Ye, 2012. The dynamic study on firm?s environmental behavior and influencing factors: An adaptive agent-based modeling approach. J. Cleaner Prod., 37: 278-287.

Peters, I. and K.H. Brassel, 2000. Integrating computable general equilibrium models and multi-agents. Proceedings of the International Conference on AI Simulation and Planning, October 4-6, 2000, Tuscon Arizona.

Tesfatsion, L., 2006. Agent-Based Computational Economics: A Constructive Approach to Economic Theory. In: Handbook of Computational Economics, Tesfatsion, L. and K.L. Judd (Eds.). Elsevier, USA., pp: 831-880.

Veit, D.J., A. Weidlich and J.A. Krafft, 2009. An agent-based analysis of the German electricity market with transmission capacity constraints. Energy Policy, 37: 4132-4144.