Similarities between "Big Data" and Traditional Chinese Medicine information

Meng Cui, Haiyan Li, Xueqin Hu

Meng Cui, Haiyan Li, Xueqin Hu, Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

Supported by the Fundamental Research Funds for the Central public welfare research institutes (No. ZZ2070822); Innovative Research Team in CACMS (No. PY1306)

Correspondence to: Dr. Haiyan Li and prof. Meng Cui, Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China. Lihy@mail.cintcm.ac.cn

Telephone: +86-10-64014411-3246

Accepted: May 1, 2014

Abstract

In this paper, we analyze the four distinct characteristics of information on Traditional Chinese Medicine (TCM), namely epistemological information, phenomenon information, overall information, and time information. These characteristics bear to some extent strong similarity to the three characteristics of "Big Data", namely integrity data, fuzzy data and correlation data, so the advent of the age of "Big Data" is bound to create good opportunities for the development of TCM informatics and is also be expected to provide methods and techniques for processing and analysis of TCM "comprehensive data".

© 2014 JTCM. All rights reserved.

Key words: Medicine, Chinese traditional; Medical informatics; Big data; Comprehensive data

INTRODUCTION

Traditional Chinese Medicine (TCM) informatics is a cross discipline of TCM and informatics. It is an emerging discipline that takes TCM information as the research subject, the motion law of TCM information as the research contents and TCM informatics methodology as the research method (pay close attention to information movement and inter-information correlation) to improve the capacity of TCM information acquisition, conversion, dissemination and application. If the physical world is made of matter, energy and information, TCM has focused on the studies of information rather than matter or energy since ancient times. That is to say, TCM has focused on studies of the changes of human body information rather than the changes of human material and energy. Similarly, the research subject of information science is also information rather than material and energy, and its development will provide methods and techniques for processing of TCM information. Therefore, TCM informatics, as an emerging cross discipline of TCM and informatics, is destined to play a significant role in promoting the development of TCM. As TCM information has four distinct characteristics, namely epistemological information, phenomenon information, overall information and time information, and these characteristics overlap with the three characteristics of "Big Data", namely integrity data, fuzzy data and correlation data, so the advent of the information age characterized by "Big Data" is bound to create good opportunities for the development of TCM informatics.

SIMILARITY BETWEEN THE "INTEGRITY" CHARACTERISTIC OF "BIG DATA" AND THE "OVERALL INFORMATION" CHARACTERISTIC OF TCM INFORMATION

"Big Data" is not a definite concept. Initially, this term refers to the amount of information to be processed is
so large that it has exceeded the amount of memory that can be used in general computer data processing. A series of articles in the special issue of "Big Data" of the journal Nature in September 2008 was a start for the term to be gradually accepted in the industry. In 2011, a report from the McKinsey Global Institute caused huge repercussions on the Internet. In March 2012, the Obama Administration announced the "Big Data Research and Development Initiative" and will invest $200 million in the project. This means a great concern and attention of the United States government to "Big Data" and marks the advent of the age of big data. "Big Data” refers to things one can do on a large-scale data base that cannot be done on a smaller one. At present, "Big Data" is mostly used in data generated in people’s daily activities. The core of "Big Data” is prediction. That is, mathematical algorithm is applied on to predict the possibility of things. One of the advantages of "Big Data” is to show concern over "integrity” rather than "sampling”. The traditional studies often infer populations through sampling, but as only samples but not populations were studied, the studies only focused on verification of prior assumptions but could not find the new relationship that had not been assumed. In the age of "Big Data”, we can analyze more data, and sometimes we even can handle all data associated with a particular phenomenon without relying on random sampling. In this way, we can more quickly and easily identify problems and thus pay more attention to the details that cannot be identified through small data studies.

TCM information is the information under the completely open environment, which pays attention to exchange and contact between human and nature or society. For example, in treatment of diseases with TCM, therapeutic measures should be suited to the time, place and individual conditions, including information exchange, connection and influence between the natural environments, social environments and human body. In terms of etiologic information of TCM, it covers the action and reaction of geographical environment, climate characteristics, social relations and mental activities on human body. Therefore, from the perspective of its attributes, TCM information is the open information.

The openness of TCM information decides its systematicness and integrity. TCM information includes the contents such as generation and change of human body information, impact of natural and social information on human body and its changes as well as the mutual impact and relevant change of extrinsic information of human body, natural information and social information, and it relatively emphasizes the overall changes of mutual influence of systems (including sub-systems). The integrity of TCM information is reflected in two main characteristics: (a) TCM information is overall information, including the extrinsic information reflecting the changes in overall functions of human body, the natural information related to the changes in overall functions of human body, including the natural environments such as seasons, days and nights, climate, regional conditions and substances such as food and medicine, as well as the social information related to the changes in overall functions of human body, such as the cultural and social relationships. It concerns about the overall information on human body, nature and society. (b) TCM information attaches relative importance to duration. Generally speaking, time is the "sustainable” property of existence of matter, and space is the "extensive” property of existence of matter. In terms of the contents of TCM information acquired, the information on human body acquired is the information on overall changes manifested on human body surface, including the overall change information forming in the long-term genetic, social and natural integration. For example, TCM emphasizes treatments based on syndrome differentiation and pays attention to changes of syndromes. Syndrome means the response state of essential organic connections at a certain stage in disease process, such as location, cause, nature and trend of disease as well as body resistance to disease. It emphasizes the stage and continuity of time and rarely acquires human disease information through the spatial segmentation concept of organic body tissue change. In terms of the understanding of Chinese herbal medicine, it emphasizes on basic information such as nature, flavor and meridian tropism of medicines, resource information such as place of origin, genus and medicinal position as well as dynamic change information demonstrating after action with human body such as processing, compatibility of medicines, functions and indications, and rarely explores the spatial attributes of medicines such as chemical composition, chemical structure and specific targets of effects. Therefore, TCM information attaches more importance to continual change of system in time and rarely carries out spatial segmentation. This is extremely similar to the characteristics of "Big Data” that focuses on the integrity rather than sampling.

**SIMILARITY BETWEEN THE "FUZZINESS" CHARACTERISTIC OF BIG DATA AND THE CHARACTERISTIC OF "EPISTEMOLOGICAL INFORMATION" OF TCM INFORMATION**

The second essence of "Big Data” is its "fuzziness". In the age of small data, the pursuit of accuracy is reasonable. As little data is collected, data need to be as accurate as possible, which still applies to some things today. However, in the age of "Big Data”, as big data can
be offset the erroneous data in small data, for some things, to quickly get a rough outline and development context is much more important than strict accuracy. In this way, people no longer need to know and understand the world based on an assumption which means an assumption on the generation mechanism and intrinsic mechanism established for a phenomenon. It will be enough as long as you can grasp the general direction of development. Certainly, this does not mean to completely abandon accuracy, but no longer be addicted to it. Ignoring appropriately the micro-level accuracy will bring a better insight at the macro level.

TCM information is epistemological information rather than ontological information, the characteristics of which are related to the subject, i.e. it is subject-object fusion information. TCM information cannot exist independently from the subject. For example, TCM symptom information is expressed in the form of "thirst, distending pain, red tongue, thin pulse etc.", which needs the perceived participation of the subject rather than "temperature of 38°C and blood pressure of 120 mm Hg" and other objective manifestations. Therefore, from the perspective of information level, TCM information is information at the epistemological level. TCM information is acquired mainly by means of "feeling" and "knowing". "Feeling" and "knowing" are two TCM information acquisition processes. "Feeling" is acquisition of objective information and "knowing" is subjective identification of information. Therefore, TCM information is the information characterized by unity of subject and object manifested through a composite mode. This kind of information is no longer simply objective information itself, but information superimposing subjective perception with integrated features. TCM information acquired is not entirely objective information but information integrating the subject and the object, and it is comprehensive information integrating a number of subjects. For example, information on human body is generally acquired by the four diagnostic methods of inspection, auscultation and olfaction, inquiry, and pulse-feeling and palpation. Inquiry is to acquire the disease information through the description of symptoms by patients, and the symptom information transmitted by patients, the organic reaction described by patients through their own feeling and the severity is related to the subjective feeling of patients.

Inspection, auscultation and olfaction, and pulse-feeling and palpation are the ways doctors get signs information of patients, and the judgment of doctors on patient’s color, smell, pulse, etc. has a close relationship with their subjective perception. If you are a re-visiting patient, then the diagnostic information, prescription information etc given by the doctor who carried out diagnosis and treatment previously will be integrated into the information acquisition process and form the TCM diagnosis information from objective information that integrates several pieces of subjective information. Subjective information is from objective information, affected by the factors of the subject itself and limited by the cognitive ability (including perceptual ability and process ability) of the subject. The sensitivity and identification degree of sense organs and the ability of memory and data processing of human body directly affect the ability to collect information from body surface, so it is only possible to make a qualitative judgment and it is difficult to make a quantitative judgment. As TCM information, as a kind of stable and unstable information integrating the object and the subject, contains a large amount of noise and redundant information, the TCM information acquired is larger than the information of the object itself and has relatively poor accuracy. This is consistent with the "fuzziness" characteristic of "Big Data".

SIMILARITY BETWEEN THE CHARACTERISTIC OF EMPHASIS ON "CORRELATION" OF BIG DATA AND THE CHARACTERISTIC OF "INCLUDING THE ESSENTIAL PHENOMENON INFORMATION" OF TCM INFORMATION

The third essence of "Big Data" is that it is no longer interested in finding causal relationships but shifts its attention to identification and use of correlation. Correlation may not be able to accurately explain why something happens, but it can suggest that this thing is happening. Correlation helps people analyze a phenomenon through identifying useful associated things rather than revealing its internal operation mechanism. The prediction based on the correlation analysis method is the core of "Big Data". As long as we can identify a significant correlation between the two phenomena, we can create enormous economic and social benefits, while it can be left over to scholars to identify why they have correlation.

The characteristic of correlation reflected in TCM is that TCM information is the phenomenon information containing the essence. Essence and phenomenon are a pair of categories revealing the correlation between internal relation and external manifestation of things. TCM diagnosis information contains the overall state of change of stable and unstable relationships caused by partial internal change of human body forming in the long-term genetic, social, and natural environment.

Essence is the internal relations of things, and it determines the nature and the development trend of things. Phenomenon is the external connection and surface characteristic of things, and it is the external manifestation of the essence of things. Essence has simplicity and profoundness and is relatively stable. Phenomenon
is rich and specific, and it is instable and prone to change. Chinese herbal medicine contains long-term genetic, and natural environment information. Furthermore, after processing and compatibility under the guidance of TCM theories, Chinese herbal medicine exhibit the stable and unstable related overall state of prescription compatibility and Chinese herbal machine interaction manifested through the overall change of human body. For example, TCM syndrome information is ontogenic information at a certain stage, so it is instable, local and individual phenomenon information. However, in diagnosis and treatment, the same treatment method will be used in response to the same syndrome information occurring in different individuals at and different stages, which shows that the rich and ever-changing syndrome information contains internally relatively stable essence information. Modern medicine highlights evidence-based medicine, strengthens the groupment of diseases and group effectiveness of medication and explores the common stable information in similar phenomena occurring in groups. The studies of pathological mechanism mostly explore the occurrence of diseases and the acting targets of drugs from the perspective of proteomics, molecular biology and genomics, and such information shows the relatively stable inherent relations of disease occurrence. Therefore, from the law of TCM information grasped, TCM information is phenomenon information containing essence information.

In the age of “Big Data”, compared with other disciplines, TCM data has no mass characteristics in quantity and is still unable to reach the exabyte, petabyte and zetabyte levels (mostly at the gigabyte and terabyte levels), but it has the distinct characteristics of “all data”. The so-called “comprehensive data” means that, during the observation of an object, all the collectable information on the object should be collected, including stable or instable essence and phenomenon information. In short, all the information should be collected and digitized. These data may not be up to the mass concept, but it is complete to a specific thing and can help us identify the detailed correlation between things. We call them “comprehensive data” (which is the high dimensional data describing the same things from various dimensions). When observing an individual or study a topic, TCM always tries to collect all the information that can be collected, which is called "phenomenon information" containing essence and phenomenon of information and stable and unstable relations. Therefore, when it is digitized, the data formed is a kind of "comprehensive data", which is a kind of data that has the characteristics of "Big Data" and can be processed by methods for processing "Big Data".

CONCLUSION
In summary, the characteristics of "Big Data" are very similar to the characteristics of TCM information. Since the early stage of its creation, the information collected by TCM has been "comprehensive data", i.e. populations rather than samples. This is reflected in that it stresses the combination of nature, earth and human being while grasping individual information, and needs to grasp all the relevant information on individuals, which is one of the characteristics of TCM information we emphasize - the integrity of information.

As the information collection methods in TCM are relative backward, it never demands too much in accuracy, and it has always highlighted the efficiency of solving practical problems. This is the phenomenon information characteristics of TCM information we emphasize. The basis of treatments based on syndrome differentiation in TCM is to identify the correlation of all kinds of information manifested by individuals and judge the syndromes, treatment and prescription medication based on this correlation rather than exploring the causal relationship. This is why when an infectious disease breaks out, TCM can always be able to calmly deal with it, and this is also the embodiment of the phenomenon and integrity characteristics of TCM information. However, before the advent of the age of "Big Data", we lack the methods and tools to process this kind of "comprehensive data". Now the advent of the age of "Big Data" is bound to provide us with many new data processing tools that enable us to process more effectively "comprehensive data" TCM is facing.

REFERENCES
8. Viktor MS, Kenneth C. Big data: a revolution that will transform how we live, work, and think. British: Hodder.
and Stoughton, 2013: 10.


11 Hu W. Thinking of the dialectic relation between the phenomena and the essential. Xue Li Lun 2011; (14): 41-42.

