INFORMATION STUDY

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
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

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".

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.[1-3]

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,[4,5] 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.[6] In March 2012, the Obama Administration announced the "Big Data Research and Development Initiative" and will invest $200 million in the project.[7] 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.[7] 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 subsystems). The integrity of TCM information is reflected in two main characteristics: (a) TCM information is overall information, including the extrinsic informa-

tion 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 emphases 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.[8,9] 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