Communication Science for the Big Data Era

Naonori Ueda

Abstract

This set of Feature Articles, on the theme of communication science that connects information and humans, introduces the research efforts and achievements of NTT Communication Science Laboratories for the *big data* era. Around the world, digital data is being created and stored in enormous quantities and the era of big data is now upon us. For research and business applications, it is vital to have some way of extracting new value from these huge and diverse data resources.

1. Introduction

The spread of social media and mobile information devices has accelerated in recent years, and huge quantities of digital data are now being produced and stored throughout the world. It reached 1.8 ZB (zettabytes: 10^{21} bytes) in 2011 and is expected to increase another fiftyfold over the next 20 years. We can safely say that the era of *big data* is now upon us. The use of big data as a buzzword can be traced back to a May 2011 report "*Big data: The next frontier for innovation, competition, and productivity*" by MGI (McKinsey Global Institute, the research division of McKinsey & Company) [1]. This report suggested that big data can lead to the creation of huge monetary value in diverse fields including healthcare, economics, and manufacturing. For example, it is estimated that the utilization of big data helps the US healthcare sector save $300 billion each year. There is also a growing trend towards the spread of cloud environments and open-source large-scale parallel distributed computing environments such as *Hadoop*, so big data is clearly set to become a driving force behind information and communications technology (ICT) innovation in this century [2].

The term big data encompasses not only large and highly diverse (atypical) data that is difficult to accumulate, edit, and store in conventional databases, but also the substantial benefits to industry and society that are gained by deriving new value from it. Or to put is another way, big data is not just about how much data you have, but also about the scenarios you use to gain new value from it and the techniques you use to implement these scenarios. For example, Indiana University, USA, has devised an analysis method that can make stock market forecasts with 87% accuracy by analyzing 9.8 million tweets from 2.7 million Twitter users. Its ability to predict the stock market from tweets can be attributed to the easy availability of vast quantities of data. This example indicates that the big data era is characterized by the profitable application of completely new analysis methods. In other words, businesses must somehow derive valuable information from the serendipitous effects of combining large quantities of seemingly unrelated information.

On the other hand, researchers need core technologies that can extract and merge latent information from large and diverse data sources in order to make predictions. In addition to these core technologies, it is also important to construct a next-generation ICT infrastructure to provide a foundation for big data. This infrastructure should support a cyber-physical system that allows computing resources to be used in cyberspace to perform advanced analysis of highly diverse data, and it should use the results to promote a real-world system in order to construct a highly efficient social system. In this set of Feature Articles, as some of the research currently under way at NTT Communication Science Laboratories (NTT CS Labs), we introduce the core technology and infrastructure technology that is essential for the big data era.

2. Cyber-physical system core technology and infrastructure technology

Machine learning techniques have been attracting attention for big data analysis applications. Simple statistical analysis is insufficient for dealing with a huge and diverse set of data, so to support a cyber-physical system we need techniques that can learn the underlying models of data (data generation mechanisms) from previous data and can use these models for learning in order to predict future events. However, the majority of conventional machine learning techniques are targeted at predefined tasks such as regression analysis and classification problems. In the big data era, what we really need is a technique that can take in a diverse variety of data and extract the latent data structures (correlation relationships and causality relationships) that exist within it. In other words, we need hypothesis discovery techniques rather than hypothesis testing techniques. At NTT CS Labs, with this in mind, we are researching relational data mining techniques based on a machine learning approach, and we are also conducting empirical studies using real data from sources such as Twitter. This is described in detail in the Feature Article "Extracting Essential Structure from Data" [3] in this issue.

As more information is made public in the big data era, issues of security and privacy become more apparent. At present, encryption and authentication techniques are used to prevent the disclosure of or tampering with data transmitted across networks. These security techniques are based on random numbers, but since most random numbers are currently pseudorandom numbers generated by numerical series calculations, it is easy to predict the random numbers that will be output if the initial value (seed) and generating formula are leaked. Thus, from this viewpoint, absolute security is not assured. An alternative method called physical random number generation produces random numbers from physical phenomena. For some years now, we have been working on a method for generating physical random numbers from the chaotic behavior of semiconductor lasers. In 2008, we achieved the world's fastest physical random number generation rate of 1.7 Gbit/s. In 2010, we implemented this technique in a miniaturized device by taking advantage of optical integrated circuit technology. Since this method is theoretically guaranteed to provide unpredictable results, it can be said to be a core security technology that provides the ultimate in security. Details can be found in the Feature Article "Fast Physical Random Number Generation Using Semiconductor Laser Chaos" [4].

Attention is also being drawn to machine-to-machine (M2M) systems in which machines connected to computers in a cyber-physical system can implement intelligent control through the mutual exchange of information without any human intervention. Sensor network technology is therefore an important part of the technical infrastructure for such systems. In the future, sensor network technology will not only make use of information derived from the sensor values themselves, such as temperature readings and chemical concentrations, but also require advanced information processing capabilities such as the ability to infer the occurrence of events from the information obtained from multiple sensors. At the CS Labs, we have developed a technique for recognizing human behavior from multiple sensors such as accelerometers, cameras, and microphones. We are also researching a method for gathering information efficiently from a large number of wireless sensor nodes. Details can be found in the Feature Article "Information Processing of Sensor Networks" [5].

3. Basic research exploring the essential nature of communication

Needless to say, conversations between humans form the basis of communication. However, a conversation is not just an exchange of linguistic information; other factors such as feelings of sympathy and antipathy (nonverbal information) also play important roles. We are researching a system and conversation scene analysis technique that utilizes techniques for extracting verbal information (speech recognition) and nonverbal information (emotion recognition) in a conversation to ascertain the circumstances under which messages are sent (when, where, who, to whom, what, how, and why (known as 6W1H)). We are working on a system called MM-Space that reproduces human head movements as physical movements of a display screen in order to allow a remote conversation participant to be represented at a different location. This system is described in detail in the Feature Article "MM-Space: Recreating Multiparty Conversation Space by Using Dynamic Displays" [6].

In addition to simple enhancements of telecommunications technology in a cyber-physical system, it is also essential to implement a communications-based society that lets anyone enjoy the convenience of this

advanced technology in a safe, secure, and enriching way. To this end, we feel that it is important to promote not only information science but also research on its human science and social science aspects. Ever since its establishment, the CS Labs has been conducting research aimed at clarifying the nature of communication. ICT must adapt to human society. To create a richly featured communication environment, we cannot simply confine our studies to information per se; we must also study the information processing mechanisms that humans use to send and receive information. Recently, we have been trying to clarify how the human brain assimilates information gathered by the senses, and we have made new discoveries related to the interaction between sight and hearing. Details can be found in the Feature Article "Hearing Sound Alters Seeing Light" [7].

References

[1] MGI (McKinsey Global Institute, the research division of McKinsey & Company), "Big data: The next frontier for innovation, competition, and productivity," 2011.

[2] H. Shinohara, "R&D to Create the Future of ICT," NTT Technical Review, Vol. 10, No. 4, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201204fa2_s.html

[3] K. Ishiguro and K. Takeuchi, "Extracting Essential Structure from Data," NTT Technical Review, Vol. 10, No. 11, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201211fa2.html

[4] K. Yoshimura, S. Shinohara, and K. Arai, "Fast Physical Random Number Generation Using Semiconductor Laser Chaos," NTT Technical Review, Vol. 10, No. 11, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201211fa3.html

[5] T. Suyama and Y. Yanagisawa, "Information Processing of Sensor Networks," NTT Technical Review, Vol. 10, No. 11, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201211fa4.html

[6] K. Otsuka, "MM-Space: Recreating Multiparty Conversation Space by Using Dynamic Displays," NTT Technical Review, Vol. 10, No. 11, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201211fa5.html

[7] T. Kawabe, "Hearing Sound Alters Seeing Light," NTT Technical Review, Vol. 10, No. 11, 2012. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201211fa6.html

Naonori Ueda

Director, NTT Communication Science Laboratories.

He received the B.S., M.S., and Ph.D. degrees in communication engineering from Osaka University in 1982, 1984, and 1992, respectively. He joined the Yokosuka Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT) in 1984. In 1994, he moved to NTT CS Labs in Kyoto, where he has been researching statistical machine learning, Bayesian statistics, and their applications to web data mining. From 1993 to 1994, he was a visiting scholar at Purdue University, Indiana, USA. He is a guest professor at the National Institute of Informatics and the Nara Advanced Institute of Science and Technology. He is a Fellow of the Institute of Electronics, Information and Communication Engineers and a member of the Information Processing Society of Japan and IEEE.