Big Data and AI

Seiki Akama

C-Republic
1-20-1, Higashi-Yurigaoka, Asao-ku, Kawasaki-shi, 215-0012, Japan

e-mail: akama@jcom.home.ne.jp

Abstract: Recently, the area called Big Data is considered to be very important for ICT (Information and Communication Technology). We also find the revival of the so-called AI (Artificial Intelligence), which aims to automate human intelligence. These two areas are closely related. In this paper, we address some interactions.
Key words: Big Data, AI, Rough Set Theory, Quantum Computing

1. INTRODUCTION

Recently, the area called *Big Data* is considered to be very important for *ICT* (Information and Communication Technology). In fact, it can play significant role in many systems. For instance, most of current search engines in *Internet* employ the techniques closely related to Big Data. It is also used in the area of *Databases*.

The idea of Big Data was proposed in the early 2000's; see Laney (2012). In Big Data, 'Big' addresses quantity as well as quality, together with associated technology. We expect to see promising applications to ICT using Big Data from science to technology.

Big Data arise from the need of 3V, namely, Volume, Velocity and Variety. Laney added Veracity to reach 4V. Later, some practical systems to use Big Data have been developed, e.g., MapReduce, Hadoop, and NoSQL.

We also find the revival of the so-called *AI* (Artificial Intelligence), which aims to automate human intelligence. It has long history since the 1950's and gives some practical applications; see Russell and Norvig (1995). The issues of AI include common-sense reasoning, knowledge representation, natural language understanding, learning, and so on.

In this paper, we consider the interactions of these two areas in view of some points, since we believe that it is of importance for the progress of these fields. Below we discuss some interesting interactions.

The plan of this paper is as follows. In section 2, we review Big Data by explaining its basic ideas. Section 3 surveys AI with listing major subfields. In section 4, we discuss the interactions of Big Data and AI and take up two interesting fields, i.e., Rough Set Theory and Quantum Computing. Finally, we give concluding remarks in section 5.

2. BIG DATA

Big Data refers to very large and complex data as well as the associated areas. Historically, it seems to emerge in the early 2000's, see Laney (2012). Now, Big Data is applied to several areas. For instance, many companies attempt to use Big Data for information management including search engines. *Internet of Things* (IoT) is another example in ICT. For IoT, Veracity of Big Data is necessary.

Healthcare is related to Big Data. It is not surprising since Data Mining has been already applied to it. Other applied areas include education, insurance, manufacturing and investment, but we here omit the details.

There are three or four essential features in Big Data, which are called *3V* and *4V*. 3V includes *Volume*, *Velocity*, and *Variety*. 4V extends 3V with *Veracity*. Big Data arise from the need of 3V, namely, Volume, Velocity and Variety. Laney added Veracity to reach 4V.

Roughly, they can be understood as follows:

Volume = Size

Velocity = Speed
Variety = Kind
Veracity = Exactness

Now, we add more words about 3V and 4V. Volume means the quantity of data, i.e., size. In fact, the size of data enhances applicability of data. This is statically justified.

Velocity is the speed of processing data. The processing contains handling, generation, recording, and publishing. Since, Big Data is usually given in real-time, it is very important.

Variety is the kind (type) of data. Currently, by the development of multimedia, there are many types of data, i.e., text, sound, image, video. And some data fuse these data.

Veracity refers to exactness (quality or value) of data. Even if we have huge data, they may contain incomplete data, we need veracity of data. It can improve data analysis.

It is clear that these features deal with 'Big' aspects. Related to these features, Big Data has many applications, including search engine, database and simulation. As mentioned above, there are some practical systems for Big Data.

In 2004, Google published the architecture of Big Data called *MapReduce*, which is based on a parallel processing model; see Dean and Ghemawat (2004). And MapReduce was adopted by *Hadoop* in 2012. These two architectures are now regarded very important in the area.

However, there are some critiques of Big Data. For example, 3V and 4V model of Big Data address computational aspects and lack understandability. Some people pointed out the novelty of Big Data. The idea of Big Data is not new. Similar ideas have been already found in Very Large Database and Datawarehouse.

In addition, the technology of Big Data is closely related to the one of Data Mining (cf. Adriaans and Zatinge (1996)). We have to consider reliability of Big Data technology. Data analysis based on Big Data depends on particular Big Data, and we may reach biased results.

Now, many countries including USA and UK promote the Big Data technology. In 2012, the Obama administration announced the initiative of R & D of Big Data. It seems obvious that to have an initiative of the Big Data technology is very useful for a country.

3. AI

AI (Artificial Intelligence) is the area of developing a system with human intelligence and its history goes back to the mid of 1950's; see Russell and Norvig (1995) for details. Now, AI enters a new phase in the sense that practical AI systems. In particular, deep learning plays a crucial role in the recent developments.

There are several subareas of AI. We can list some of them as follows:

Common-sense Reasoning
Knowledge Representation
Natural Language Understanding
Machine Learning
Vision
Robotics

Common-sense reasoning treats human daily reasoning. There are many types of human reasoning. For instance, by mathematical reasoning we can prove theorems of mathematics. In practice, automating mathematical reasoning is called *theorem-proving*. However, it is known that to automate common-sense reasoning is more difficult.

Knowledge representation is the area of representing human knowledge. There are many such frameworks, e.g. semantic networks, frame and script. Knowledge bases extend databases in that they can store human knowledge.

Natural language understanding (also called natural language processing) is to understand our daily language and it is a basis of human interface in AI systems. To construct an AI system, it has to understand our languages. The so-called *machine translation* is one of the important subjects in natural language understanding.

Machine learning studies methods of learning from many data by a computer. There are different types of learning methods and they can be handled by the so-called neural networks which emulate neurons of our brain. *Deep-learning* can be classified as a method of machine learning.

Vision concerns how a computer can understand image data and is closely related to *image processing* in computer science. Some AI systems need techniques in vison since they should understand image data.

Note that many practical computer vison systems. have been developed

Robotics is to develop a (human-like) robot which imitates a human. A robot must do human intelligent activities. Namely, it has to understand languages, images and sounds, to perform many types of reasoning, to communicate with human, and so on. In this sense, robotics uses many AI techniques.

There are other subareas including Game, Planning and Evolutional Computing, but we here omit their explosion.

We know that there are basically two different approaches to AI in a certain sense. One is symbolic which processes our knowledge as symbols. The other is numerical which represents our knowledge numerically. They have both merits and demerits. However, we need to integrate both approaches, because they are complementary in the study of AI.

4. INTERACTIONS

There are intimate connections between Big Data and AI. We consider the interactions of these two areas in view of some points, since we believe that it is of importance for the progress of these fields. Belo, we discuss some interesting interactions.

First, some AI approaches clearly need Big Data. For example, deep learning can establish some types of our intelligent activities. But for doing so, we have to learn Big Data by the method.

Second, to deal properly with Big Data in a computer, we must use very fast computation. There are some hardware solutions to very fast computation. This can dramatically enhance the usability of Big Data.

The first interaction can be solved by using some AI technique. In fact, such techniques have been studied in the area of Data Mining in databases (or knowledge bases); see Adriaans and Zantinge (1998). Its aim is to extract useful information from many data. Although many Data Mining techniques are well known, the so-called *rough set theor***y** can serve as a basis for Data Mining to model (vague) knowledge; see Pawlak (1991).

We know that rough set theory can be seen as a generalization of standard set theory. It is important in that it has many applications for various areas with firm mathematical foundations and it seems to be of use to handle Big Data. For example, Akama et al. (2018) addressed reasoning based on rough set, which can cover various types of common-sense reasoning.

The second interaction means the need of new type of hardware. In other words, we need non-von Neumann computers. This is because conventional (von Neumann) computers have limitation in efficiency.

We believe that the so-called *quantum compute***r** is attractive in this context; see Akama (2015). Quantum computers enable us to give very fast computation, but at the present time we never see *practical* quantum computers.

From these considerations, the interaction of AI and fast computation seems to be very important. We know that little work has been done in this line. And, it is extremely difficult to explore it.

5. CONCLUSIONS

We summarize that Big Data is obviously a promising field for ICT. We pointed out that there are at least two interactions of Big Data and AI. Indeed, we could investigate other interactions, but we think that the interactions considered in this paper is interesting and essential.

Finally, we notice that the current Big Data technologies appear to face some limitations. For instance, they cannot appropriately cope with the notions of time, uncertainty, incompleteness, inconsistency, and so on.

Uncertainty, incompleteness and inconsistency of Big Data are closely connected with Veracity. Temporal aspects of Big Data should not be neglected, because Big Data involve time. Surely, such approaches as agents and probabilistic models can lead to interesting solutions for overcoming difficulties mentioned above.

REFERENCES

[1] P. Adriaans and D. Zantinge, *Data Mining*, Addison-Wesley, Reading (1996)

[2] S. Akama, *Elements of Quantum Computing*. Springer, Heidelberg (2015)

[3] S. Akama, T. Murai and Y. Kudo, *Reasoning with Rough Sets*. Springer, Heidelberg (2018)

[4] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters, Google Inc. (2014)
[5] D. Laney, The importance of 'Big Data': A Definition, Gartner (2012).
[6] Z. Pawlak, *Rough Sets*, Kluwer, Dordrecht (1991)
[7] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall Saddle River (1995)
[8] http://hadoop.apache.org/ (Hadoop)