BD 1-2 The fourth paradigm of scientific research

1

Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering

School of Computer Science, in Beijing Institute of Technology, in this session, we will discuss about The fourth paradigm of scientific research, which is data intensive research methods.

2

Jim gray, he is the founder of Relational database and he is also a Nautical sport enthusiast.

Jan,28,2007 he was disappeared on the sea forever, and just 11 days before that, Jan,17,2007

At the NRC-CSTB, National Research Council-Computer Science and Telecommunications Board meeting held in Mountain View, California, he delivered his last lecture, "The Revolution of Scientific Method," Divide scientific research into four types of paradigms. (Paradigm, a certain norm that must be followed or a routine that everyone uses), Followed by Experimental induction, model deduction, simulation and data-intensive scientific discovery (Data-Intensive Scientific Discovery). Among them, the last "data-intensive" is what we now call "scientific big data".

3

1） Thousand years ago, science was Empirical, people do the science by describing natural phenomena. Like in ancient China, farming was done by observing celestial phenomena.

2 ）last few hundred years, theoretical branch, using models, generalizations.

From the primitive drilling wood to make fire, it developed to the early stage of scientific development during the Renaissance， represented by Galileo.

It is difficult to complete a more accurate understanding of natural phenomena. Scientists began to try to simplify the experimental model as much as possible, remove some complex interference, and leave only the key factors (for example, “smooth enough", " enough long time", "sufficiently thin air" and so on in our physics learning, which is the inexplicable condition description), and then sum up through calculus, this is the second paradigm of scientific research.

This research paradigm lasted until the end of the 19th century, and it was perfect. Newton's three laws successfully explained classical mechanics, Maxwell's theory successfully explained electromagnetism, and the classic physics building was magnificent.

But after the emergence of quantum mechanics and the theory of relativity, which focused on theoretical research, surpassing experimental design with extraordinary brain thinking and complex calculations. As the difficulty of verifying theories and higher and higher economic investment, scientific research began to appear incapable.

3). In the middle of the 20th century, von Neumann proposed a modern electronic computer architecture, and the model of using electronic computers to simulate scientific experiments has been rapidly popularized. People can deduce more and more complex phenomena by simulating.

Typical cases include simulated nuclear tests and weather forecasts. As computer simulation increasingly replaces experiments, it has gradually become a conventional method of scientific research, which is the third paradigm of scientific research.

A major step in the use of computational science was taken when it was realized that nonmathematical models could be simulated as well as mathematical ones. That is, rules other than mathematical equations could be incorporated into computer programs and stepped through simulated time "to see what happens" to the state of the model. For example, traffic simulations can predict where bottlenecks and other problems may occur. Even simple rules can generate complex behavior, making it very valuable to observe the output of the simulation to understand what the rules imply. On the other hand, given a set of data, either observed or human-entered, can computers help us find rules to explain the data?

4) The future development trend of science is that with the explosive growth of data, computers will not only be able to do simulations, but also analyze and summarize, and generate theories. The data-intensive paradigm should be separated from the third paradigm and become a unique scientific research paradigm. This method of scientific research is called the fourth paradigm. The fourth paradigm of scientific research and the third paradigm of scientific research. both use computers to perform calculations. What is the difference between these two paradigms of scientific research?

“What is the scientific problem?", "What are the scientific hypotheses?" This is first to propose possible theories, then collect data, and then verify them through calculations.

The fourth paradigm based on big data is to first have a large amount of known data, and then calculate the previously unknown theory. In the "Big Data Era" written by Victor Meyer-Schoenberg, it is clearly pointed out that the biggest change in the big data era is to abandon the desire for causality, which are Cause and result, and focus on correlations instead.

(The paradigm is "human brain + computer", the human brain is the protagonist主角, the lead, and the fourth paradigm is "computer + human brain", the computer is the protagonist. However, to discover the causal connection between things, in most cases which is always difficult. The causal connection we humans derive is always based on past knowledge, obtain a "deterministic" mechanism decomposition, and then build a new model for derivation. However, this kind of past experience and common sense may be incomplete, and may even neglect important variables intentionally or unintentionally.)

4

According to the existing mechanism understanding, the formation of haze weather is not only related to the chemical composition of the Source atmosphere, but also related to the terrain, wind direction,

temperature, Humidity and meteorological factors.

Only these limited parameters have exceeded the ability of conventional monitoring, and can only be simplified and artificially removed some seemingly unimportant factors, and only some simple parameters are retained. Will those seemingly unimportant parameters play a vital role under certain conditions?

If we consider the spatial heterogeneity of different parameters, is the spatial distribution of these weather stations reasonable and sufficient? From this point of view, if we can obtain more comprehensive data, we may be able to truly make more scientific predictions. This is the starting point of the fourth paradigm, and perhaps the fastest and most practical way to solve problems.

So, how will the fourth paradigm be studied? In the era of rampant mobile terminals and rapid development of sensors, the trend of the future seems to be in sight. Now, our mobile phones can monitor temperature and humidity, and can locate spatial coordinates. Soon there may be sensing devices that can monitor atmospheric environment chemistry and PM2.5. These mobile monitoring terminals increase the spatial coverage of the measurement, and at the same time a huge amount of data has been generated. Using these data, we can analyze the causes of the haze, and finally make a better prediction.

The emergence of this kind of massive data not only exceeds the understanding and cognitive ability of ordinary people, but also brings huge challenges to computer science itself. Therefore, when the data volume of these large-scale calculations exceeds 1PB, the traditional storage system has been unable to meet the read and write needs of massive data processing, and the bottleneck of data transmission I/O bandwidth becomes more and more prominent.

However, simply dividing the data into blocks cannot meet the needs of data-intensive computing, and is contrary to the original intention of big data analysis. Therefore, the biggest problem many specific researches at present faced is not lack of data, but too much data without knowing how to deal with it.

5

Another example of **Abandoning causality, taking correlations.**

2004, In Wal-Mart shopping list, Wal-Mart recorded all the useful information, including consumption amount, items in the shopping basket, the specific time of purchase, and even the weather of the purchase. And they find whenever the seasonal hurricane comes, not only the sales of flashlights will increase, but the sales of egg tarts will also increase.

Therefore, when seasonal storms come, egg tarts close to hurricane supplies, and increased sales. It is hard to explain the reason.

In other words, as long as you know the "what", you don't need to know the "why". This has overturned human thinking conventions for thousands of years, and is said to pose a brand-new challenge to human cognition and the way of communicating with the world. Because humans always think about the causal connection between things, and are not so sensitive to data-based correlations; on the contrary, computers can hardly understand causation by themselves, and are very good at correlation analysis. So, we can understand this.

In the "Big Data Era" written by Victor Meyer-Schoenberg, it is clearly pointed out that the biggest change in the big data era is to abandon the desire for causality, which are Cause and result, and focus on correlations instead.

6

In this session, we learned the fourth paradigm of scientific research, which is based on big data and first have a large amount of known data, and then calculate the previously unknown theory. And the biggest change in the big data era is to abandon the desire for causality, which are Cause and result, and focus on correlations instead.

**Thank you for your attention, if you have any question, feel free to connect me.**