A Conversation with Professor Bole Shi

Bole Shi
Fudan University
Shanghai 200433, China
bshi@fudan.edu.cn

More than twenty years have passed since the concept of KDD has been coined in the early 1990's. In these 20 years, KDD has brought both nice surprises and frustration for academic research and industrial applications.

Looking back in the past, we note that in the past 20 years, data mining has been a popular topic in academic research, but in industrial applications, we have heard few exciting examples after the "beer and diapers" one. This is in part due to the general pattern of technological advances, in which the advance of academic research is ahead of industrial applications. On the other hand, we cannot ignore the fact that in this past 20 years, in the world, especially in China, data processing and analytics have focused more on data accumulation and integration phase, but the need for wide-spread data mining has yet to come. Currently, after so many years of accumulation and preparation, whether in China or in the world, the scale and complexity of available data has far exceeded our expectations, and further development of data mining will finally take the central stage of research and development. In this view, we expect to see more advance in data mining in terms of its ability to handle greater data types, larger scales, diverse business needs, application varieties and cross-disciplinary integration.

- **Scale and Variety**

In the academic circle around 20 years ago, data mining and knowledge discovery were still limited by scale to around million record levels. However, in the recent 10 years, new data collection methods have developed rapidly, especially with the development of Internet based services and businesses. New Internet based user-interaction models are increasingly accepted by users at large, such that data collection is no longer limited to small-scaled sampling based on dedicated experts. New user generated data such as Twitter/Weibo, social networks and other new applications make it possible to massively and rapidly collect the user data, which in turn boosted the need for large-scale data mining on big data. Recent studies have revealed that many effective methods on small-scale data cannot maintain their advantages on large-scale data. In 2009 China Database Conference, researchers from Google research showed that some techniques believed to be outdated previously can demonstrate their superior performance on large data. With the development of cloud computing, new computational platforms and framework are emerging. A natural question for us to ask is: should we develop novel methods that are targeted for big data?

- **Business Needs**

Around 10 years ago, a frequently discussed topic among data mining researchers is how to get their developed methods to be applied to the real problems. In many businesses and enterprises, when discussing issues on KDD with CIOs, a feedback often heard was: "we currently have no need for data mining." Many large businesses such as banks, a central issue at the time was to integrate different datasets and databases into a data center. Thus, they would not have enough resources to spend on data analytics, and many take a sideline when the issue comes to data mining. Data mining at the time is equivalent to setting up data warehouses, performing OLAP analysis. There were companies that developed software products in data mining, especially from large companies such as IBM, who developed Intelligent Miner products, as well as companies such as SAS and SPSS, etc., but still these products were more academic than industrial, and at that time, they were not popular yet on the market. However, these products have served an important purpose of setting up the groundwork for what's to come. Today, with the emergence of new concepts such as Intelligent Earth, Internet of Things, etc., the needs for integrating diverse data efficiently and effectively have become real, and by the same token, the popularity of these concepts is evidence that the market is getting ready for wide spread data mining. We believe that after this long period of being dormant, the real need for data mining will truly and rapidly arrive.

- **Application Background**

Ten years ago, a large part of the data mining research work is relatively broad and general, where many research projects without considering the specific domains and background knowledge. we carried out association analysis on traffic accidents somewhere in China, but the system concluded that black cars were likely to be involved in accidents. A deeper analysis revealed that cars come in different colors more in the recent years. Around 10 years ago, black cars were the majority, causing the association rule to be strong. This story shows that data mining without actual application background knowledge is likely to be misleading, or even 'shocking'. Take timing series analysis as another example. Many research works in this area tend to use the stock market analysis as the motivation. However, without understanding the basic operations of the stock market, analytical results tend to return only `non-surprising' results that are of no use. However, the situation is changing. In recent years, with the deepening of our understanding on data mining technology, more and more experts in the field have gradually mastered data analytic skills. By combining with domain knowledge, many data mining techniques have become important technological support for many key industries, including product recommendation, search services, opinion analysis, etc.

- **Cross-disciplinary Integration**

Data mining integrates several research areas, including databases, statistics, artificial intelligence and other technologies. In the decades of development and research, the data mining technology has absorbed a variety of technology and ideas. Starting from the initial simple types of data analysis in data mining, this field has

gradually extended to analysis of data involving image, video, audio and other multimedia data. The structures being analyzed evolved from simple tuples of data and simple data structure to more complex data structures including trees, graphs as well as other complex data structures. The data to be mined have diversified to multiple heterogeneous types. These developments also presented new challenges to data mining research and applications.

Compared to 10 years ago, we now have much better opportunity and more powerful tools. Of course, we also face more difficult problems and more new challenges. What we need is more in-depth domain knowledge and do more solid research work. But we must bear in mind that data mining is an application-driven field, and as such it cannot deviate from real-world applications and practice.

About the author:

Bole Shi is a professor at Fudan University, China.