A Conversation with Professor Shan Wang et al.

Shan Wang, Cuiping Li and Hong Chen

Key Laboratory of the Ministry of Education for Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China, and

School of Information, Renmin University of China, Beijing 100872, China

1. Please share with us your view on the history and important milestones of the Chinese KDD research and application areas.

Compared to other countries, the research of Data Mining (DM) in China began a little later. In early 1990s, China's NSFC (Natural Science Foundation of China) started to support research projects in data mining field. In 2001, Jiawei Han's book "data mining: concepts and techniques" was introduced into China by the Machinery Industry Press and was translated into Chinese shortly.

One year later, the Chinese government launched the Dragon Star Plan [1]. It is aimed to organize a group of oversea Chinese scholars to come back to China to teach U.S. graduate level courses systematically on a particular area around the universities in China. These scholars usually had got some achievements and had certain positions in the United States' Academia. With the support of this plan, since 2002, quite a few world-famous DM researchers such as Jiawei Han, Qiang Yang, Jian Pei, Hui Xiong were invited to China to teach DM courses in several universities. All these (the book and courses) significantly promoted the popularization of DM technology in China.

In 2005, organized by the China Computer Federation (CCF) and the China Artificial Intelligence Association, the first China Conference on Data Mining (CCDM) was held in Beijing. In the same year, with the support of NSFC, the first international conference of Advanced Data Mining and Applications (ADMA) was held in Wuhan. These meetings provided a communication and cooperation platform for the vast DM researchers.

At present, DM research in China is blooming. Many research institutes and universities, including Tsinghua University, Peking University, Renmin University of China, Harbin Institute of Technology, Northeastern University, Fudan University, and Institute of Computing Technology of Chinese Academy of Sciences, do a lot of work on DM research and application, and gain many significant research results.

In 2002, for the first time, researchers from the Mainland China published three papers on the KDD conference, which was a breakthrough in the history of Chinese DM research. After that, almost every year, Mainland Chinese scholars' papers can be seen in KDD proceedings. Since 2007, the paper count has grown gradually and reached to 13 in 2010. For the other two important DM international conferences ICDM and SDM, the situation is similar. Not only is the quantity but also the quality of the Mainland Chinese scholars' paper improved every year. For example, in 2010, authors from Renmin University of China won the Best Paper Award of the SDM Conference, one of the top three conferences in data mining area.

After years of continuous research, many mainland DM researchers have come forward with more results. Some of these researchers are very young. For example, Dr Jianyong Wang from Tsinghua University has made a remarkable achievement in DM algorithm research with his paper; according to Google Scholar [2], the total citation of his papers on data mining has exceeded 3000.

In summary, the development of DM in China can be divided into three stages. In the first stage, the

DM knowledge were brought in through the book and courses, which accelerated the popularity of the data mining knowledge in China; in the second stage, both theoretical innovation and the introduction of knowledge are important. Based on the digestion of the knowledge, it is possible for us to make some theoretical innovation and break some core technologies. In the third stage, the technology self-development is more important. With the appearance of various kinds of industrial applications, it is necessary to develop products with independent intellectual property and copyright protection to achieve the goal of protecting national information security and promoting the information industry in China.

2. Please describe your expertise and contribution to KDD.

In our group, we carry out DM research together with Data Warehouse (DW) and On-Line Analytical Processing (OLAP). Early in 1996, we published a series of articles in the "Computer World" journal, which introduced data warehouse related techniques. Two years later, we published the first data warehouse textbook "Data Warehouse and On-Line Analytical Processing", which explained the techniques of DW, OLAP and DM. In 2001, we applied for the establishment of DB and BI Research Center of the Ministry of Education (MOE). It was approved and passed the evaluation successfully in 2003. In 2004, cooperated with well-known DW company NCR, we set up the RUC-NCR joint laboratory, and conducted a sequence of cooperation on DW research, training and application promotion.

From 2000 to 2005, we conducted a wide range of research in DW and DM fields. Our research interests include multi-dimensional data model, cube computing, materialized view, index selection, intelligent data analysis, classification, association rule mining, etc. We published dozens of papers (including the one published on KDD'04 [3]), and developed a parallel data warehouse prototype system ParaWare. These works were supported by two NSFC projects ("DW Technology Research" and "Analysis-oriented High-performance DB Key Technique Research"), one National 863 project ("Domain-oriented Data Analysis and Mining Technique Research"), and one MOE project ("Analysis-oriented Three-high and One-big DB key Techniques Research").

After that, due to the diversity and complexity of data, we extended our research interests to profit-based mining (supported by the youth project of NSFC "Research on the technology of the profit-based data analysis and mining"), network and stream mining (supported by the major program of NSFC "Research on the general network knowledge editor and topic semantic network"), text and multimedia mining (supported by the key program of NSFC "Research on the theory and method of Network information fusion and knowledge service"), and other fields. In order to utilize the advantage of new hardware platform, we carry out research on memory-based OLAP (supported by the Beijing municipal education commission manufacture-learning-research cooperation project "main memory-based OLAP"), GPU-based OLAP (supported by the MOE Ph. D. Programs Foundation project "GPU-based OLAP"), and so on. We conducted research on topic semantic network construction, knowledge fusing and service, profit-based data mining, frequent mining, what-if query processing, etc. Our papers are published on SIGMOD'06, VLDB'07, DKE'09, KDD'10, TKDE'11, etc. In addition, we got the best paper awards of SDM'10 and ADMA'10.

The paper we published on SIGMOD'06 [4] needs to be mentioned since it is the first SIGMOD paper whose first author came from a Chinese mainland organization. For the first time, this paper proposed the concept of dominant relationship analysis, which can be used to do economic analysis. It is showed in Google Scholar that it has been cited for 78 times by SIGMOD, VLDB, SIGKDD, ICDE, TKDE, etc. Prof. H. V. Jagadish, from Electronic Engineering and Computer Science department at University of Michigan, gave an evaluation on this paper in Digital Review. He said: "This paper, in a very clever way, ties these two ideas together".

In addition, we have had strong academic cooperation with Microsoft Research Asia (MSRA), HP Labs of China, and so on. For example, we developed "Story Teller" system under the support of MSRA cooperation project. This system can detect current hot topics and their evolution by analyzing the user click-stream logs. The related demo paper was published on KDD'10. The "Story Teller" system was selected to attend the presentation of 2011 Microsoft

Global Education Summit which was the unique system coming from Chinese mainland..

3. Please share with us your view on the future of KDD both in China and the world.

After the development of more than twenty years, data mining has absorbed many latest research results and grown into an independent research branch. Currently, the research and application status of this field is still in the stage of exploration. With the fast development of network and database techniques, and with the appearance of various kinds of industrial applications, data mining techniques will face the following challenges:

(1) Data Analysis and Mining on Massive Uncertain Heterogeneous Data

With the fast development of the Internet, not only the quantity of the data, but also the scale of Internet applications is increasing rapidly. In different ways, these applications store and manage data. Consequently, BIG heterogeneous data is produced. This kind of data has the special properties of disorder and uncertainty; this makes it hard for people to dig out the useful information. In the future, data mining research should provide effective ways to integrate, analyze, and mine such kind of uncertain and heterogeneous data.

(2)Privacy Protection in Data Mining

In carrying out data mining operations, how to protect the personal privacy is one serious and hard problem [5]. Since some sensitive information stored in the network is usually collected without personal awareness, mining on these data may result in serious violation of personal privacy. With the continual improvement of data mining algorithms and tools, this issue becomes more and more important. Some work has been already done on this topic, but many problems still remain unsolved. Thus, a further research is needed.

(3) Data Mining under the Mobile Environment

Mobile Internet is bringing a profound revolution to the information industry. With the arrival of the 3G era, it is an inevitable trend that a large number of mobile phones or other mobile devices need frequently to access the Internet to get information. Although now mobile users can access Internet to use Google and other search services, but due to the limitation of the cell phone, users often cannot effectively get their desired results. This is because in many cases, the mobile users' query requests are relevant to their locations. In addition, the cell phone has small screen and limited computing resource. These should be specially considered when mobile data mining is performing.

(4) Integration of Data Mining with the Database/Data warehouse

When designing a data mining system, a key issue is how to integrate it with the existing DB/DW systems. Since most data in DB/DW has been well integrated, it is easier to mine task-related, high-quality knowledge based on these systems. Otherwise, data mining systems will have to spend a lot of time on finding, collecting, cleaning and transforming data. To integrate data mining into DB/DW effectively, the most important thing is to identify common data mining primitives and implement them in the DB/DW systems. But this is not easy, and further research is needed.

(5) Data Mining Application Exploration

With the development of data mining technology, the application range of data mining becomes wider and wider. It is used not only in traditional fields such as retail, telecommunications, and banking to do marketing analysis, customer relationship management, customer behavior and market forecast, but also in agriculture, electricity, tax, biology, astrophysics, chemistry, medicine and other fields. However, at many times, users are unsatisfied with the result of data mining because the result fails to meet users' expectations. For business managers, it is very difficult to understand the probabilities or rules embedded in the data mining result. Professional data analysts are needed in this case to help understand or further explain the result. So that more work should be done to make the mining process more intelligent, and to make the mining result more direct and visible.

4. REFERENCES

[1] The Dragon Star Project, http://dragonstar.ict.ac.cn/

[2]http://scholar.google.com/scholar?q=jianyong+wang&hl=zh-CN&lr=

[3] Cuiping Li, Gao Cong, Tung Kum Hoe, Shan Wang, Incremental Maintenance of Quotient Cube for Median, ACM-SIGKDD 2004 International Conference, August 22-25,2004, Seattle, USA

[4] Cuiping Li, Beng Chin Ooi, Anthony K. H. Tung, Shan Wang. "DADA: A Data Cube for Dominant Relationship Analysis", ACM-SIGMOD 2006, Chicago, USA，2006

[5] Privacy – An Evolving Challenge, http://www.scl.org/site.aspx?i=ne21077

[6] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition. 2006

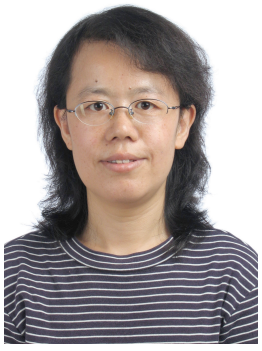
[7] Jiawei Han and Jing Gao, Research Challenges for Data Mining in Science and Engineering, in H. Kargupta, et al., (eds.), Next Generation of Data Mining, Chapman & Hall, 2009

[8] Gregory Piatetsky, Interview with Jon Kleinberg, KDD Explorations, Volume 9, Issue 2

About the authors:

Shan Wang is a Professor of Computer Science in the School of Information at Renmin University of China, in Beijing, China. She finished her undergraduate studies in Physics at the Peking U. in 1968 and received her Master Degree from RUC in 1981 in CS. Prof. Wang has been taken actively part in research and development work in database technology over 30 years. During her work, she advised more than 100 Ph.D. and master students. She participated more than 50 large research projects and as the leader of them, Prof. Wang's research involves various aspects of data management technologies, including High Performance Parallel Database, OLAP and DW, Mobile DBS, Grid Data Management and Searching Databases with Keywords. Her current research interests include Main-memory DB, Video Data Management and Analytics, Data Intensive Computing, DW and BI. Many of her research achievements have awarded by Ministry of Science and Technology, Ministry of Education, Beijing Government.

Cuiping Li received a B.E. degree from Xi'an Jiao Tong University, China, in 1994 and an M.E. degree from Xi'an Jiao Tong University, China, in 1997. In 2003, she received her Ph.D. from the Institute of Computing Technology, CAS. She is currently a Professor of Renmin University of China. Her current research interests include database systems, data warehouse, and data mining.



Hong Chen received a B.E. degree from Renmin University of China in 1986 and a M.E. degree from Renmin University of China, in 1989. In 2000, she received her Ph.D. from the Institute of Computing Technology, CAS. She is a professor in the School of Information, Renmin University of China. Her current research interests include database systems, data warehouse, and wireless sensor networks.