A Conversation with Professor Jianzhong Li

Jianzhong Li
Harbin Institute of Technology
Beijing 100084, China
lijzh@hit.edu.cn

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

The research on knowledge discovery and data mining (KDD) in China started a few years later than some other countries. In 1993, the National Natural Science Foundation of China (NSFC) funded the first research project in the field of KDD. In the nearly 20 years of development, a large number of research institutes and universities have been active in carrying out innovative research on the theory and applications of KDD, including the Tsinghua University, Peking University, Fudan University, Nanjing University, Institute of Computing Technology in Chinese Academy of Sciences, Sichuan University, Harbin Institute of Technology and so on.

The Chinese academic journals that publish the latest research results in the field of KDD in China include Chinese Journal of Computer Science, Journal of Computer Science and Technology, Journal of Software, Journal of Computer Research and Development, Chinese Journal of Electronics, to name a few.

The National Database Academic Conference (NDBC), hosted by the China Computer Federation (CCF) Technical Committee on Databases (CCF TCDB), is the first academic conference in China that presents the latest research and industrial papers in the field of KDD. Until now, the NDBC conference has been successfully held for 28 times. Every since the 14th NDBC conference held in 1997, a session on KDD has been set up specifically to provide a platform for researchers to exchange their ideas on KDD research.

From 2005, the CCF Technical Committee on Artificial Intelligence and Pattern Recognition and the Chinese Association for Artificial Intelligence (CAAI) Technical Committee on Machine Learning have jointly hosted the China Conference on Data Mining (CCDM) once every two years. The 1st to the 4th CCDM conferences were held in 2005, 2007, 2009 and 2011, in Beijing, Zhengzhou, Yantai and Guangzhou, respectively.

China's KDD research community has had close cooperation with KDD research communities worldwide. A number of academic conferences on KDD, such as PAKDD 1999 (Beijing), PAKDD 2007 (Nanjing), PAKDD 2011 (Shenzhen) and ADMA, have been successfully held in China in recent years. The premium conferences on databases, SIGMOD 2007 and ICDE 2009, that also presented research papers on KDD, have been held in Beijing and Shanghai, respectively. The year 2012 will be a memorable year in the history of China's KDD research community because the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012), the premium conference on knowledge discovery and data mining, will be held in Beijing, China.

2. Please describe your expertise and contribution to KDD.

The Data-Intensive Super-Computing (DISC) research center of Harbin Institute of Technology (HIT) has been carrying out the research on databases, knowledge discovery and data mining, data quality, wireless sensor networks and cyber-physical systems for about twenty years. Here, in the last five years, our group has made significant progress in the research on mining uncertain graph data, and has published some papers in the premium journals and conferences on data mining such as IEEE Transactions on Knowledge and Data Engineering (TKDE), KDD, ICDE and CIKM. In 2009, Zhaonian Zou, one of my Ph.D. students at that time, presented his paper "Frequent Subgraph Pattern Mining on Uncertain Graph Data" in CIKM. This paper is the first one on mining uncertain graph data. His thesis "Algorithms for Mining Uncertain Graph Data" studies the semantic model of uncertain graph data, the computational complexities of typical mining problems on uncertain graph data and the algorithms for mining uncertain graph data.

In recent years, a large amount of graph data has been collected in real applications, in which a large amount of useful knowledge is buried. However, in practice, due to the limitations of data acquisition techniques, data imprecision and data incompleteness, uncertainties are ubiquitous in graph data. Graph data accompanied with uncertainties is called *uncertain graph data*, e.g., protein-protein interaction (PPI) networks in bioinformatics, topologies of wireless sensor networks, and so on. For PPI networks, due to the low reliability of high-throughput biological experiments, each PPI, i.e., edge in the network, obtained via high-throughput experiments does not definitely exist but has some uncertainty that indicates the possibility of the PPI existing in practice. There is a large amount of useful knowledge hidden in uncertain graph data. Thus, it is very important to discover knowledge from uncertain graph data, which is also known as *uncertain graph mining*. However, existing graph mining methods and uncertain data mining methods can't deal with uncertain graph mining problems.

In our work, we proposed the model of uncertain graph data, the algorithms for mining frequent subgraph patterns on uncertain graph data and the algorithm for mining top-k maximal cliques.

In our research on frequent subgraph pattern mining over uncertain graph data, we investigated this problem under the expected semantics and the probabilistic semantics, respectively. For the expected semantics, a measure called *expected support* is proposed to evaluate the significance of a subgraph pattern, and the problem of mining frequent subgraph patterns is formally stated on uncertain graph data under the expected semantics based on expected support. It is rigorously proved that this problem is

NP-hard and that it is #P-hard to compute the expected support of a subgraph pattern. To reduce computational complexity, our method tries to compute an ε-approximate set of frequent subgraph patterns including all frequent subgraph patterns and a fraction of infrequent subgraph patterns but with expected support no less than proportion ε than the expected support threshold *minsup*. Actually, an approximate mining algorithm is proposed to compute an (ε, δ)-approximate set of frequent subgraph patterns. Any frequent subgraph pattern is contained in this set with a high probability, and any infrequent subgraph pattern with expected support less than $(1-\varepsilon) \cdot$ minsup is contained in this set with a low probability.

For frequent subgraph pattern mining over uncertain graph data under the probabilistic semantics, a measure called φ-*frequent probability* is proposed to evaluate the significance of a subgraph pattern, and the problem of mining frequent subgraph patterns is formally stated on uncertain graph data under the probabilistic semantics on the basis of φ-frequent probability. It is rigorously proved that this problem is NP-hard and that it is #P-hard to compute the φ-frequent probability of a subgraph pattern. To reduce computational complexity, our method tries to compute an ε-approximate set of frequent subgraph patterns including all frequent subgraph patterns and a fraction of infrequent subgraph patterns but with φ-frequent probability no less than ε than the φ-frequent probability threshold τ. Actually, an approximate mining algorithm is proposed to compute an (ε, δ)-approximate set of frequent subgraph patterns. Any frequent subgraph pattern is contained in this set with probability at least $((1-\delta)/2)^s$, where s is the number of edges of the subgraph pattern, and any infrequent subgraph pattern with φ-frequent probability less than $\tau-\varepsilon$ is contained in this set with probability at most $\delta/2$.

We have also investigated top-k maximal clique mining on uncertain graphs. A measure called *maximal-clique probability* is proposed to evaluate the probability of a subset of vertices being a maximal clique, and the problem of mining top-k maximal cliques is formally defined on uncertain graphs under the probabilistic semantics. This problem is rigorously proved to be NP-hard. A method is developed for computing the maximal-clique probability of a subset of vertices in polynomial time. On the basis of this method, a branch-and-bound algorithm is developed to quickly discover top-k maximal cliques, which uses a number of efficient pruning techniques and some preprocessing techniques. The application of this algorithm in protein complex prediction is able to increase the accuracy of prediction.

Besides our research on uncertain graph mining, we have also studied the data mining problems on data streams and wireless sensor networks.

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

Nowadays, ubiquitous computing, wireless sensor networks (WSN), cyber-physical systems (CPS) and the Internet of Things (IoT) have been attracting increasingly more attentions from the research community and the industry. In this new generation of computing infrastructures, massive data will be continuously generated, transported, stored and processed, from which large quantity of new knowledge about the physical world could be obtained via data mining. However, the quantity of the data generated in these computing infrastructures will be several orders of magnitude larger than the quantity of the data we have seen so far, and hence far exceed the processing capability of the existing data mining algorithms. Therefore, a large number of novel algorithms with sub-linear running time should be developed to meet the requirement of the new computing systems. To this end, new models, theories, techniques and methodologies should be established in the future to make it possible to discover new knowledge about the world from the astronautic amount of data collected from the world.

About the author:

Jianzhong Li is a professor and the chairman of the Department of Computer Science and Engineering at the Harbin Institute of Technology, China. He worked in the University of California at Berkeley as a visiting scholar in 1985. From 1986 to 1987 and from 1992 to 1993, he was a scientist in the Information Research Group in the Department of Computer Science at Lawrence Berkeley National Laboratory, USA. He was also a visiting professor at the University of Minnesota at Minneapolis, Minnesota, USA, from 1991 to 1992 and from 1998 to 1999. His current research interests include database management systems, data warehousing, data mining, and wireless sensor networks. He has authored three books and published more than 200 papers in refereed journals and conference proceedings, such as VLDB Journal, Algorithmica, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Parallel and Distributed Systems, Parallel and Distributed Database, SIGMOD, VLDB, ICDE, INFOCOM, ICDCS. He has been involved in the program committees of major computer science and technology conferences, including SIGMOD, VLDB, ICDE, INFOCOM, ICDCS, and WWW. He has also served on the editorial boards for distinguished journals, including Knowledge and Data Engineering, and refereed papers for varied journals and proceedings.