A Conversation with Professor Bo Zhang

Bo Zhang
Tsinghua University
Beijing 100084, China
dcszb@mail.tsinghua.edu.cn

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

In the 11th international joint conference on artificial intelligence (IJCAI'89), in 1989, Piatetsky-Shapiro and colleagues led a seminar on knowledge discovery in databases, which signaled the coming of age for the field of data mining. Soon after, in 1993, data mining research officially begun by first being supported by the National Natural Science Foundation of China (NSFC). This was viewed as the beginning of data mining research in China. Thus, China was among the pioneers in data mining research given that the time gap was only 4 years. Afterwards, major research activities on both fundamental research and practical applications can be found in many research institutions and universities, including Tsinghua University, Peking University, Institute of Computing Technology in Chinese Academy of Science, Nanjing University, and many others. In these institutions, applications of data mining cover a broad range of areas including finance, business, transportation, medicine, energy, etc. However, before the 21th century, due to the shortage of data, China's KDD research was still limited in terms of theoretical depth and types of applications. This period can be considered a warm-up period, in which Chinese researchers were learning and following international major KDD research directions.

The rapid and sustainable development of China's Internet has brought about a turning point for KDD research. This represented a major boost for KDD research and development in China at the beginning of the century. In 2002, the total number of Netizens in China has reached 55.6 million, which was the second largest in the world. In 2008, China's Internet users hit 221 million, ranking as the world's largest. This trend has continued rapidly; recently, this number has reached 500 million. This tremendous number of Internet users, together with the huge amount of data, provided a very rich and interesting data mining research, particularly due to factors such as China's special culture, language, and social environment. Consequently, research on data mining theory and its applications in China has leap-frogged. During this period, important achievements have been made in several research institutions, including Tsinghua University, Nanjing University, the Institute of Computing Technology in Chinese Academy of Science, Fudan University, etc. In term of applications, many outstanding KDD products appeared. For example, the most recent Chinese-language input methods on computers have adopted technologies including massive-scale data mining on the Internet. The data mining activities include analysis based on dynamically updated word and phrase dictionaries, correction and learning based on user behavior models. This has resulted in a much better computer interface for Chinese users when they type Chinese on keyboards. As a result, ever since 2006, these input methods have now totally replaced the traditional Chinese input methods. Some leading products such as Sogou's Chinese Pinyin software have gained more than 80% of market share, dramatically increasing the efficiency of computer usage for Chinese users. Similar products include Youdao dictionary, which automatically builds a Chinese-English dictionary based on the result of mining on the online bilingual study materials.

2. Please describe your expertise and contribution to KDD.

I am part of the State Key Laboratory of Intelligent Technology and Systems at Tsinghua University, which is a national-level lab in China that was founded in February, 1990. Since its establishment, we have been conducting lots of research work on data mining. Especially from the beginning of this century, along with the rapid development of the Internet in China, we have made advances in both theoretical research and practical applications on mining useful knowledge from the huge amount of Internet data. The followings are some of our achievements:

- **Web data extraction and structure learning theory**

We have conducted systematic research on the topics of statistical relational learning [2], and on its complex data modeling frameworks – probabilistic graphical models and Markov logic networks [3].

At the theoretical level, we have proposed a general probabilistic graphical model learning framework called maximum-entropy discrimination Markov networks (MaxEnDNet, or simply, MEDN), which integrates both maximum-margin learning and maximum likelihood estimation. This framework combines and extends the merits of the existing methods based on either maximum-margin learning or maximum likelihood estimation [5]. Our Bayesian-style formulation of MaxEnDNet also provides natural extensions to deal with missing data when performing maximum-largin learning [9] as well as to leverage the advantages of non-parametric Bayesian techniques to deal with the hard model selection problem in learning a maximum margin model, such as resolving the unknown number of components in learning a mixture of SVM classifiers or determining the unknown dimensionality of latent features in learning a latent SVM model [17][18].

At the application level, we have applied relational learning to Web data extraction. Based on our research of probabilistic graphical models, we have proposed a probabilistic framework for Web data extraction and learning [4] [6] [7] [8]. By modeling the Web structures, our methods could achieve high extraction accuracy and robustness. We also developed a method called StatSnowball [12], which combines the discriminative ability of Markov logic networks and the bootstrapping framework of Snowball [1], to effectively model and extract entity relationships on Internet data with only a few labeled data (i.e. samples with

supervised information). StatSnowball has been applied to an entity-relationship search engine developed by Microsoft Research Asia [23][22].

- **Predictive latent subspace learning**

One of the core tasks in data mining is to extract the potential structures and latent relationships among noisy data that can reveal the properties of the data. Nowadays, more and more labeled or weakly labeled data can be freely obtained from the Web via methods such as crowdsourcing. Traditional methods including unsupervised feature analysis cannot meet all of our requirements. In order to solve this problem, we have proposed methods for maximum margin-based predictive latent subspace learning [10] [14], and applied them to sentiment analysis and image classification problems. We have also extended the methods to multi-view learning [15].

- **Sparse learning in high dimensions**

When dealing with data in high-dimensional spaces, both data mining and machine learning suffer from the "curse of dimensionality". Structure learning of probabilistic graphical models is another problem of learning in high-dimensions that is challenged by how to effectively control the exponential growing model space. In order to solve these problems, we proposed effective sparse learning methods with appropriate sparsity-inducing regularization terms to handle the high dimensional data in bioinformatics, computer vision and text mining [11][16]; and we also developed a fast algorithm for automatically learning the structure of Markov networks [13].

- **Active and Broad Collaboration with Academia and Industry**

We have actively maintained close collaboration with major companies in China's Internet industry. For example, we have established the "Tsinghua-Sohu Joint Laboratory on Search Technology" and "Tsinghua-Tencent Internet Innovation Technology associated Laboratory". The collaboration provides us with numerous large-scale Practical Web data, as well as the ability to innovate at the frontier of technology development in China's Internet area, allowing us to have a good understanding of the application requirements and the latest trends in the Internet industry. Furthermore, we have built collaboration with international research institutes. Through international cooperation platforms, including "Tsinghua-Waterloo joint research center for internet information acquisition" and "Tsinghua-NUS Extreme Search Research Center", we have collaborated with many international counterparts to conduct research on Internet data mining.

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

Although tremendous progress has been made in data mining, there are still many important issues that need to be resolved.

First of all, machine learning, which provides the theoretical foundations and tools for data mining, still has serious flaws. Machine learning was originally designed to find the potential semantics and rules by analyzing empirical data. Yet currently, it is based on statistics and classical information theory, using apparent co-occurrence frequency values as the evidence of the inner relationship among the data. Clearly, high co-occurrence frequency does not necessarily represent inner causal relationship (i.e., the semantics). The difference between human intelligence and artificial intelligence is that human beings can capture the semantics in data with better results. Therefore, to attain human-like intelligence, a future direction for machine learning is cross-disciplinary research among classical information theory, cognitive science, psychology, etc. In fact, some valuable preliminary work has been done at present. For example, some scholars have successfully applied machine learning and data mining algorithms on data in cognitive science to interpret the phenomenon of human learning (e.g. feature learning [19] and vocabulary learning [20]). Some other scholars have successfully used the theoretical and empirical principles derived from cognitive science to guide the research on machine learning theory and algorithms [21]. However, there is still a long way to go before we can address practical problems.

Second, the problem of how to deal with "Big data", "multimodal", "social media" is also a challenge. To handle large-scale data mining efficiently, we need to make continuous improvements on both data modeling and algorithms, e.g., relational learning. Although we have achieved some success in this direction, the existing methods still suffer from high computational complexity and limited accuracy of learning algorithms. In the future, it will become a major direction of data mining to extend the traditional data mining models (e.g. K-means clustering and SVM classification), as well as new statistical models (e.g. Markov networks on relational learning and Markov logic networks) to online learning versions and distributed computing versions.

Finally, China's Internet industry and user population are growing in a strong momentum. Meanwhile, with the increase of Internet penetration rate, net citizens' internet-dependence increases and they are more willing to engage in Interaction with each other on the Internet. In order to meet the requirements for effectively dealing such data, user content-oriented data mining research, especially social media demand-oriented data mining research, may constitute another very important future research direction.

4. REFERENCES

[1] E. Agichtein and L. Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In: Proceedings of the International Conference on Digital Library, 2000.

[2] L. Getoor and B. Taskar. Introduction to Statistical Relational Learning. The MIT Press, 2007.

[3] M. Richardson and P. Domingos. Markov Logic Networks. Machine Learning, 62, 107-136, 2006.

[4] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. 2D Conditional Random Fields for Web Information Extraction. In: Proceedings of the 22th International Conference on Machine Learning (ICML'05), Bonn, Germany, 2005, 1049-1056.

[5] J. Zhu and E.P. Xing. Maximum Entropy Discrimination Markov Networks. Journal of Machine Learning Research (JMLR), 10(Nov): 2531-2569, 2009.

[6] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction. In: Proceedings Of the 12th ACM conference on Knowledge Discovery and Data Mining (SIGKDD'06), Philadelphia, USA, 2006, 494-503.

[7] J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen. Dynamic Hierarchical Markov Random Fields for Integrated Web Data Extraction. Journal of Machine Learning Research (JMLR), 9(Jul): 1583-1614, 2008.

[8] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and H.-W. Hon. Webpage Understanding: an Integrated Approach. In: Proceedings Of the 13th ACM conference on Knowledge Discovery and Data Mining (SIGKDD'07), San Jose, USA, 2007, 903-912.

[9] J. Zhu, E.P. Xing, and B. Zhang. Partially Observed Maximum Entropy Discrimination Markov Networks. In: Advances in Neural Information Processing Systems (NIPS'08), Vancouver, Canada, 2008.

[10] J. Zhu, A. Ahmed, and E.P. Xing. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classificatioin. In: Proceedings of the 26th International Conference on Machine Learning (ICML'09), Montreal, Canada, 2009, 1257-1264.

[11] J. Zhu and E.P. Xing. On the Primal and Dual Sparsity in Markov Networks. In: Proceedings of the 26th International Conference on Machine Learning (ICML'09), Montreal, Canada, 2009, 1265-1272.

[12] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen. StatSnowball: a Statistical Approach to Extracting Entity Relationships. In Proc. of WWW, 2009.

[13] J. Zhu, N. Lao, and E.P. Xing. Grafting-Light: Fast, Incremental Feature Selection and Structure Learning of Markov Random Fields. In: Proceedings Of the 16th ACM conference on Knowledge Discovery and Data Mining (SIGKDD'10), Washington DC, USA, 2010, 303-311.

[14] J. Zhu, L. Li, L. Fei-Fei and E.P. Xing. Large Margin Learning of Upstream Scene Understanding Models. In: Advances in Neural Information Processing Systems (NIPS'10), Vancouver, Canada, 2010.

[15] N. Chen, J. Zhu and E.P. Xing. Predictive Subspace Learning for Multi-view Data: a Large Margin Approach. In: Advances in Neural Information Processing Systems (NIPS'10), Vancouver, Canada, 2010.

[16] S. Lee, J. Zhu and E.P. Xing. Detecting eQTLs using Adaptive Multi-task lasso. In: Advances in Neural Information Processing Systems (NIPS'10), Vancouver, Canada, 2010.

[17] J. Zhu, N. Chen, and E.P. Xing. Infinite SVM: a Dirichlet Process Mixture of Large-margin Kernel Machines. In: Proceedings of the 28th International Conference on Machine Learning (ICML'11), Bellevue, Seattle, 2011.

[18] J. Zhu, N. Chen, and E.P. Xing. Infinite Latent SVM for Classification and Multi-task Learning. In: Advances in Neural Information Processing Systems (NIPS'11), Granada, Spain, 2011.

[19] J. Austerweil and T. Griffiths. Analyzing Human Feature Learning as Nonparametric Bayesian Inference. In: Advances in Neural Information Processing Systems (NIPS'10), Vancouver, Canada, 2009.

[20] F. Xu and J. Tenenbaum. Word Learning as Bayesian Inference. Psychological Review, 114(2), 245-272, 2007.

[21] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'07), 29(3), 411-426, 2007.

[22] http://renlifang.msra.cn/

[23] http://entitycube.research.microsoft.com/

About the author:

Bo Zhang is a Professor of Computer Science and Technology Department of Tsinghua University in China. He is a Fellow of Chinese Academy of Science (CAS), a Vice-chairman of academic committee of Information Science and Technology College in Tsinghua University, and Chairman of Intelligent Control Professional Committee of Chinese Automation Association.
http://www.csai.tsinghua.edu.cn/personal_homepage/zhang_bo/index.html