A Conversation with MSRA Researchers

Wei-Ying Ma
Microsoft Research Asia
Beijing, China
wyma@microsoft.com

Zheng Chen
Microsoft Research Asia
Beijing, China
zhengc@microsoft.com

Hang Li
Microsoft Research Asia
Beijing, China
hangli@microsoft.com

Tie-Yan Liu
Microsoft Research Asia
Beijing, China
tyliu@microsoft.com

Zaiqing Nie
Microsoft Research Asia
Beijing, China
znie@microsoft.com

Haixun Wang
Microsoft Research Asia
Beijing, China
haixunw@microsoft.com

Ji-Rong Wen
Microsoft Research Asia
Beijing, China
jrwen@microsoft.com

Xing Xie
Microsoft Research Asia
Beijing, China
xingx@microsoft.com

Yu Zheng
Microsoft Research Asia
Beijing, China
yuzheng@microsoft.com

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

Ten years ago, KDD research was still in its infancy in China. Things have changed significantly. With the push from the technological advancement in academia and the pull from the explosive growth of application needs in industry, KDD research is flourishing. At Microsoft Research Asia, we have been conducting research in many areas related to KDD research, including web search, data mining, information retrieval, multimedia mining, natural language processing, and visualization. In addition to publishing papers in KDD and developing technologies for commercial products, we have also contributed to the talent development in China by supervising students and growing young researchers who later become well known in the field related to KDD in universities and industries.

2. Please describe your expertise and contribution to KDD.

At Microsoft Research Asia, we have many groups working on KDD related research. Our major contributions are in the following areas:

- **Entity-level Web Search:** We developed the first technique to analyze Web pages using visual cues and use the information to model the Web and extract structured data from Web pages. With these advanced Web-analysis techniques, we developed a next-generation search engine that goes beyond traditional page-level relevance ranking. By extracting and integrating information about real-world entities such as people, places and things (e.g. products) from billions of public Web pages, our system creates a paradigm shift on Web search by enabling search queries, relevance ranking, and browsing and navigation of search results at the level of entities. We also built the Microsoft academic search engine based on entity-level search technologies. It provides many innovative ways to retrieve rank and explore scientific papers, conferences, journals, and authors based on their importance and relationship. (MSRA member: Zaiqing Nie, Ji-Rong Wen, Wei-Ying Ma)
- **Search Infrastructure for Web-scale Data Mining and Knowledge Discovery:** We initiated an effort in Microsoft to develop an infrastructure for web-scale data mining and knowledge discovery for web search. Different from traditional Internet services, web search involves myriad offline computations to analyze the data at a very large scale, and an infrastructure for "scale" experiments is often required to evaluate the effectiveness of newly invented algorithms in a semi-real environment. Such an infrastructure is also critical for supporting massive web mining, knowledge discovery, and asynchronous metadata exchange in a search engine pipeline so that the cycle of idea formulation, experimentation, and deployment can be iterated quickly. (MSRA member: Ji-Rong Wen, Wei-Ying Ma)
- **Relevance in Web Search:** Relevance is one of the most important factors for web search, and it has been observed that many hard cases in search relevance are due to term mismatch between query and document (e.g., query "ny times" does not match well with document only containing `new york times'). We have developed advanced technologies to address the grand challenge, by conducting better query and document understanding, and performing better matching between enriched query and document representations. The technologies that we developed include those for query spelling error correction, large-scale topic modeling, key-phrase extraction in query and document, and learning of matching between query and document. All of them can significantly improve the relevance in search. Our query spelling error correction methods using log linear model and CRF are considered state-of-the-art methods for the task. Our topic modeling technique called RLSI has much better scalability than existing techniques (MSRA member: Hang Li)
- **Interactive Knowledge Mining through Crowdsourcing:** Traditional knowledge mining approaches aim to automatically discover and integrate knowledge from structured data sources or unstructured text documents. However, automated algorithms still perform poorly in some knowledge mining tasks. To address this problem, we have been developing an interactive knowledge mining framework

called iKnoweb that enables users to interact with and contribute to automated knowledge mining tools such as entity extractors and name disambiguation systems. The system interacts with users to retrieve the knowledge in their minds and keeps learning through interacting with people. As more users interact with the system, more knowledge will be accumulated. More user interactions also help make the system more powerful and human task easier. (MSRA member: Zaiqing Nie)

- **Knowledge Base:** We have put considerable efforts on building knowledge bases and using them to support a wide range of applications. Projects such as Probase focus on building general-purpose knowledge bases, and we also have efforts on creating domain-specific knowledge bases. It is our belief that in order to make applications more intelligent, data mining techniques need to be able to integrate and manipulate human knowledge. We have been using these techniques and knowledge bases to improve search and ads related applications. (MSRA member: Haixun Wang)
- **Graph Data Mining:** We have developed a series of technologies to perform data mining (cluster and rank) on large-scale, information-rich, and heterogeneous graph data. In KDD 2005, we published a technology to co-cluster star-structured Web data. We formulated the problems as the fusion of multiple pairwise co-clustering sub-problems with the constraint of the star structure, and proposed the concept of consistent bipartite graph co-partitioning, and developed an efficient algorithm based on semi-definite programming. In KDD 2006, we published a technology to detect events from click-through logs based on the evolution pattern in the data. We proposed a dual graph representation for the data, which can encode both semantic information and temporal information into its structure and metadata. We then designed a two-phase graph cut algorithm to partition the dual graph for event detection. In KDD 2011, we published a parametric graph ranking technology to optimize the ranking in a large-scale graph with rich metadata, according to partial supervision. We show that it is possible to efficiently learn the optimal parameters of the model and the optimal ranking scores of the nodes using a MapReduce framework, by leveraging the sparsity of the graph. (MSRA member: Tie-Yan Liu, Hang Li)
- **Search Log Mining:** Several advanced and ground-breaking methods of search log mining have been proposed from our team. For example, we proposed algorithms for query clustering and query expansion based on the click-through bi-partite graph, which are among the first studies on the issue and are highly cited in the community. The work on context aware search based on mining of search logs initiated a new approach to web search and one of the papers won the best application award paper at KDD'08. (MSRA member: Ji-Rong Wen)
- **Behavioral Targeting (BT):** BT aims to understand the user interests, tasks as well as preference through mining the user profile and user behavioral log. We aim to deliver the right ad and right recommendations to the right audience at the right time. Though this area is relatively young in KDD, it is attracting increasingly attention from both academia and industry due to its huge market potential. We have published a book chapter and many papers which have influenced many other researchers to conduct research in this area. We expect to see an even brighter future of BT research in KDD. (MSRA member: Zheng Chen)
- **Urban Computing:** We launched the project of urban computing that aims to understand city dynamics by mining large-scale traffic and geographical datasets. An example is that we glean the problematic (or less-effective) urban planning in a city using GPS trajectories generated by a large number of taxicabs. This project has been featured twice by MIT Technology Review in 2011. The project resulted in several best paper awards and fostered a few real-world prototype systems that are making impact to urban planning and intelligent transportation systems. (MSRA member: Yu Zheng, Xing Xie)

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

KDD is going to have a bright future. We know statistics-based approaches have achieved great successes in the last decade. For example, machine translation based on statistical models of big data is revolutionary both in its approach and in its effectiveness. I consider these triumphs are a celebration of the success of KDD, machine learning, and AI in the past decade.

The coming decade is going to be more challenging and more exciting for KDD. As we celebrate its success, we also need to acknowledge its current limitations. For example, statistics-based approaches still do not "understand" the meaning of the problem. This affects both the depth and breadth of the problems that statistics based approaches can handle. Thus, how to effectively and efficiently capture the inherent structure of the input, instead of treating it as a huge, flat array of independent features might be the biggest challenge for KDD and machine learning.

As for the future of web search and mining, there will be more opportunities for disruptive innovations. For example, emerging consumer devices such as smartphones and tablets are increasingly driving user activities into apps and services that offer superior capabilities for the completion of tasks. In addition, many such technologies do not necessarily live inside the traditional web browser. The new generation of mobile devices, equipped with a variety of sensors and location-based technologies, is connecting people to information not only in the virtual world, but also in the physical world by bridging between the two. Social networks are also being integrated into search, making it easier to find answers and share information, while natural user interface technologies such as touch, voice, and gesture recognition are creating yet further entry points for information access beyond the traditional search box. These new trends and their profound impact on how people interact with information will unlock many new opportunities for researchers in KDD to take the field to a new height.

About the authors:

Dr. Wei-Ying Ma is an Assistant Managing Director at Microsoft Research Asia where he oversees multiple research groups in the area of Web Search, Data Mining, and Natural Language Computing. He and his team of researchers have developed many key technologies that have been transferred to Microsoft's Bing Search Engine. He has published more than 250 papers at international conferences and journals. He is a Fellow of the IEEE and a Distinguished Scientist of the ACM. He currently serves on the editorial boards of ACM Transactions on Information System (TOIS) and ACM/Springer Multimedia Systems Journal. In recent years, he served as program co-chair of WWW 2008, program co-chair of Pacific Rim Conference on Multimedia (PCM) 2007, and general co-chair of Asia Information Retrieval Symposium (AIRS) 2008. He is the general co-chair of ACM SIGIR 2011.

Wei-Ying Ma(http://research.microsoft.com/en-us/people/wyma/), Zheng Chen (http://research.microsoft.com/en-us/people/zhengc/), Hang Li (http://research.microsoft.com/en-us/people/hangli/), Tie-Yan Liu (http://research.microsoft.com/en-us/people/tyliu/), Zaiqing Nie (http://research.microsoft.com/en-us/um/people/znie/), Haixun Wang (http://wis.cs.ucla.edu/~hxwang/), Ji-Rong Wen (http://research.microsoft.com/en-us/um/people/jrwen/), Xing Xie (http://research.microsoft.com/en-us/people/xingx/), and Yu Zheng (http://research.microsoft.com/en-us/people/yuzheng/), are all from Microsoft Research Asia.