Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Data Science for Social Good
2014 KDD Highlights

Wei Wang
Department of Computer Science, University of California, Los Angeles
weiwang@cs.ucla.edu

Introduction

As the premier international forum for data science, data mining, knowledge discovery and big data, the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) brings together researchers and practitioners from academia, industry, and government to share their ideas, research results and experiences. Partnered with Bloomberg, it celebrated its 20th years in 2014 with the theme "Data Science for Social Good".

Among academic conferences, the KDD conference typically has an emphasis on research motivated by real-world applications. It is this synergy of research in areas such as algorithms, computational geometry, database, graph theory, machine learning, natural language processing, statistics, visualization, and many others when applied to problems arising in diverse fields, such as the Web, medicine, biology, and marketing, driving our field forward and making it vibrant and fun. The breadth of topics covered in the 2014 research program is truly comprehensive and nicely balanced among social and information networks, data mining for social good, graph mining, statistical techniques for big data, topic modeling, recommender systems, data streams, scalable methods, Web mining, clustering, feature selection, applications to health care and medicine, public safety, advertising, social analytics, personalization, workforce analytics, health, and many more.

Research Highlights

In the following, we highlight some research advances presented at the 2014 KDD conference.

Scalable Methods for Big Data

Scalability is crucial to big data applications and can be addressed by clever sampling and approximation, parallel computing, and/or levering sparsity properties. Substantial efforts were devoted to scale up graph algorithms on pageRank estimation (Lofgren 2014), betweenness centrality (Yoshida 2014), graph sampling (Ahmed 2014), and graph partition (Bourse 2014). Li and co-authors (Li 2014) investigated this issue in the context of topic modeling. Sampling is employed in topic modeling inference in order to associate latent variables with observations. Leveraging the sparsity property, they proposed an efficient algorithm that approximates a dense, slowly changing distribution buy the combination of Metropolis-Hastings step, use of sparsity, and amortized constant time sampling via Walker's alias method. It scales linearly to the number of instantiates topics in the document rather than the total number of topics, leading to an order of magnitude speedup. This algorithm is generic, and has wide applications in statistical modeling. This paper was recognized with the *Best Research Paper Award*.

Social Media Analysis

Social media analysis continues to attract attentions from researchers and practitioners. Topic modeling, especially in the context of discovery of topics and monitoring topic evolution in social media, is a popular topic. Tong and co-authors (Tong 2014) developed a Topics Crowd Service (TCS) model to discover latent topics in crowd-oriented service data, using belief propagation and pairwise sketching. Schubert and co-authors (Schubert 2014) investigated the problem of early detection of emerging topics before they become hot tags in social media streams. Tsytsarau and co-authors (Tsytsarau 2014) studied the dynamics of news events and their relation to changes of sentiment expressed

on relevant topics. They proposed a framework that models the behavior of news and social media in response to events as a convolution between event's importance and media response function, specific to media and event type. Sudhof and co-authors (Sudhof 2014) showed that modeling dependencies between emotional states can yield insights into human emotion and support more powerful sentiment analysis.

Social and Information Networks

Social networks has been a hot research area in the KDD community for many years, thanks to the tremendous success in industry. However, the research focuses have moved from link prediction, reachability query, and community detection in static networks to modeling the dynamics of information diffusion over the networks (Embar 2014; Kurashima 2014; Wang 2014) and event detection, forecasting, and organization in dynamic and heterogeneous networks of, for instance, sensors or smart devices (Chen 2014; Rozenshtein 2014; Li 2014).

Applications to Healthcare and Medicine

Electronic health records (EHRs) are becoming an increasingly important source of detailed patient information. Effective integration and efficient analysis of EHRs can aid in solving many of the healthcare problems: making informed clinical decisions, improving patient safety, and facilitating investigations and knowledge discovery. Ho and co-authors (Ho 2014) proposed a sparse non-negative tensor factorization method to automatically derive phenotypes that represents complex interactions between several sources and maps to existing medical concepts and easily understood by a medical professional. Ghassemi and co-authors (Ghassemi 2014) demonstrated that latent topic-derived features from electronic health record of ICU patients were very effective in predicting patient hospital mortality.

Conclusion

In summary, we observed a substantial increase in the number of papers related to data mining for social good, as well as works on scaling up algorithms to deal with big data. It is extremely encouraging to see that data mining is branching out to many application areas of societal importance: public policy, education, healthcare, medicine, smart cities, climate, green energy, the Internet of things, and many others. We hope that this trend of data mining affecting all aspects of our society will definitely continue in the future.

References

Ahmed, N., Duffield, N., Neville, J., and Kompella, R. 2014. Graph Sample and Hold: A Framework for Big-Graph Analytics, *ACM SIGKDD,* 1446-1455.

Bourse, F., Lelarge, M., and Vojnovic, M. 2014. Balanced Graph Edge Partition, *ACM SIGKDD,* 1456-1465.

Chen, F. and Neill, D. 2014. Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs, *ACM SIGKDD,* 1166-1175.

Embar, V., Pasumarthi, R., and Bhattacharya, I. 2014. A Bayesian Framework for Estimating Properties of Network Diffusions, *ACM SIGKDD,* 1216-1225.

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., and Szolovits, P. 2014. Unfolding Physiological State: Mortality Modelling in Intensive Care Units, *ACM SIGKDD,* 75-84.

Ho, J., Ghosh, J., and Sun, J. 2014. Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization, *ACM SIGKDD,* 115-124.

Li, A., Ahmed, A., Ravi, S., and Smola. A. 2014. Reducing the Sample Complexity in Topic Models, *ACM SIGKDD*, 891-900.

Li, K., Lu, W., Bhagat, S., Lakshmanan, L., and Yu, C. 2014. On Social Event Organization, *ACM SIGKDD,* 1206-1215.

Kurashima, T., Iwata, T., Takaya, N., and Sawada, H. Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks, *ACM SIGKDD,* 1236-1245.

Lofgren, P., Banerjee, S., Goel, A., and Seshadhri, C. 2014. FAST-Ppr: Scaling Personalized PageRank Estimation for Large Graphs, *ACM SIGKDD,* 1436-1435.

Rozenshtein, P., Anagnostopoulos, A., Gionis, A., and Tatti, N. 2014. Event Detection in Activity Networks, *ACM SIGKDD,* 1176-1185.

Schubert, E., Weiler, M., and Kriegel, H. 2014. SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds, *ACM SIGKDD*, 871-880.

Sudhof, M., Emilsson, A., Maas, A., and Potts, C. 2014. Sentiment Expression Conditioned by Affective Transitions and Social Forces, *ACM SIGKDD*, 1136-1145.

Tong, Y., Cao, C., and Chen, L. 2014. TCS: Efficient Topic Discovery over Crowd-Oriented Service Data, *ACM SIGKDD*, 861-870.

Tsytsarau, M., Palpanas, T., and Castellanos, M. 2014. Dynamics of News Events and Social Media Reaction, *ACM SIGKDD*, 901-910.

Wang, S., Hu, X., Yu, P., and Li, Z. MMrate: Inferring Multi-Aspect Diffusion Networks with Multi-Pattern Cascades, *ACM SIGKDD,* 1246-1255.

Yoshida, Y. 2014. Almost Linear-Time Algorithms for Adaptive Betweenness Centrality Using Hypergraph Sketches, *ACM SIGKDD*, 1416-1425.