Preface

5th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications

http://bigdata-mining.org/

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications The aim of this workshop is to bring together people from both academia and industry to present their most recent work related to big-data issues, and exchange ideas and thoughts in order to advance this big-data challenge, which has been considered as one of the most exciting opportunities in the past 10 years.

Big data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real-time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (call data record)-based processing for billing purposes only. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help improve quality of life and make our world a better place. For example, after we get up every morning, in order to optimize our commute time to work and complete the optimization before we arrive at office, the system needs to process information from traffic, weather, construction, police activities to our calendar schedules, and perform deep optimization under the tight time constraints. In all these applications, we are facing significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models.

August 2016

Wei Fan, Albert Bifet, Jesse Read, Qiang Yang and Philip Yu
BigMine 2016 Program co-Chairs
http://bigdata-mining.org/

BigMine 2016 Workshop Organization

Workshop Chairs

Wei Fan
Baidu Research Big Data Lab
E-mail: wei.fan at gmail.com

Albert Bifet
Telecom-ParisTech
E-mail: albert.bifet at telecom-paristech.fr

Jesse Read
Telecom-ParisTech
E-mail: jesse.read at telecom-paristech.fr

Qiang Yang
Hong Kong University of Science and Technology
E-mail: qyang (at) cse (dot) ust (dot) hk

Philip Yu
University of Illinois at Chicago
E-mail: psyu at cs.uic.edu

Organizers

- Albert Bifet, Telecom-ParisTech
- Wei Fan, Baidu Research Big Data Lab
- Jing Gao, University at Buffalo
- Le Gruenwald, University of Oklahoma
- Dimitrios Gunopulos, University of Athens
- Geoff Holmes, University of Waikato
- Latifur Khan, University of Texas at Dallas
- Dekang Lin, Google
- Jesse Read, Telecom ParisTech
- Deepak Turaga, IBM T.J. Watson Research
- Qiang Yang, Hong Kong University of Science and Technology
- Philip Yu, University of Illinois at Chicago

- Kun Zhang, Xavier University of Louisiana
- Xiatian Zhang, TalkingData. Ltd.
- Yuanchun Zhou, Chinese Academy of Sciences

Treasury

- Xiaoxiao Shi, University of Illinois at Chicago
- Jing Gao, SUNY Buffalo

Program Committee

- Vassilis Athitsos, University of Texas at Arlington
- Roberto Bayardo, Google
- Francesco Bonchi, Yahoo! Labs Barcelona
- Liangliang Cao, IBM
- Hong Cheng, The Chinese University of Hong Kong
- Alfredo Cuzzocrea, ICAR-CNR & University of Calabria
- Ian Davidson, SUNY
- Gianmarco De Francisci Morales, Yahoo Labs Barcelona
- Nan Du, Georgia Institute of Technology
- Joao Gama, University Porto
- Ricard Gavaldà, Universitat Politècnica de Catalunya
- Fosca Giannotti, ISTI-CNR
- Bart Goethals, University of Antwerp
- Jiawei Han, University of Illinois at Urbana-Champaign
- Marwan Hassani, Aachen University
- Georges Hebrail, Electricité de France - EDF R&D
- Steven C.H. Hoi, Nanyang Technological University
- Dino Ienco, UMR TETIS, Irstea, Montpellier
- Siddhartha Jonnalagadda, Mayo Clinic
- Murat Kantarcioglu, University of Texas at Dallas

- George Karypis, University of Minnesota
- Steve Ko, SUNY at Buffalo
- Vipin Kumar, University of Minnesota, Twin Cities
- Jianhui Li, Computer Network Information Center,Chinese Academy of Sciences
- Cindy Xide Lin, University of Illinois at Urbana-Champaign
- Shou-De Lin, National Taiwan University
- Qiang Ma, Yahoo
- Michael May, Fraunhofer IAIS
- Hassan Ozdemir, Panasonic R&D
- Themis Palpanas, University of Trento
- Fernando Perez-Cruz, University Carlos III
- Bernhard Pfahringer, University of Waikato
- Jesse Read, Aalto University
- Chandan K. Reddy, Wayne State University
- Cyrus Shahabi, USC
- Ashok Srivastava, NASA
- Frederic Stahl, University of Reading
- Jian-Tao Sun, Microsoft Research Asia
- Jie Tang, Tsinghua University
- Hanghang Tong, Carnegie Mellon University
- Joaquin Vanschoren, Eindhoven University of Technology
- Haifeng Wang, Baidu
- Bo Wang, Nanjing University of Aeronautics & Astronautics
- Yi Wang, Tencent
- Xian Wu, Microsoft
- Tian Wu, Baidu
- Zhenghua Xue, Chinese Academy of Sciences
- Gui-Rong Xue, Shanghai Jiao Tong University

- Xifeng Yan, University of California at Santa Barbara
- Rong Yan, Facebook
- Aden Yuen, Tencent
- Demetris Zeinalipour, University of Cyprus
- Xingquan Zhu, University of Technology, Sydney

List of Subreviewers

- Fabiola Pereira
- Priya Govindan

Invited Keynote Speakers

Charles Elkan – University of California and Amazon

Title: From Practice to Theory in Learning from Massive Data

This talk will discuss examples of how Amazon applies machine learning to large-scale data, and open research questions inspired by these applications. One important question is how to distinguish between users that can be influenced, versus those who are merely likely to respond. Another question is how to measure and maximize the long-term benefit of movie and other recommendations. A third question, is how to share data while provably protecting the privacy of users. Note: Information in the talk is already public, and opinions expressed will be strictly personal.

Bio: *Charles Elkan is a professor of computer science at the University of California, San Diego, currently on leave as Amazon Fellow and leader of machine learning for Amazon in Seattle and Silicon Valley. In the past, he has been a visiting associate professor at Harvard. His published research has been mainly in machine learning, data science, and computational biology; the MEME algorithm that he developed with Ph.D. students has been used in over 3000 published research projects in biology and computer science. He is fortunate to have had inspiring undergraduate and graduate students who are in leadership positions now such as vice president at Google.*

Joseph Bradley – Apache Spark PMC

Title: Foundations for Scaling ML in Apache Spark

Apache Spark has become the most active open source Big Data project, and its Machine Learning library MLlib has seen rapid growth in usage. A critical aspect of MLlib and Spark is the ability to scale: the same code used on a laptop can scale to 100's or 1000's of machines. This talk will describe ongoing and future efforts to make MLlib even faster and more scalable by integrating with two key initiatives in Spark. The first is Catalyst, the query optimizer underlying DataFrames and Datasets. The second is Tungsten, the project for approaching bare-metal speeds in Spark via memory management, cache-awareness, and code generation. This talk will discuss the goals, the challenges, and the benefits for MLlib users and developers. More generally, we will reflect on the importance of integrating ML with the many other aspects of big data analysis.

About MLlib: MLlib is a general Machine Learning library providing many ML algorithms, feature transformers, and tools for model tuning and building workflows. The library benefits from integration with the rest of Apache Spark (SQL, streaming, Graph, core), which facilitates ETL, streaming, and deployment. It is used in both ad hoc analysis and production deployments throughout academia and industry.

Bio: *Joseph Bradley is a Software Engineer and Apache Spark PMC member working on machine learning and graph processing at Databricks. Previously, he was a postdoc at UC Berkeley after receiving his Ph.D. in Machine Learning from Carnegie Mellon U. in 2013. His research included probabilistic graphical models, parallel sparse regression, and aggregation mechanisms for peer grading in MOOCs.*

Hanghang Tong – Arizona State University

Title: Inside the Atoms: Mining a Network of Networks and Beyond.

Networks (i.e., graphs) appears in many high-impact applications. Often these networks are collected from different sources, at different times, at different granularities. In this talk, I will present our recent work on mining such multiple networks. First, we will present two models - one on modeling a set of inter-connected networks (NoN); and the other on modeling a set of inter-connected co-evolving time series (NoT). For both models, we will show that by treating networks as context, we are able to model more complicate real-world applications. Second, we will present some algorithmic examples on how to do mining with such new models, including ranking, imputation and prediction. Finally, we will demonstrate the effectiveness of our new models and algorithms in some applications, including bioinformatics, and sensor networks.

Bio: *Hanghang Tong is currently an assistant professor at School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University since August 2014. Before that, he was an assistant professor at Computer Science Department, City College, City University of New York, a research staff member at IBM T.J. Watson Research Center and a Post-doctoral fellow in Carnegie Mellon University. He received his M.Sc and Ph.D. degree from Carnegie Mellon University in 2008 and 2009, both majored in Machine Learning. His research interest is in large scale data mining for graphs and multimedia. He has received several awards, including one 'test of time' award (ICDM 10-Year highest impact paper award), four best paper awards and four 'best of conference'. He has published over 100 referred articles and more than 20 patents.*

He is co-editor of the book "Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques" published by Chapman & Hall/CRC Press.