Preface

4th 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. This 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.

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.

August 2015

Wei Fan, Albert Bifet, Qiang Yang and Philip Yu
BigMine 2015 Program co-Chairs
http://bigdata-mining.org/

BigMine 2015 Workshop Organization

Workshop Chairs

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

Albert Bifet
Huawei Noah's Ark Lab
E-mail: abifet at cs.waikato.ac.nz

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, Huawei Noah's Ark Lab
- 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
- 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
- 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
- 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

- Ayoade Gbadebo
- Gbadebo Ayoade
- Mohit Sharma
- Agoritsa Polyzou
- Yaliang Li
- Chuishi Meng

Sponsors

Gold Sponsors

- TalkingData

TalkingData
Mobile·Data·Value

Invited Keynote Speakers

Xiatian Zhang (TalkingData)

Title: Making Data Talk

TalkingData is China's largest independent Big Data service platform with focus on the mobile Internet. TalkingData offers the best-in-class Big Data products and services varying from mobile app analytics, mobile ad tracking, mobile game analytics, mobile market intelligence, DMP (Data Management Platform), industry consulting, etc. Today, 80% of the Top 50 developers in China rely on TalkingData to track their app metrics, analyze user data points, and optimize monetization. Industry giants such as China Merchants Bank, Citic Bank, Ping An Group also rely on TalkingData's enterprise Big Data solution to build their core data infrastructure towards the mobile era. In this Talk, Xiatian Zhang will introduce the work of TalkingData to mine big mobile data.

Bio: *Xiatian Zhang is Chief Data Scientist of TalkingData at Beijing since 2013. He received the M.S. degree in Computer Science from Being University of Post and Teleommunication in 2007. Before he joined TalkingData, he has worked in IBM CRL, Tencent, and Huawei on research and applications of machine learning and data mining.*

Bernhard Pfahringer, University of Waikato, New Zealand

Title: Why Big Data miners should care about Stream Mining

This talk will provide big data miners with a brief introduction to stream mining. Stream mining is concerned with online learning from non-stationary data sources. This presentation will highlight issues in stream mining, especially around proper evaluation. I will argue that many, if not all, big data mining endeavours will encounter similar issues, and might therefore want borrow ideas and algorithms from data stream mining research area.

Bio: *Bernhard Pfahringer received his PhD degree from the University of Technology in Vienna, Austria, in 1995. He is currently a Professor with the Department of Computer Science at Waikato University in New Zealand. His interests span a range of data mining and machine learning sub-fields, with a focus on streaming, randomization, and complex data.*

Francesco Bonchi, Yahoo Labs, Catalonia

Title: Learning the strength of social influence

With the success of online social networks and microblogging platforms such as Facebook, Tumblr, and Twitter, the phenomenon of influence-driven propagations, has recently attracted the interest of computer scientists, sociologists, information technologists, and marketing specialists. Starting from one of the key problems in this area, i.e. the identification of influential users, we will then focus on the needed step of learning the strength of social influence along each link of a social network. We will discuss why this step is important and challenging. We will present algorithms for learning the social influence in a streaming context, topic models for social influence, and discuss privacy issues associated to this type of analysis.

Bio: *Francesco Bonchi is Director of Research at Yahoo Labs in Barcelona, Spain, where he is leading the Web Mining Research group. His recent research interests include mining query-logs, social networks, and social media, as well as the privacy issues related to mining these kinds of sensible data. In the past he has been interested in data mining query languages, constrained pattern mining, mining spatiotemporal and mobility data, and privacy preserving data mining.*

He is member of the ECML PKDD Steering Committee, Associate Editor of the newly created IEEE Transactions on Big Data (TBD), of the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Intelligent Systems and Technology (TIST), Knowledge and Information Systems (KAIS), and member of the Editorial Board of Data Mining and Knowledge Discovery (DMKD). He has been program co-chair of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010). Dr. Bonchi has also served as program co-chair of the first and second ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007 and 2008), the 1st IEEE International Workshop on Privacy Aspects of Data Mining (PADM 2006), and the 4th International Workshop on Knowledge Discovery in Inductive Databases (KDID 2005).

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

Vincent S. Tseng, National Chiao Tung University, Hsinchu, Taiwan

Title: In-Depth View of Some Key Challenges in Big Data Mining: Perspective from Practical Experiences

Nowadays, large volume of data is being collected at unprecedented and explosive scale in a broad range of application areas. Analytics on such Big Data deliver amazing value and can drive nearly every aspect of our life, including retail, financial services, mobile services, manufacturing, life sciences, etc. Decisions that previously were based on hypothetical models or just unreliable guesswork can now be made effectively and efficiently by utilizing the big data itself. New wave of revolution in information technology has jumped into this Big Data era with new opportunities and challenges arisen. In this talk, I will investigate some key challenges in Big Data Analytcs through in-depth observations from various aspects covering data preprocessing, key feature discovery, learning and modeling, post-processing, etc. Experiences from practical projects in different domains including biomedicine, social media, e-commerce, manufacturing, etc., will also be shared. Finally, some emerging research topics and potential opportunities underlying this topic will also be addressed accordingly.

Bio: *Vincent S. Tseng is currently a Professor at Department of Computer Science in National Chiao Tung University, Taiwan. He has also been the chair for IEEE CIS Tainan Chapter since 2013. He served as the president of Taiwanese Association for Artificial Intelligence during 2011-2012 and acted as the director for Institute of Medical Informatics of NCKU (National Cheng Kung University) during 2008 and 2011. Dr. Tseng received his Ph.D. degree with major in computer science from National Chiao Tung University, Taiwan, in 1997. After that, he joined Computer Science Division of University of California at Berkeley as a postdoctoral research fellow during 1998-1999. He has a wide variety of research interests covering data mining, big data, biomedical informatics, mobile and Web technologies. He has published more than 300 research papers in referred journals and conferences as well as 15 patents (held and filed). He has been on the editorial board of a number of journals including IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Knowledge Discovery from Data, IEEE Journal of Biomedical and Health Informatics, etc. He has also served as chairs/program committee members for a number of premier international conferences related to data mining and intelligent computing, including KDD, ICDM, SDM, PAKDD, ICDE, CIKM, IJCAI, etc. In recent years, Dr. Tseng has also been serving to oversee the planning on Big Data Analytics for the governmental and industrial units in Taiwan. He is also the recipient of 2014 K. T. Li Breakthrough Award.*

Latifur Khan, University of Texas at Dallas, USA

Title: Stream Data Mining and Applications: A Big Data Perspective

Data streams are continuous flows of data. Examples of data streams include network traffic, sensor data, call center records and so on. Data streams demonstrate several unique properties that together conform to the characteristics of big data (i.e., volume, velocity, variety and veracity) and add challenges to data stream mining. In this talk we will present an organized picture on how to handle various data mining techniques in data streams. Most existing data stream classification techniques ignore one important aspect of stream data: arrival of a novel class. We address this issue and propose a data stream classification technique that integrates a novel class detection mechanism into traditional classifiers, enabling automatic detection of novel classes before the true labels of the novel class instances arrive. Novel class detection problem becomes more challenging in the presence of concept-drift, when the underlying data distributions evolve in streams. In this talk we will show how to make fast and correct classification decisions under this constraint. Furthermore, we will present a semi supervised framework which exploits change detection on classifier confidence values to update the classifier intelligently with limited labeled training data.

We will present a number of stream classification applications such as website fingerprinting, real time monitoring, evolving insider threat detection and textual stream classification.

This research was funded in part by USA National Science Foundation (NSF), NASA, Air Force Office of Scientific Research (AFOSR), Sandia National Lab (via DOE) and Raytheon.

Bio: *Dr. Latifur Khan is currently a full Professor (tenured) in the Computer Science department at the University of Texas at Dallas, USA where he has been teaching and conducting research since September 2000. He received his Ph.D. and M.S. degrees in Computer Science from the University of Southern California (USC) in August of 2000, and December of 1996 respectively. Dr. Khan is an ACM Distinguished Scientist. He has received prestigious awards including the IEEE Technical Achievement Award for Intelligence and Security Informatics. Dr. Khan has published over 200 papers in prestigious journals, and in peer reviewed conference proceedings. Currently, his research area focuses on big data management and analytics, data mining, complex data management including geo-spatial data and multimedia data. More details can be found at:* www.utdallas.edu/~lkhan/

Jingrui He, Arizona State University, USA

Title: Heterogeneous Learning: Algorithms and Applications

Data heterogeneity is common across many high-impact real applications, ranging from security to manufacturing, from healthcare to traffic analytics, and it is closely related to the 'Variety' aspect of big data. Such heterogeneity can be presented in a variety of forms, including task heterogeneity, where multiple related tasks may form a hierarchical structure; view heterogeneity, where information is being collected from various sources; instance heterogeneity, where the label of a single example can be decomposed into a set of heterogeneous labels associated with composing instances; etc. In this talk, I will introduce how we model data heterogeneity from a holistic perspective. In particular, I will hinge on multiple applications, discuss the major challenges being shared by these applications that are related to data heterogeneity, and introduce our proposed techniques for addressing these challenges.

Bio: *Jingrui He is an Assistant Professor in the School of Computing, Informatics, DecisionSystems Engineering at Arizona State University. She received the Ph.D degree from School of Computer Science, Carnegie Mellon University in 2010. Her research interests include heterogeneous machine learning and rare category analysis with applications in semiconductor manufacturing, social media analysis, healthcare, traffic prediction, etc. She is the recipient of the 2014 IBM Faculty Award and is the author of the book on Analysis of Rare Categories (Springer-Verlag, 2011). She has served on the organizing/senior program committees of many conferences, including ICML, KDD, IJCAI, ICDM, SDM, etc.*