17th Asia-Pacific Web Conference

(APWeb 2015)

September 18-20,

Guangzhou, China

**Foreword**

Welcome to the 17th Asia-Pacific Web Conference (APWeb 2015). As the 17th edition of the conference, APWeb has become a leading conference to host top researchers, as well as industrial players to exchange the experience with edge-cutting technologies for web-based applications and web data processing. The prosperity of big data has also propelled the technical advances of computer science on web technologies, which is reflected by the passions presented by the attendants and speakers in APWeb. For the first time for APWeb, China Telecom, as one of the largest telecommunication companies in China, participates as a main sponsor and organizer of this academia conference, and actively contributes to the first data challenge competition in APWeb. Driven by the huge interests from industry, the research investigations on web technologies are moving towards practical technologies to support the potential businesses in the real world. Text analytics and recommendation systems have been the most popular topics discussed in the research works in the conference, which contains both research and practical values with the emergence of Internet economy in the world.

The APWeb conference provides a unique venue for researchers and practitioners from all over the world. The conference successfully attracts submissions from Canada, China, Hong Kong, India, Japan, Luxembourg and other countries. The program committee accepts 67 submissions in the research track, 3 submissions in industrial track and 8 submissions in the demonstration track. All these high-quality works cover a wide spectrum of web-related data management problems, and provide a thorough view on the rapid advances of technical solutions.

We would like to thank all authors for their kind contributions. We would like to express our ultimate gratitude to the program committee members for their huge efforts on reviewing and discussions on the submissions. Special thanks to our keynote speakers, Prof. Elisa Bertino, Prof. Wenfei Fan and Prof. Xindong Wu. We also would like to express our gratitude to the local organizers and sponsors.

We hope the success of APWeb 2015 enhances the research on web-based data management technologies, consolidates the connections between academia and industry, and creates sparks with idea exchange between all participants.

APWeb 2015 Program Co-Chairs,

Reynold Cheng, Bin Cui, Zhenjie Zhang

**General Co-Chairs**

|  |  |
| --- | --- |
| Kang Cai | China Telecom Co. Ltd., China |
| Zhifeng Hao | Guangdong University of Technology, China |
| Xuemin Lin | University of New South Wales, Australia |

**Program Co-Chairs**

|  |  |
| --- | --- |
| Reynold Cheng | The University of Hong Kong, Hong Kong |
| Bin Cui | Peking University, China |
| Zhenjie Zhang | Advanced Digital Sciences Center, Singapore |

**Workshop Co-Chairs**

|  |  |
| --- | --- |
| Xiaokui Xiao | Nanyang Technological University, Singapore |
| Jian Yin | Sun Yat-Sen University, China |

**Demo Co-Chairs**

|  |  |
| --- | --- |
| Xiaoli Li | Institute for Infocomm Research, Singapore |
| Xiaojun Xiao | Guangzhou Useease Information Technology Co. Ltd. |
| Rong Zhang | East China Normal University, China |

**Panel Co-Chairs**

|  |  |
| --- | --- |
| Zexiong Ma | China Telecom Co. Ltd., China |
| Jeffrey Xu Yu | The Chinese University of Hong Kong, Hong Kong |
| Zhiming Zhong | China Telecom Co. Ltd., China |

**DYL Series Chair**

|  |  |
| --- | --- |
| Aoying Zhou | East China Normal University, China |

**Industry Co-Chairs**

|  |  |
| --- | --- |
| Hongyang Chao | Sun Yat-Sen University, China |
| Junjun Hu | China Telecom Co. Ltd., China |
| Ying Yan | Microsoft Research Asia, China |

**Publication Co-Chairs**

|  |  |
| --- | --- |
| Ruichu Cai | Guangdong University of Technology, China |
| Jia Xu | Guangxi University, China |

**Big Data Challenge Co-Chairs**

|  |  |
| --- | --- |
| Kang Chen | China Telecom Co. Ltd., China |
| Baozhong Wang | China Telecom Co. Ltd., China |

**Publicity Co-Chairs**

|  |  |
| --- | --- |
| Ilaria Bartollini | University of Bologna, Italy |
| Yin Yang | Hamad bin Khalifa University, Qatar |
| Bin Zhou | University of Maryland, Baltimore County, USA |

**Local Organization Co-chairs**

|  |  |
| --- | --- |
| Xiaohui Guo | China Telecom Co. Ltd., China |
| Jie Ling | Guangdong University of Technology, China |
| Lijuan Wang | Guangdong University of Technology, China |
| Wen Wen | Guangdong University of Technology, China |

**Research Student Symposium Co-Chairs**

|  |  |
| --- | --- |
| Xiaohui Yu | Shandong University, China |
| Wenjie Zhang | University of New South Wales, Austrilia |

**Research Program Committee**

|  |  |
| --- | --- |
| Toshiyuki Amagasa | University of Tsukuba, Japan |
| **Zhifeng Bao** | University of Tasmania, Australia |
| **Arnab Bhattacharya** | Indian Institute of Technology, Kanpur, India |
| **Sourav Bhowmick** | Nanyang Technological University, Singapore |
| **Tru Hoang Cao** | Ho Chi Minh City University of Technology, Vietnam |
| **Kang Chen** | China Telecom Co, China |
| **Rui Chen** | Hong Kong Baptist University, Hong Kong |
| **Wenliang Chen** | Soochow University, China |
| **Yueguo Chen** | Renmin University, China |
| **James Cheng** | The Chinese University of Hong Kong, Hong Kong |
| **Zhihong Chong** | Southeast University, China |
| **Bing Tian Dai** | Singapore Management University, Singapore |
| **Markus Endres** | Univresity of Augsburg, Germany |
| **Yuan Fang** | Institute for Infocomm Research, Singapore |
| **Tom Z. J. Fu** | Advanced Digital Sciences Center, Singapore |
| **Gabriel Ghinita** | University of Massachusetts at Boston, USA |
| **João Bártolo Gomes** | Institute for Infocomm Research, Singapore |
| **Jun Gao** | Peking University, China |
| **Shenghua Gao** | Shanghait Tech University, China |
| **Yu Gu** | Northeast University, China |
| **Lifang He** | University of Illinois at Chicago, USA |
| **Bao-Quoc Ho** | VNU-HCMUS, Vietnam |
| **Liang Hong** | Wuhan University, China |
| **Haibo Hu** | Hong Kong Baptist University, Hong Kong |
| **Zhuolin Jiang** | Noah's Ark Lab, Huawei Technologies, Hong Kong |
| **Mizuho Iwaihara** | Waseda University, Japan |
| **Yoshiharu Ishikawa** | Nagoya University, Japan |
| **Yiping Ke** | Nanyang Technological University, Singapore |
| **Hady Lauw** | Singapore Management University, Singapore |
| **Sangkeun Lee** | Oak Ridge National Laboratory, USA |
| **Wookey Lee** | Inha University, Korea |
| **Feng Li** | Microsoft Research, USA |
| **Guoliang Li** | Tsinghua University, China |
| **Chen Lin** | Xiamen University, China |
| **Zhiyong Lin** | Guangdong Polytechnic Normal University, China |
| **Bo Liu** | Guangdong University of Technology, China |
| **Hua Lu** | Aalborg University, Denmark |
| **Mihai Lupu** | Vienna University of Technology, Austria |
| **Shuai Ma** | Beihang University, China |
| **Zakaria Maamar** | Zayed University, United Arab Emirates |
| **Alexander Markowetz** | Bonn University, Germany |
| **Anirban Mondal** | Xerox Research Center India, India |
| **Yang-Sae Moon** | Kangwon National University, Korea |
| **Yasuhiko Morimoto** | Hiroshima University, Japan |
| **Shinsuke Nakajima** | Kyoto Sangyo University, Japan |
| **Hiroaki Ohshima** | Kyoto University, Japan |
| **Sanghyun Park** | Yonsei University, Korea |
| **Dhaval Patel** | IIT-R, India |
| **Stavros Papadopoulos** | Intel Labs, USA |
| **Chenxi Qi** | China Telecom Co, China |
| **Jianzhong Qi** | University of Melbourne, Australia |
| **Weining Qian** | East China Normal University, China |
| **Krishna Reddy** | IIIT Hyderabad, India |
| **Chaofeng Sha** | Fudan University, China |
| **Nan Tang** | Qatar Computing Research Institute, Qatar |
| **Quan Thanh Tho** | Ho Chi Minh City University of Technology, Vietnam |
| **James Thom** | RMIT, Australia |
| **Quan Thanh Tho** | Ho Chi Minh City University of Technology, Vietnam |
| **Alex Thomo** | University of Victoria, Australia |
| **Leong Hou U** | University of Macau, Macau |
| **Taketoshi Ushiama** | Kyushu University, Japan |
| **Quang-Hieu Vu** | Khalifa University, United Arab Emirates |
| **Baozhong Wang** | China Telecom Co, China |
| **Hongzhi Wang** | Harbin Institue of Technology, China |
| **Junhu Wang** | Griffith University, Australia |
| **Raymond Wong** | Hong Kong University of Science and Technology, Hong Kong |
| **Xiaochun Yang** | Northeast University, China |
| **Xiaowei Yang** | South China University of Technology, China |
| **Yunjie Yao** | East China Normal University, China |
| **Hongzhi Yin** | University of Queensland, Australia |
| **Man Lung Yiu** | Hong Kong Polytechnical University, Hong Kong |
| **Shanshan Ying** | Advanced Digital Sciences Center, Singapore |
| **Haruo Yokota** | Tokyo Institute of Technology, Japan |
| **Xiaoyan Yang** | Advanced Digital Sciences Center, Singapore |
| **Ting Yu** | Qatar Computing Research Institute, Qatar |
| **Ganzhao Yuan** | South China University of Technology, China |
| **Dongxiang Zhang** | National University of Singapore, Singapore |
| **Meihui Zhang** | Singapore Univrsity of Technology and Design, Singapore |
| **Rui Zhang** | University of Melbourne, Australia |
| **Wen Zhang** | Wuhan University, China |
| **Vicent Zheng** | Advanced Digital Sciences Center, Singapore |
| **Jia Zhu** | South China Normal University, China |
| **Feida Zhu** | Singapore Management University, Singapore |
| **Lei Zou** | Peking University, China |

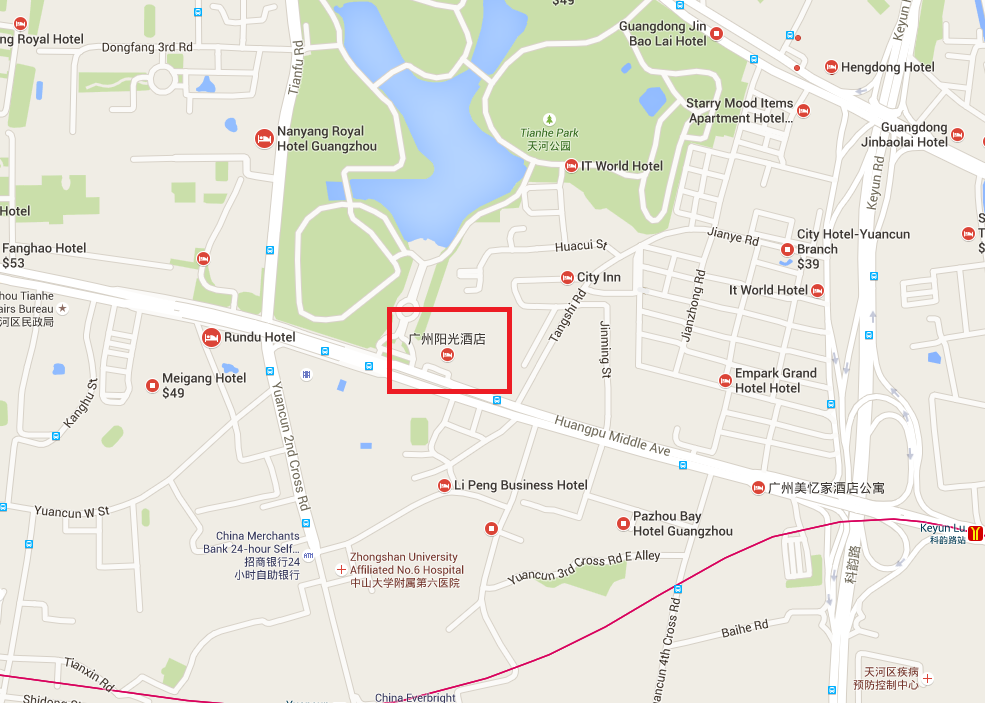
**Industrial Program Committee**

|  |  |
| --- | --- |
| Zhao Cao | IBM Research, China |
| Peng Jiang | Alibaba Group, China |
| Bin Liu | Google Co., USA |
| Shin-Ming Liu | Intel Labs, China |
| Tantan Liu | Google Co., USA |
| Yue Lu | Twitter Co., USA |
| Ping Luo | HP Research Labs, China |
| Junfeng Pan | Facebook Co., USA |

**Demonstration Program Committee**

|  |  |
| --- | --- |
| Ming Gao | East China Normal University, China |
| Xiaofeng He | East China Normal University, China |
| Kyoung-Sook Kim | National Institute of Advanced Industrial Science and Technology, South Korea |
| Sanjay Kumar Madria | Missouri University of Science and Technology, USA |
| Yang-Sae Moon | Kangwon National University, South Korea |
| Wei Pan | Northwestern Polytechnical University, China |
| Bin Yang | Aalborg University, Denmark |
| Kai (Keivn) Zheng | University of Queensland, Australia |

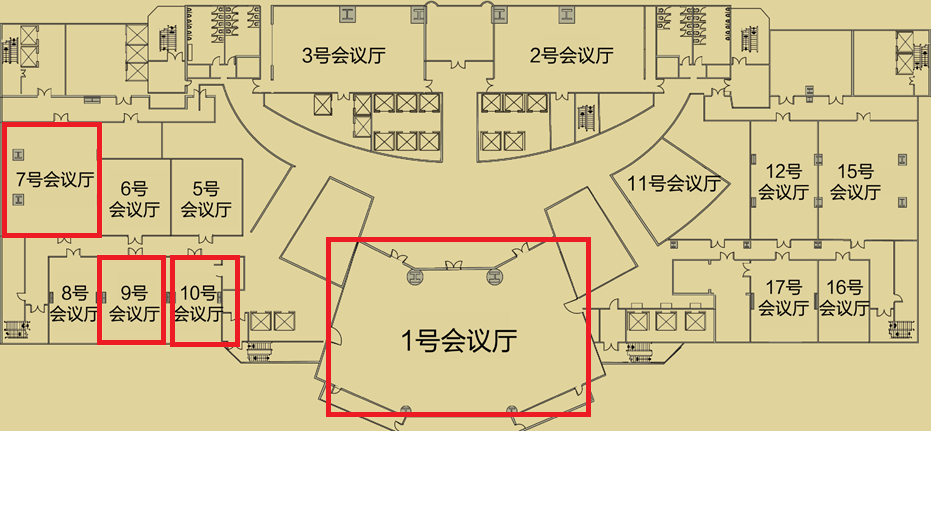
**Conference Venue**



Soluxe Hotel Guangzhou is situated in the central business district between Huangpu Ave. and Guangzhou Ave. It is right next to the glamorous Tianhe Park, and faces the famed Pazhou International Exhibition Centre across the Pearl River. There are more than 20 bus lines, and the two stations (Yuancun station and Keyun Road metro station) near the hotel, which offer the best convenience on transportation. The hotel is opened in 2011, and offers 403 rooms and suits for quality accommodations.

**Address**: No.199 Huangpu Avenue Middle, Tianhe District, 510655, Guangzhou, China

**Tel**:+86-20-38018888



All sessions are held on Level 7 (Room 1, Room 7, Room 9 and Room 10) of Yangguang Hotel

**Overall Schedule**

Day 1

|  |  |  |  |
| --- | --- | --- | --- |
| 9:30-10:30 | BDAT 2015 | WDMA 2015 | BSD 2015 |
| 10:30-11:00 | Coffee Break | | |
| 11:00-12:00 | BDAT 2015 | Tutorial 1 | Tutorial 4 |
| 12:30-13:30 | Lunch | | |
| 13:30-15:30 | Industry 1 | Tutorial 2 | Tutorial 3 |
| 15:30-16:00 | Coffee Break | | |
| 16:00-17:30 | Research 1 | Research 2 | Research 3 |
| 19:00-21:00 | Reception | | |

Day 2

|  |  |  |  |
| --- | --- | --- | --- |
| 8:30-9:00 | Opening Ceremony | | |
| 9:00-10:00 | Keynote 1 | | |
| 10:00-10:30 | Coffee Break | | |
| 10:30-12:00 | Research 4 | Research 5 | Research 6 |
| 12:00-13:30 | Lunch | | |
| 13:30-14:30 | Keynote 2 | | |
| 14:30-15:00 | Coffee Break | | |
| 15:00-16:30 | Research 7 | Research 8 | Research 9 |
| 16:30-17:00 | Coffee Break | | |
| 17:00-18:00 | Research 10 | Research 11 | Research 12 |
| 19:00-22:00 | Banquet | | |

Day 3

|  |  |  |  |
| --- | --- | --- | --- |
| 9:00-10:00 | Keynote 3 | | |
| 10:00-10:30 | Coffee Break | | |
| 10:30-11:30 | Panel | | |
| 11:30-13:00 | Lunch | | |
| 13:00-14:30 | Research 13 | Research 14 | Research 15 |
| 14:30-15:00 | Coffee Break | | |
| 15:00-16:00 | Distinguished Young Lecturer Series | Demonstrations | APWeb Data Challenge |
| 16:00-16:30 | Coffee Break | | |
| 16:30-17:50 | Research 16 | Research 17 | Research 18 |

**Session Schedule**

Day 1 (September 18, 2015):

9:00 – 10:30

Workshop Session 1: Big data applications in Telecoms (BDAT 2015)

Workshop Session 2: Web Data Mining and Applications (WDMA 2015)

Workshop Session 3: Big Social Data (BSD 2015)

10:30 – 11:00 Coffee break

11:00 – 12:00

Workshop Session 4: Big data applications in Telecoms (BDAT 2015)

Tutorial Session 1: Indexing for Querying Metric Spaces

Tutorial Session 4: On Repairing Structural Problems In Semi-structured Data

12:30 – 13:30 Lunch

13:30 – 15:30

Industrial Session 1: Cloud and Real-Time Computation

Tutorial Session 2: Influence computation in spatial databases

Tutorial Session 3: Providing Retrievability Guarantees in Cloud-Based Storage Services

15:30 – 16:00 Coffee Break

16:00 – 17:30

Research Session 1: Social Prediction

Research Session 2: Text Analysis I

Research Session 3: Learning from Data

19:00 – 21:00 Reception

Day 2 (September 19, 2015)

8:30 – 9:00

Opening Ceremony

9:00 – 10:00

Keynote 1: Big Data - Security and Privacy

10:00 – 10:30 Coffee Break

10:30 – 12:00

Research Session 4: Hardware, Schema and Query

Research Session 5: Topic Model

Research Session 6: Metric Learning

12:00 – 13:30 Lunch and Posters

13:30 – 14:30

Keynote 2: Personalized Web News Filtering and Summarization

14:30 – 15: 00 Coffee Break

15:00 – 16:30

Research Session 7: Semantics

Research Session 8: Query Processing

Research Session 9: Recommendation System I

Demo Session

16:30 – 17:00 Coffee Break

17:00 – 18:00

Research Session 10: Semanic Web and Knowledge Base

Research Session 11: Text Analysis II

Research Session 12: MapReduce

19:00 Banquet

Day 3 (September 20, 2015)

9:00 – 10:00

Keynote 3: Querying Big Data: From Theory and Practice

10:00 – 10:30 coffee break

10:30 – 11:30

Panel Session: Cradle of Bubble? Where Is Big Data Age Going?

11:30 – 13:00 Buffet Lunch

13:00 – 14:30

Research Session 13: Social Network

Research Session 14: Recommendation System II

Research Session 15: Location-Based Service

14:30 – 15:00 Coffee Break

15:00 – 16:00

Distinguished Young Lecturer Session

Demonstration Session

APWeb Big Data Challenge Session

16:00 – 16:30 Coffee Break

16:30 – 17:50

Research Session 16: Graph Processing

Research Session 17: Data Mining

Research Session 18: Influence Maximization

**Research Session Information**

Research Session 1: Social Prediction (chaired by **Shanshan Ying**)

Hashtag Sense Induction Based on Co-Occurrence Graphs*, Mengmeng Wang, Mizuho Iwaihara*

Early Outbreak Detection in Social Networks, *Chuan Zhou*

A Multi-view Retweeting Behaviors Prediction in Social Networks, *Bo Jiang, Ying Sha, Lihong Wang*

Research Session 2: Text Analysis I (chaired by **Xiaoyan Yang**)

A Multi-news Timeline Summarization Algorithm Based on Aging Theory, Jie Chen, Zhendong Niu, Hongping Fu

Extended Strategies for Document Clustering with Word Co-occurrences, *Yang Wei, Jinmao Wei, Zhenglu Yang*

User Generated Content Oriented Chinese Taxonomy Construction, *Jinyang Li, Chengyu Wang, Xiaofeng He, Rong Zhang, Ming Gao*

Research Session 3: Learning from Data (chaired by **Xiaowei Yang**)

Discovering Restricted Regular Expressions with Interleaving, *Feifei Peng, Haiming Chen*

Spare Part Demand Prediction based on Context-aware Matrix Factorization, *Jianwei Ding, Yingbo Liu, Yuan Cao, Li Zhang, Jianmin Wang*

Online Feature Selection based on Passive-Aggressive Algorithm with Retaining Features, *Hai-Tao Zheng, Haiyang Zhang*

Batch Mode Active Learning For Geographical Image Classification, *Bo Du, Zengmao Wang, Wenbin Hu, Lefei Zhang, Liangpei Zhang, Dacheng Tao*

Research Session 4: Hardware, Schema and Query (chaired by **Wen Wen**)

Efficient Buffer Management for PCM-Enhanced Hybrid Memory Architecture, *Kaimeng Chen, Peiquan Jin, Lihua Yue*

DistDL: A Distributed Deep Learning Service Schema with GPU Accelerating, *Jianzong Wang, Lianglun Cheng*

Overlapping Schema Summarization based on Multi-label Propagation, *Man Yu, Chao Wang, Xiangrui Cai, Yanlong Wen, Ying Zhang, Xiaojie Yuan*

RDQS: A Relevant and Diverse Query Suggestion Generation Framework, *Yichi Zhang, Hai-Tao Zheng*

Research Session 5: Topic Model (chaired by **Lijuan Wang**)

An Online Inference Algorithm for Labeled Latent Dirichlet Allocation, *Qiang Zhou, Heyan Huang, Xianling Mao*

Multi-Label Emotion Tagging for Online News by Supervised Topic Model, *Ying Zhang, Lili Su, Zhifan Yang, Xue Zhao, Xiaojie Yuan*

Distinguishing Specific and Daily topics, *Tao Ge, Wenzhe Pei, Baobao Chang, Zhifang Sui*

A Supervised Parameter Estimation Method of LDA, Zhenyan Liu, *Dan Meng*,*Weiping Wang*,*Chunxia Zhang*

Research Session 6: Metric Learning (chaired by **Ruichu Cai**)

Tree-Based Metric Learning for Distance Computation in Data Mining, *Ming Yan, Zhang Yan, Hongzhi Wang*

Learning Similarity Functions for Urban Events Detection by Mining Hotline Phone Records, *Pengjie Ren, Peng Liu, Zhumin Chen, Jun Ma, Xiaomeng Song*

A New Similarity Measure Between Semantic Trajectories based on Road Networks, *Wu Xia, Yuanyuan Zhu, Shengchao Xiong, Yuwei Peng, Zhiyong Peng*

Multiple Attribute Aware Personalized Ranking, *Weiyu Guo, Shu Wu, Liang Wang, Tieniu Tan*

Research Session 7: Semantics (chaired by **Xiaofang Zhou**)

An Ensemble Matchers based Rank Aggregation Method for Taxonomy Matching, *Hailun Lin, Yuanzhuo Wang, Yantao Jia, Jinhua Xiong, Peng Zhang, Xueqi Cheng*

Random-based Algorithm for Efficient Entity Matching, *Pingfu Chao, Zhu Gao, Yuming Li, Junhua Fang, Rong Zhang, Aoying Zhou*

Research on Semantic Disambiguation in Treebank, *Lin Miao, Xueqiang Lv, Yunfang Wu, Yue Wang*

An Self-Learning Rule-based Approach for Sci-tech Compound Phrase Entity Recognition, *Tingwen Liu, Yang Zhang, Yang Yan, Jinqiao Shi, Li Guo*

Research Session 8: Query Processing (chaired by **Reynold Cheng**)

Efficient Location-dependent Skyline Queries in Wireless Broadcast Environments, *Yingyuan Xiao, Pengqiang Ai, Hongya Wang, Ching-Hsien Hsu, Wenxiang Cui*

Efficient Algorithms for Distance-based Representative Skyline Computation in 2D Space, *Taotao Cai, Rong-Hua Li, Jeffrey Xu Yu, Rui Mao, Yadi Cai*

A Compression-Based Filtering Mechanism in Content-Based Publish/Subscribe System, *Qin Liu, Yiwen Zheng, Kaile Wang*

Hashing Multi-Instance Data from Bag and Instance Level, *Yao Yang, Xin-Shun Xu, Xiaolin Wang, Shanqing Guo, Lizhen Cui*

Research Session 9: Recommendation System I (chaired by **Xuemin Lin**)

MATAR: Keywords Enhanced Multi-Label Learning for Tag Recommendation, *Licheng Li, Yuan Yao, Feng Xu, Jian Lu*

Simple is Beautiful: An Online Collaborative Filtering Recommendation Solution with Higher Accuracy, *Feng Zhang, Ti Gong, Victor Lee, Gansen Zhao, Guangzhi Qu*

PDMA: A Probabilistic Framework for Diversifying Recommendation Lists, *Yang-Hui Yan, Hai-Tao Zheng, Ying-Min Zhou*

A Semi-Supervised Solution for Cold Start Issue on Recommender Systems, *Zhifeng Hao, Yingchao Cheng, Ruichu Cai, Wen Wen, Lijuan Wang*

Research Session 10: Semantic Web and Knowledge Base (chaired by **Lei Zou**)

On The Marriage of SPARQL and Keywords, *Peng Peng, Lei Zou, Dongyan Zhao*

On Coherent Indented Tree Visualization of RDF Graphs, *Qingxia Liu, Gong Cheng, Yuzhong Qu*

Knowledge Base Completion Using Matrix Factorization, *Wenqiang He, Yansong Feng, Lei Zou, Dongyan Zhao*

Research Session 11: Text Analysis II (chaired by **Zhenjie Zhang**)

Matching Reviews to Object Based on 2-Stage CRF, *Yongxin Zhang, Qingzhong Li, Dequan Wang, Yanhui Ding, Congli Liu, Zhongmin Yan*

Sentiment Word Identification with Sentiment Contextual Factors, *Jiguang Liang, Xiaofei Zhou, Yue Hu, Li Guo, Shuo Bai*

Sentiment Classification for Chinese Product Reviews Based on Semantic Relevance of Phrase, *Heng Chen, Hai Jin, Pingpeng Yuan, Lei Zhu, Hang Zhu*

Research Session 12: MapReduce (chaired by **Yin Yang**)

Distributed XML Twig Query Processing using MapReduce, *Xin Bi, Guoren Wang, Xiangguo Zhao, Zhen Zhang, Shuang Chen*

Large-Scale Graph Classification based on Evolutionary Computation with MapReduce, *Zhanghui Wang, Yuhai Zhao, Guoren Wang, Yurong Cheng*

A Lightweight Evaluation Framework for Table Layouts in MapReduce Based Query Systems, *Feng Zhu, Jie Liu, Lijie Xu, Dan Ye, Jun Wei, Tao Huang*

Research Session 13: Social Network (chaired by **Zhixu Li**)

Sleep Quality Evaluation of Active Microblog Users, *Kai Wu, Jun Ma, Zhumin Chen, Pengjie Ren*

Analysis of Subjective City Happiness Index Based on Large Scale Microblog Data, *Kai Wu, Jun Ma, Zhumin Chen, Pengjie Ren*

Location Sensitive Friend Recommendation in Social Network, *Xueqin Sui, Zhumin Chen, Jun Ma*

A Secure and Efficient Framework for Privacy Preserving Social Recommendation, *Shushu Liu, An Liu, Zhixu Li, Lei Zhao, Guanfeng Liu, Pengpeng Zhao, Jiajie Xu*

Research Session 14: Recommendation System II (chaired by **Yunjun Gao**)

Effective Citation Recommendation by Unbiased Reference Priority Recognition, *Wenyang Lu, Yubin Yang, XiaoJiao Mao, Qihai Zhu*

Trustworthy Collaborative Filtering through Downweighting Noise and Redundancy, *Qiuxiang Dong, Zhi Guan, Zhong Chen*

User Behavioral Context-Aware Service Recommendation for Personalized Mashups in Pervasive Environments, *Wei He, Guozhen Ren, Lizhen Cui, Hui Li*

Online Personalized Recommendation Based on Streaming Implicit User Feedback, *Zhisheng Wang, Qi Li, Ye Liu, Wei Liu, Jian Yin*

Research Session 15: Location-Based Service (chaired by **Muhammad Aamir Cheema**)

Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks, *Yang Zhang, Jun Pang*

Hybrid-LSH for Spatio-Textual Similarity Queries, *Mingdong Zhu, Derong Shen, Ling Liu, Ge Yu*

Answering Spatial Approximate Keyword Queries on Disks, *Jinbao Wang, Donghua Yang, Yuhong Wei, Hong Gao, Jianzhong Li, Ye Yuan*

Reverse Direction-based Surrounder Queries, *Xi Guo, Yoshiharu Ishikawa, Aziguli Wulamu, Yonghong Xie*

Research Session 16: Graph Processing (chaired by **Jianliang Xu**)

GSCS - Graph Stream Classification with Side Information, *Amit Mandal, Mehedi Hasan, Anna Fariha, Chowdhury Farhan Ahmed*

An Extended Graph-based Label Propagation Method for Readability Assessment, *Zhiwei Jiang, Gang Sun, Qing Gu, Lixia Yu, Daoxu Chen*

AILabel: A Fast Interval Labeling Approach for Reachability Query on Very Large Graphs, *Feng Shuo, Derong Shen*

Graph-based Hybrid Recommendation Using Random Walk and Topic Modeling, *Yang-Hui Yan, Hai-Tao Zheng, Ying-Min Zhou*

Research Session 17: Data Mining (chaired by **Jeffrey Yu**)

Mining Weighted Frequent Itemsets with the Recency Constraint, *Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong*

Boosting Explicit Semantic Analysis by Clustering Paragraph Vectors of Wikipedia Articles, *Hai-Tao Zheng, Wenzhen Wu*

Probabilistic Frequent Pattern Mining by PUH-Mine, *Wenzhu Tong, Carson Leung, Dacheng Liu, Jialiang Yu, Fan Jiang*

Learning to Hash for Recommendation with Tensor Data, *Qiyue Yin, Shu Wu, Liang Wang*

Research Session 18: Influence Maximization (chaired by **Yoshiharu Ishikawa**)

A Co-Ranking Framework to Select Optimal Seed Set for Influence Maximization in Heterogeneous Network, *Yashen Wang, Heyan Huang, Chong Feng, Xianxiang Yang*

UserGreedy : Exploiting the Activation Set to Solve Influence Maximization Problem, *Wenxin Liang, Chengguang Shen, Xianchao Zhang*

Minimizing the Cost to Win Competition in Social Network, *Ziyan Liu, Xiaguang Hong, Zhaohui Peng, Zhiyong Chen, Weibo Wang, Tianhang Song*

Ad Dissemination Game in Ephemeral Networks, *Lihua Yin, Yunchuan Guo, Yanwei Sun, Junyan Qian, Thanos Vasilakos*

Industrial Session 1: Cloud and Real-Time Computation (chaired by **Tom Fu**)

Hybrid Cloud Deployment of an ERP-based Student Administration System, *Simon K.S. Cheung*

A Benchmark Evaluation of Enterprise Cloud Infrastructure, *Yong Chen, Xuanzhong Xie, Jiayin Wu*

A Fast Data Ingestion and Indexing Scheme for Real-time Log Analytics, *Haoqiong Bian, Yueguo Chen, Xiongpai Qin, Xiaoyong Du*

Demonstration Session:

A Fair Data Market System with Data Quality Evaluation and Repairing Recommendation, *Xiaoou Ding, Hongzhi Wang, Dan Zhang, Jianzhong Li, Hong Gao*

HouseIn: A Housing Rental Platform with Non-Redundant Information Integrated from Multiple Sources, *Jian Zhou, Zhixu Li, Qiang Yang, Jun Jiang, An Liu, Guanfeng Liu, Lei Zhao, Jia Zhu*

ONCAPS: An Ontology-based Car Purchase Guiding System, *Jianfeng Du, Jun Zhao, Jiayi Cheng, Qingchao Su, Jiacheng Liang*

A Multiple Trust Paths Selection Tool in Contextual Online Social Networks, *Linlin Ma, Guanfeng Liu, Guohao Sun, Lei Li, Zhixu Li, An Liu, Lei Zhao*

EPEMS: An Entity Matching System For E-Commerce Products, *Lei Gao, Pengpeng Zhao, Victor S. Sheng, Zhixu Li, An Liu, Jian Wu, Zhiming Cui*

PPS-POI-Rec: A Privacy Preserving Social Point-of-Interest Recommender System, *Xiao Liu, An Liu, Guanfeng Liu, Zhixu Li, Jiajie Xu, Pengpeng Zhao, Lei Zhao*

Incorporating Contextual Information into a Mobile Advertisement Recommender System, *Ke Zhu, Yingyuan Xiao, Pengqiang Ai, Hongya Wang, Ching-Hsien Hsu*

**Detailed Program**

Keynote Session 1:

**Title**: Big Data - Security and Privacy

**Speaker**: *Prof. Elisa Bertino* (Purdue University, USA)

**Abstract**: Technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, and to quickly process and analyze huge amounts of data from which to extract information critical for security-related tasks. In the area of cyber security, such tasks include user authentication, access control, anomaly detection, user monitoring, and protection from insider threat. By analyzing and integrating data collected on the Internet and Web one can identify connections and relationships among individuals that may in turn help with homeland protection. By collecting and mining data concerning user travels and disease outbreaks one can predict disease spreading across geographical areas. And those are just a few examples; there are certainly many other domains where data technologies can play a major role in enhancing security. The use of data for security tasks is however raising major privacy concerns. Collected data, even if anonymized by removing identifiers such as names or social security numbers, when linked with other data may lead to re-identify the individuals to which specific data items are related to. Also, as organizations, such as governmental agencies, often need to collaborate on security tasks, data sets are exchanged across different organizations, resulting in these data sets being available to many different parties. Apart from the use of data for analytics, security tasks such as authentication and access control may require detailed information about users. An example is multi-factor authentication that may require, in addition to a password or a certificate, user biometrics. Recently proposed continuous authentication techniques extend access control system. This information if misused or stolen can lead to privacy breaches. It would then seem that in order to achieve security we must give up privacy. However this may not be necessarily the case. Recent advances in cryptography are making possible to work on encrypted data – for example for performing analytics on encrypted data. However much more needs to be done as the specific data privacy techniques to use heavily depend on the specific use of data and the security tasks at hand. Also, current techniques are not still able to meet the efficiency requirement for use with big data sets. In this talk we will discuss methods and techniques to make this reconciliation possible and identify research directions.

**Biography**: Elisa Bertino is professor of computer science at Purdue University, and serves as Director of Purdue Cyber Center and Research Director of the Center for Information and Research in Information Assurance and Security (CERIAS). She is also an adjunct professor of Computer Science & Info tech at RMIT. Prior to joining Purdue in 2004, she was a professor and department head at the Department of Computer Science and Communication of the University of Milan. She has been a visiting researcher at the IBM Research Laboratory (now Almaden) in San Jose, at the Microelectronics and Computer Technology Corporation, at Rutgers University, at Telcordia Technologies. Her recent research focuses on database security, digital identity management, policy systems, and security for web services. She is a Fellow of ACM and of IEEE. She received the IEEE Computer Society 2002 Technical Achievement Award and the IEEE Computer Society 2005 Kanai Award. She is currently serving as EiC of IEEE Transactions on Dependable and Secure Computing

Keynote Session 2:

**Title**: Personalized Web News Filtering and Summarization

**Speaker**: *Prof. Xindong Wu* (University of Vermont, USA; Hefei University of Technology, China)

**Abstract**: Information on the World Wide Web is congested with large amounts of news contents. Recommending, filtering, and summarization of Web news have become hot topics of research in Web intelligence, aiming to find interesting news for users and give concise content for reading. This talk presents our research on developing the Personalized News Filtering and Summarization system (PNFS). An embedded learning component of PNFS induces a user interest model and recommends personalized news. Two Web news recommendation methods are proposed to keep tracking news and find topic interesting news for users. A keyword knowledge base is maintained and provides real-time updates to reflect the news topic information and the users interest preferences. The non-news content irrelevant to the news Web page is filtered out. A keyword extraction method based on lexical chains is proposed that uses the semantic similarity and the relatedness degree to represent the semantic relations between words. Word sense disambiguation is also performed in the built lexical chains. Experiments on Web news pages and journal articles show that the proposed keyword extraction method is effective. An example run of our PNFS system demonstrates the superiority of this Web intelligence system.

**Biography**: Xindong Wu is a Yangtze River Scholar in the School of Computer Science and Information Engineering at the Hefei University of Technology (China), a Professor of Computer Science at the University of Vermont (USA), and a Fellow of the IEEE and AAAS. He received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and Web information exploration. Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), and an Editor-in-Chief of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE, by the IEEE Computer Society) between 2005 and 2008. He served as Program Committee Chair/Co-Chair for ICDM '03 (the 2003 IEEE International Conference on Data Mining), KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management), and IEEE/ACM ASONAM '14 (the 2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining). He received the 2012 IEEE Computer Society Technical Achievement Award "for pioneering contributions to data mining and applications", and the 2014 IEEE ICDM 10-Year Highest-Impact Paper Award.

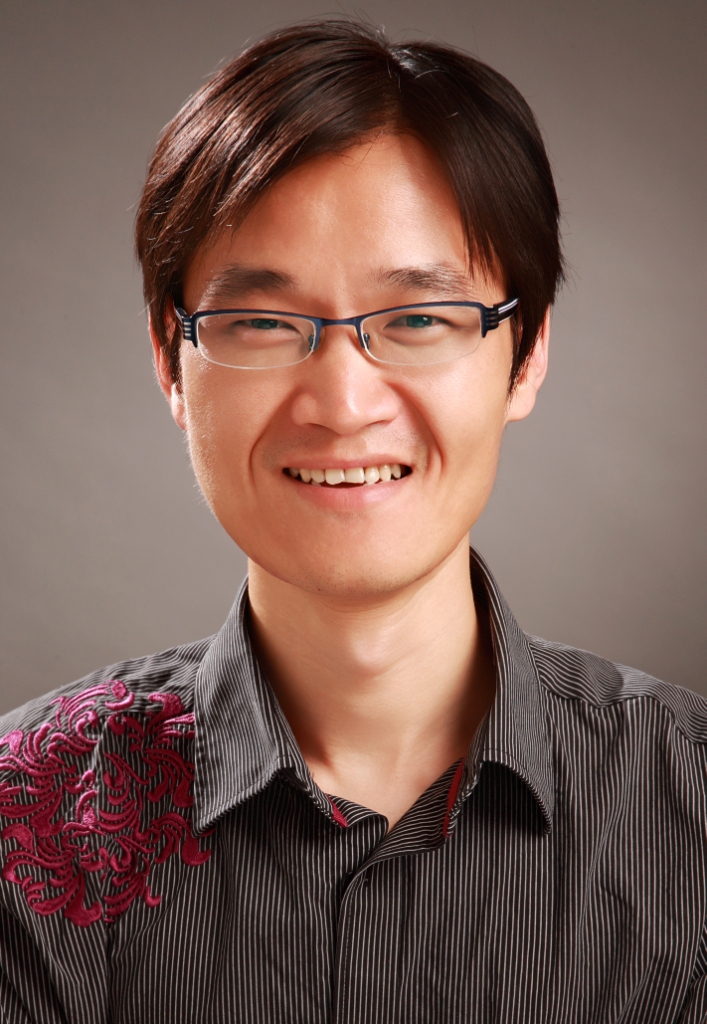
Keynote Session 3:

**Title**: Querying Big Data: From Theory and Practice

**Speaker**: *Prof. Wenfei Fan* (University of Edinburgh, UK; Beihang University, China)

**Abstract**: Querying big data is a departure from our familiar database techniques and even from the classical computational complexity theory. Several questions arise. What query classes are tractable in the context of big data? When is it feasible to query big data by accessing small data? Is there any practical technique for querying big data? When exact answers are beyond reach in practice, is approximate query answering possible within our constrained resources and with accuracy guarantees? This talk aims to provide an overview of recent advances in tackling these questions.

**Biography**: Professor Wenfei Fan is the Chair of Web Data Management in the School of Informatics, University of Edinburgh, UK, and the director of the International Research Center on Big Data, Beihang University, China. Prior to his move to the UK, he worked for Bell Labs, Lucent Technologies in the USA. He received his PhD from the University of Pennsylvania, USA, and his MS and BS from Peking University, China. Professor Fan is a Fellow of the Royal Society of Edinburgh, UK, a Fellow of the ACM, USA, a National Professor of the 1000-Talent Program and a Yangtze River Scholar, China. He is a recipient of the Alberto O. Mendelzon Test-of-Time Award of ACM PODS 2015 and 2010, ERC Advanced Fellowship in 2015, the Roger Needham Award in 2008, the Outstanding Overseas Young Scholar Award in 2003, the NSF Career Award in 2001, and several Best Paper Awards (VLDB 2010, ICDE 2007, and Computer Networks 2002). His current research interests include database theory and systems, in particular big data, data quality, data integration, distributed query processing, query languages, recommender systems, social networks and Web services.

Tutorial Session 1:

**Title**: Indexing for Querying Metric Spaces

**Speaker**: *Dr. Yunjun Gao* (Zhejiang University, China)

**Abstract**: Spatial queries, including similarity search, similarity joins, and aggregate k nearest neighbors queries, are useful in many areas, such as multimedia retrieval, data integration, and computational biology, to name but a few. However, they are not yet supported well by commercial DBMS. This may be due to the complex data types involved and the needs for flexible similarity criteria seen in real applications. In this talk, we propose an efficient disk-based metric access method, the Space-filling curve and Pivot-based B+-tree (SPB-tree), to accelerate query processing and support a wide range of data types and similarity metrics. The SPB-tree uses a small set of so-called pivots to reduce significantly the number of distance computations, utilizes a space-filling curve to cluster the data into compact regions, thus improving storage efficiency, and employs a B+-tree with minimum bounding box information as the underlying index. The SPB-tree also utilizes a separate random access file to efficiently manage a large and complex data. By design, it is easy to integrate the SPB-tree into an existing DBMS. In this talk, we also present efficient similarity search, similarity join, and aggregate k nearest neighbor query algorithms and corresponding cost models based on the SPB-tree. In addition, we develop a distributed geo-textual image retrieval and recommendation system (I2RS), which employs SPB-trees to index geo-textual images, and utilizes metric similarity queries, including time-aware range and k nearest neighbor queries, to provide a variety of geo-textual image retrieval and recommendation services.

**Biography**: Yunjun Gao received the PhD degree in computer science from Zhejiang University, China, in 2008. He is currently an associate professor in the College of Computer Science, Zhejiang University, China. Prior to joining the faculty, he was a Postdoctoral research Fellow at the Singapore Management University during 2008-2010, and a Visiting Scholar or Research Assistant at the Nanyang Technological University, Simon Fraser University, and City University of Hong Kong, respectively. His research interests include spatial and spatio-temporal databases, metric and missing/uncertain data management, and spatio-textual data processing. He has published papers in Journals and conferences including TODS, VLDBJ, TKDE, SIGMOD, VLDB, ICDE, and SIGIR. He is a member of the ACM and the IEEE, and a senior member of the CCF.

Tutorial Session 2:

**Title**: Influence computation in spatial databases

**Speaker**: *Dr. Muhammad Aamir Cheema* (Monash University, Australia)

**Abstract**: Consider a set of facilities (e.g., restaurants, fuel stations) and a set of users (e.g., diners, drivers). A user may issue a spatial query to find the facilities matching her preferences. In this tutorial, we focus on three most important queries issued by the users: 1) k nearest neighbors queries; 2) top-k queries; and 3) skyline queries. We say that a user u is interested in a facility q if q is one of the answers for the spatial query issued by u. Then, the influence set of a query facility is defined to be the set of all the users that are interested in q. Influence computation has attracted significant attention in the past few years due to its importance in many fields such as marketing, decision support systems and location based services. For example, a restaurant owner may want to find the users that are interested in her restaurant to send special deals. In this tutorial, we first highlight the importance of influence computation in spatial databases using various examples. Then, we introduce three most notable influence related queries: reverse k nearest neighbor (RkNN) queries, reverse top-k queries and reverse skyline queries. Finally, we present key ideas and insights in the most significant techniques to answer these queries.

**Biography**: Muhammad Aamir Cheema is a Lecturer at Clayton School of Information Technology, Monash University, Australia. He obtained his PhD from UNSW Australia in 2011. He has significant research experience in the areas of location-based services, spatio-temporal databases, mobile and pervasive computing and probabilistic databases and publishes frequently in top-tier conferences and journals such as PVLDB, SIGMOD, ICDE, VLDBJ, TODS, TKDE etc. He has won several prestigious awards and honours including 2012 Malcolm Chaikin Prize for Research Excellence in Engineering, 2014 Deans Award for Excellence in Research by an ECR, two CiSRA best-paper-of-the-year awards (2009 and 2010), two one-of-the-best-papers awards (ICDE 2010 and ICDE 2012), two best paper awards (WISE 2013 and ADC 2010), and a poster-of-the-day award (ICDE 2011). He was the publication chair for DASFAA 2015, PC co-chair for ADC 2015, local co-chair for APWeb 2013 and workshop co-chair for MSTD 2013.

Tutorial Session 3:

**Title**: Providing Retrievability Guarantees in Cloud-Based Storage Services

**Speaker**: *Dr. Yin Yang* (Hamad Bin Khalifa University, Qatar)

**Abstract**: Nowadays users store increasingly large amounts of data in cloud-based storage services such as Google Drive, Apple iCloud Drive and Microsoft OneDrive. For important data, the user often wants to verify that her data stored over the cloud have not been corrupted or compromised, which we call data retrivability. Simple approaches to this problem tend to be inefficient (e.g., the user could download the entire dataset and verify it locally, which incurs high communication overhead) or ineffective (e.g., the user could download and verify random blocks of the data, which tends to miss small-scale data corruptions). Meanwhile, traditional data authentication methods often burden the cloud service with heavy computations, leading to increased costs for the service. This tutorial introduces to the audience several state-of-the-art solutions for verifying data retrievability. We will overview their main ideas, and compare their efficiency, effectiveness and security guarantees. Additionally, we will discuss several interesting issues and application scenarios that have not been sufficiently addressed by existing methods.

**Biography**: Yin Yang is an assistant professor at the College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Qatar. Before joining HBKU, he worked as a research scientist at the Advanced Digital Sciences Center, a research lab of University of Illinois at Urbana-Champaign, located in Singapore. He has published extensively in data security and privacy, especially in query authentication and differential privacy. His recent research interests include security and privacy issues in cloud computing.

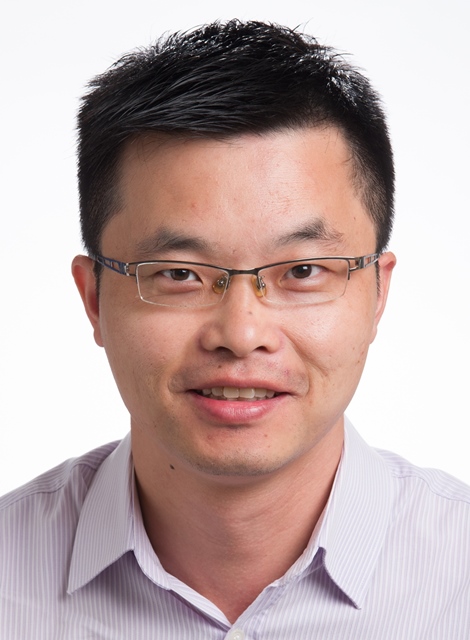
Tutorial Session 4:

**Title**: On Repairing Structural Problems In Semi-structured Data

**Speaker**: *Dr. Shanshan Ying* (Advanced Digital Sciences Center, Singapore)

**Abstract**: Semi-structured data such as XML are popular for data interchange and storage. However, many XML documents have improper nesting where open- and close-tags are unmatched. Since some semi-structured data (e.g., Latex) have a flexible grammar and since many XML documents lack an accompanying DTD or XSD, we focus on computing a syntactic repair via the edit distance. To solve this problem, we propose a dynamic programming algorithm which takes cubic time. While this algorithm is not scalable, well-formed substrings of the data can be pruned to enable faster computation. Unfortunately, there are still cases where the dynamic program could be very expensive; hence, we give branch-and-bound algorithms based on various combinations of two heuristics, called MinCost and MaxBenefit, that trade off between accuracy and efficiency. Finally, we experimentally demonstrate the performance of these algorithms on real data.

**Biography**: Shanshan Ying is a Postdoctoral Fellow at Advanced Digital Sciences Center, University of Illinois at Urbana Champaign. Shanshan received her Ph.D. in computer science under the supervison of Prof. Anthony Tung from the School of Computing, National University of Singapore, in 2014. Before that, she graduated with a B.S. degree from the Software Engineering Instituite, East China Normal University in 2008. Her research interests cover social network, data cleaning, in particular on semi-structured data.

Distinguished Young Lecturer Session:

**Title**: Authenticated Query Services in Outsourced Cloud Environments

**Speaker**: *Dr. Jianliang Xu* (Hong Kong Baptist University, Hong Kong)

**Abstract**: With the advent of Big Data, there has been a rising trend of outsourcing data management to cloud service providers, which provide query services to clients on behalf of data owners. While the cloud model is appealing in terms of scalability, elasticity and self-manageability, it brings about new challenges to query processing. In particular, the service provider can be untrustworthy or compromised, thereby returning incorrect or incomplete query results to clients, intentionally or not. Therefore, empowering clients to authenticate query results is imperative for outsourced databases. In this talk, we will start by a review of representative query authentication techniques. Then, we will present some of our recent efforts on authenticated query processing with privacy preservation requirements and in multi-source settings. We will also discuss some possible future research directions.

**Biography**: Jianliang Xu is a Professor in the Department of Computer Science, Hong Kong Baptist University. He received the BEng degree in computer science and engineering from Zhejiang University and the PhD degree in computer science from Hong Kong University of Science and Technology. He held visiting positions at Pennsylvania State University, State College, PA, USA and Fudan University, Shanghai, China. His research interests include data management, database security & privacy, mobile computing, and networked systems. He has published more than 150 technical papers in these areas, many of which appeared in leading journals and conferences including SIGMOD, PVLDB, ICDE, TODS, TKDE, and VLDBJ, with an h-index of 36 (Google Scholar). He was a recipient of IEEE ICDE Outstanding Reviewer Award (2010) and HKBU Faculty Performance Award for Outstanding Young Researcher (2012). He has served as a vice chairman of ACM Hong Kong Chapter and is a senior member of IEEE. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering (TKDE).



**Title**: Research Trends in Location-Based Services

**Speaker**: *Dr. Muhammad Aamir Cheema* (Monash University, Australia)

**Abstract**: Due to the immense increase in the usage of smartphones and other GPS enabled devices, cheap wireless network bandwidth and the availability of huge amount of data related to the locations, location-based services have become ubiquitous. Skyhook reported that the number of location-based applications being developed each month is increasing exponentially. Some applications of the location-based services include navigation, mobile yellow pages, location-based advertising, emergency services, geosocial networking, location-based games, travel planning and location-based billing. This talk will present a summary of the research in location-based services and is focused on a broader audience not necessarily familiar with the research in spatial databases. First, some of the most fundamental location-based queries will be introduced. Then, the emerging trends and research directions in the location-based services will be discussed.

**Biography**: Muhammad Aamir Cheema is a Lecturer at Clayton School of Information Technology, Monash University, Australia. He obtained his PhD from UNSW Australia in 2011. He has significant research experience in the areas of location-based services, spatio-temporal databases, mobile and pervasive computing and probabilistic databases and publishes frequently in top-tier conferences and journals such as PVLDB, SIGMOD, ICDE, VLDBJ, TODS, TKDE etc. He has won several prestigious awards and honours including 2012 Malcolm Chaikin Prize for Research Excellence in Engineering, 2014 Deans Award for Excellence in Research by an ECR, two CiSRA best-paper-of-the-year awards (2009 and 2010), two one-of-the-best-papers awards (ICDE 2010 and ICDE 2012), two best paper awards (WISE 2013 and ADC 2010), and a poster-of-the-day award (ICDE 2011). He was the publication chair for DASFAA 2015, PC co-chair for ADC 2015, local co-chair for APWeb 2013 and workshop co-chair for MSTD 2013.

Panel Session:

**Title**: Big Data in China: Where Are We Moving Toward?

**Speakers**: Elisa Bertino (Purdue University, USA), Kang Chen (China Telecom Ltd. Co., China), Wenfei Fan (University of Edinburgh, UK; Beihang University, China), Zhifeng Hao (Guangdong University of Technology, China), Xuemin Lin (New South Wales University, Australia), Jeffrey Xu Yu (The Chinese University of Hong Kong, Hong Kong); Xiaofang Zhou (University of Queensland, Australia)

Research Session 1: Social Prediction

**Hashtag Sense Induction Based on Co-Occurrence Graphs***, Mengmeng Wang* (Waseda University, Japan), *Mizuho Iwaihara* (Waseda University, Japan)

Abstract: Twitter hashtags are used to categorize tweets for improving search categorizing topic. But the fact that people can create and use hashtags freely leads to a situation such that one hashtag may have multiple senses. In this paper, we propose a method to induce senses of a hashtag in a particular time frame. Our assumption is that for a sense of a hashtag the context words around it are similar. Then we design a method that uses a co-occurrence graph and community detection algorithm. Both words and hashtags are nodes of the co-occurrence graph, and an edge represents the relation of two nodes co-occurring in the same tweet. A list of words with a high node degree representing a sense is extracted as a community of the graph. Different from natural language words, not all of what we obtain is senses; some of them are just highly related topics, due to the arbitrary use of hashtags. We take Wikipedia disambiguation list page as word sense inventory to refine the results by removing non-sense topics. The result is evaluated by the way of information retrieval.

**Early Outbreak Detection in Social Networks**, *Chuan Zhou* (Institute of Information Engineering, Chinese Academy of Science, China)

Abstract: In this paper, we investigate the problem of placing sensors in a social network to get quickly informed about contagious outbreaks, i.e., placing k sensors in a network in order to minimize the time until a contaminant - starting from a random node in the network - is detected. We aim to optimize the Sensor Placement from two complementary directions. One is to improve the original greedy algorithm and its extensions to reduce sensor selection time, and the other is to propose a new Quickest Path heuristic that can shorten the detection time. We test and compare our algorithms with previous algorithms on four real data sets. Experimental results show that 1) the new greedy algorithm is more efficient than existing greedy algorithms in terms of selection time, 2) the quickest path heuristic obtains less detection time than centrality-based heuristics, and is as effective as the greedy algorithms, and 3) the new heuristic has the potential to scale well to large networks, having low detection time and selection time.

**A Multi-view Retweeting Behaviors Prediction in Social Networks**, *Bo Jiang* (Institute of Information Engineering, Chinese Academy of Science, China)*, Ying Sha* (Institute of Information Engineering, Chinese Academy of Science, China)*, Lihong Wang* (Institute of Information Engineering, Chinese Academy of Science, China)

Abstract: Retweeting is the most prominent feature in online social networks. It allows users to reshare another user's tweets for her followers and bring about second information diffusion. Predicting retweeting behaviors is an important and essential task for advertising product launch, hot event detection and analysis of human behavior. However, most of the methods and systems have been developed for modeling the retweeting behaviors, it has not been fully explored for this problem. In this paper, we first cast the problem of retweeting behavior prediction as a classification task and propose a formally definition. We then systematically summarize and extract a lot of features, namely user status, content, temporal, and social tie information, for predicting users' retweeting behaviors. We incorporate these features into Support Vector Machine (SVM) model for our prediction problem. Finally, we conduct extensive experiments on a real world dataset collected from Twitter to validate our proposed approach. Our experimental results demonstrate that our proposed model can improve prediction effectiveness by combining the extracted features compared to the baselines that do not.

Research Session 2: Text Analysis I

**A Multi-news Timeline Summarization Algorithm Based on Aging Theory**, *Jie Chen*(Beijing Institute of Technolog, China), *Zhendong Niu* (Beijing Institute of Technolog, China; University of Pittsburgh, USA)*, Hongping Fu* (Beijing Institute of Technolog, China)

Abstract: This paper focuses on the problem of news event timeline summary in Multi-Document Summarization, which aims to summarize multi-news regarding the same event in timeline. Most traditional methods to solve this problem consider the text surface features and topic- related features, such as length of each sentence, position of the sentence in the document, number of topic words, etc. Traditional methods ignored that every event has its life circle including birth, growth, maturity and death. In this paper, we present a novel approach for summarizing multi-news regarding the same topic. The proposed approach consider both the traditional features and the life circle feature of each event. Our approach consists of four steps. First, sentences and their publish date are extracted from each news article. Second, several pretreatment work are done to these sentences to reduce the influence of noises like synonyms. Third, life circle features and other four categories of features which are common used in this field are collected. Finally, SVM model is used to train these features to recognize the summary sentence of the news document. We have tested our approach on the public datasets, DUC-2002 and TAC-2010, and the results show that our approach is more effective in summarizing multi-news in timeline than existing methods.

**Extended Strategies for Document Clustering with Word Co-occurrences**, *Yang Wei*(Nankai University, China)*, Jinmao Wei*(Nankai University, China)*, Zhenglu Yang*(Nankai University, China)

Abstract: To tackle the sparse data problem of the bag-of-words model for document clustering, recent strategies have been proposed to enrich a document with the relatedness of all the words in a corpus to the document, where the relatedness is estimated by the weighted sum of word co-occurrences. However, the relatedness is overestimated without eliminating the overlaps between word co-occurrences. This paper demonstrates that the weighted sum strategy gives the upper bound of the theoretic degree of relatedness. Two strategies are further proposed to approach the theoretic degree of relatedness. The first strategy is established under the extreme assumption that all the words in a document co-occur with each other. By considering the specificities of words, the second strategy gives several extended versions of the weighted sum strategy. Substantial experiments verify that the document clustering incorporated with the extended strategies achieve a significant performance improvement compared to the state-of-the-art techniques.

**User Generated Content Oriented Chinese Taxonomy Construction**, *Jinyang Li*(East China Normal University, China)*, Chengyu Wang*(East China Normal University, China)*, Xiaofeng He*(East China Normal University, China)*, Rong Zhang*(East China Normal University, China)*, Ming Gao*(East China Normal University, China)

Abstract: The taxonomy is one of the basic components in knowledge graphs as it establishes types of classes and semantic relations among the classes. Taxonomies are normally constructed either manually, or by language-dependent rules or patterns for type and relation extraction or inference. Existing work on building taxonomies for knowledge graphs is mostly in English language environment. In this paper, we propose a novel approach for large-scale Chinese taxonomy construction based on user generated content. We take Chinese Wikipedia as the data source, develop methods to extract classes and their relations mined from user tagged categories, and build up the taxonomy using a bottom-up strategy. The algorithms can be easily applied to other Wiki-style data sources. The experiments show that the constructed Chinese taxonomy achieves better results in both quality and quantity.

Research Session 3: Learning from Data

**Discovering Restricted Regular Expressions with Interleaving**, *Feifei Peng*(Institute of Software, Chinese Academy of Sciences, China)*, Haiming Chen*(Institute of Software, Chinese Academy of Sciences, China)

Abstract: Discovering a concise schema from given XML documents is an important problem in XML applications. In this paper, we focus on the problem of learning an unordered schema from a given set of XML examples, which is actually a problem of learning a restricted regular expression with interleaving using positive example strings. Schemas with interleaving could present meaningful knowledge that cannot be disclosed by previous inference techniques. Moreover, inference of the minimal schema with interleaving is challenging. The problem of finding a minimal schema with interleaving is shown to be NP-hard. Therefore, we develop an approximation algorithm and a heuristic solution to tackle the problem using techniques different from known inference algorithms. We do experiments on real-world data sets to demonstrate the effectiveness of our approaches. Our heuristic algorithm is shown to produce results that are very close to optimal.

**Spare Part Demand Prediction based on Context-aware Matrix Factorization**, *Jianwei Ding*(Tsinghua University, China)*, Yingbo Liu*(Tsinghua University, China)*, Yuan Cao*(Tsinghua University, China)*, Li Zhang*(Tsinghua University, China)*, Jianmin Wang*(Tsinghua University, China)

Abstract: Maintenance spare part is used to replace and update the damaged and old components in the equipment. Forecasting spare part demand is notoriously difficult, as demand is typically intermittent and lumpy. Meanwhile, with the development of the sensor and internet technology, numerous condition monitoring systems are used to monitor the working condition of equipment, generating a large variety of monitor data at runtime. In this paper, we propose a Spare Part Demand (SPD) model based on a context-aware matrix factorization approach. The SPD mode incorporates historical spare part demands, the correlation between spare part demands and working places, and the correlation between spare part demands and monitor data. We evaluate our method based on extensive experiments using historical spare demands of one important component from more than 10000 concrete pump trucks and monitor data generated by part of these pump concrete pump trucks over a period of 9 months. The results demonstrate the advantages of our method over the previous studies, validating the contribution of our method.

**Online Feature Selection based on Passive-Aggressive Algorithm with Retaining Features**, *Hai-Tao Zheng*(Tsinghua University, China)*, Haiyang Zhang*(Tsinghua University, China)

Abstract: Feature selection is an important topic in data mining and machine learning, and has been extensively studied in many literature. Unlike traditional batch learning methods, online learning is more efficient for real-world applications. Most exist-ing studies of online learning require accessing all the features of training instanc-es, but in real world, it is often expensive to acquire the full set of attributes. In online feature selection process, when a training instance arrive, a fixed small number of features will be selected, and then the other features will be ignored. However, those ignored features may be useful and selected in later instances. If we only consider the new instances for these special features, it will lead to ex-treme errors. To address these issues, we improved a novel algorithm with Pas-sive-Aggressive Algorithm and retaining features. Then we evaluate the perfor-mance of the proposed algorithms for online feature selection on several public datasets, and we can see from the experiments that our algorithm consistently surpassed the baseline algorithms for all the situations.

**Batch Mode Active Learning For Geographical Image Classification**, *Bo Du*(Wuhan Univresity, China)*, Zengmao Wang*(Wuhan Univresity, China)*, Wenbin Hu*(Wuhan Univresity, China)*, Lefei Zhang*(Wuhan Univresity, China)*, Liangpei Zhang*(Wuhan Univresity, China), *Dacheng Tao*(University of Technology, Sydney, Australia)

Abstract: In this paper, an innovative batch mode active learning by combining discriminative and representative information for hyperspectral image classification with support vector machine is proposed. In the past years, the batch mode active learning mainly exploits different query functions, which are based on two criteria: uncertainty criterion and diversity criterion. Generally, the uncertainty criterion and diversity criterion are independent of each other, and they also could not make sure the queried samples identical and independent distribution. In the proposed method, the diversity criterion is focused. In the innovative diversity criterion, firstly, we derive a novel form of upper bound for true risk in the active learning setting, by minimizing this upper bound to measure the discriminative information, which is connected with the uncertainty. Secondly, for the representative information, the maximum mean discrepancy(MMD) which captures the representative information of the data structure is adopt to match the distribution of the labeled samples and query samples, to make sure the queried samples have a similar distribution to the labeled samples and guarantee the queried samples are diversified. Meanwhile, the number of new queried samples is adaptive, which depends on the distribution of the labeled samples. In the experiment, we employ two benchmark remote sensing images, Indian Pines and Washington DC. The experimental results demonstrate the effective of our proposed method compared with the state-of-the-art AL methods.

Research Session 4: Hardware, Schema and Query

**Efficient Buffer Management for PCM-Enhanced Hybrid Memory Architecture**, *Kaimeng Chen*(University of Science and Technology of China, China)*, Peiquan Jin*(University of Science and Technology of China, China; Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, China)*, Lihua Yue*(University of Science and Technology of China, China; Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, China)

Abstract: Recently, phase change memory (PCM) has been considered in memory architecture to serve as an extension to DRAM, due to its special properties of byte-addressibility, low energy consumption, and high read performance. However, PCM has a lower write speed than DRAM. Besides, it has a limited write endurance. Therefore, the co-existence of PCM and DRAM in main memory urges a careful buffer-management policy to avoid frequent writes to PCM. To address this problem, we present the first approach that reduces PCM writes by efficient page exchanges and page replacements. Specially, we propose two clock data structures to maintain DRAM and PCM pages, and devise a page exchange method to make recently-updated pages reside in DRAM. In addition, differing from previous studies that do not consider the influence of page replacements on PCM writes, we present a new page replacement algorithm to reduce page replacement on PCM. With this mechanism, we can reduce PCM writes efficiently while keeping a high hit ratio. We conduct trace-driven experiments on both synthetic and real traces. The experimental results suggest that our proposal can greatly reduce PCM writes and maintain a high hit ratio for PCM/DRAM-based hybrid memory architecture.

**DistDL: A Distributed Deep Learning Service Schema with GPU Accelerating**, *Jianzong Wang*(Guangdong University of Technology, China)*, Lianglun Cheng*(Guangdong University of Technology, China)

Abstract: Deep Learning is a term coined by the industry and academia which combines the broad field of machine learning with the application of deep neural networks in the big dat era. Recently, the ability to train neural networks has resulted in state-of-the-art performance in many areas of computer vision, speech recognition, recommended system, and behavioural analysis etc. Existing deep learning systems are inadequate for scaling, especially in current cloud distributed infrastructures where the nodes are distributed across multiple geographical location. In this paper, we have presented DistDL, a novel distributed deep learning service schema in order to reduce the training time consumption and achieve extremely parallelism with data and model. We also took into consideration GPU inside DistDL by leverage the remarkable capacity of heterogeneous. The results of our experiments in the benchmarking suggest that DistDL is adaptive to various scaling patterns with the same accuracy while minimising training time by leveraging the GPU computation capability, up to 80% decrease.

**Overlapping Schema Summarization based on Multi-label Propagation**, *Man Yu*(NanKai University, China)*, Chao Wang*(NanKai University, China)*, Xiangrui Cai*(NanKai University, China)*, Yanlong Wen*(NanKai University, China)*, Ying Zhang*(NanKai University, China)*, Xiaojie Yuan*(NanKai University, China)

Abstract: Modern databases are usually composed of hundreds of tables. Querying an unfamiliar database is a tall order for users before they truly understand its schema. A schema summary can help to provide a succinct overview of the schema and improve the usability of databases. Existing summarization methods only focus on each element in a database belongs to one topic, ignores the fact that some elements may belong to multiple topics. This paper come up with a new method of generating overlapping summaries. It is the very first work to address the task as far as we know. We formulate overlapping schema summarization first and then introduce multi-label propagation algorithm in community detection to achieve several groups. To refine the partition, we cluster the groups additionally using hierarchical clustering algorithm. Finally, we find representative tables in each cluster to annotate the schema summary. The extensive experiments on both benchmark database and real-world database show that our approach not only achieves higher accuracy but also generates more meaningful summary.

**RDQS: A Relevant and Diverse Query Suggestion Generation Framework**, *Yichi Zhang*(Tsinghua University, China)*, Hai-Tao Zheng*(Tsinghua University, China)

Abstract: Traditional query suggestion methods mainly leverage click-through information to find related queries as recommendations, without considering the semantic relatedness between queries. In addition, few studies use click-through distribution in diversifying query suggestions. To address these issues, we propose a novel and effective framework to generate relevant and diversified query suggestions. We combine query semantics and click-through information together to generate query suggestion candidates which are highly relevant to original query, we use click-through distribution to diversify the candidates. We evaluate our method on a large-scale search log dataset of a commercial engine, experimental results indicate that our framework has significantly improved the relevance and diversity of suggested queries by comparing to four baseline methods.

Research Session 5: Topic Model

**An Online Inference Algorithm for Labeled Latent Dirichlet Allocation**, *Qiang Zhou* (Beijing Institute of Technology, China)*, Heyan Huang* (Beijing Institute of Technology, China)*, Xianling Mao* (Beijing Institute of Technology, China)

Abstract: Using topic models to analyze documents is a popular method in text mining. Labeled Latent Dirichlet Allocation(Labeled LDA) is one of them that is widely used to model tagged documents and to solve relevant problems such as tagged document visualization, snippet extraction and so on. However, traditional batch inference for Labeled LDA, which runs over entire document collection, is computationally expensive and not suitable for large scale corpora and text streams. In this paper, we develop an efficient online algorithm for Labeled LDA, called online Labeled LDA(online-LLDA). It is based on particle filter, a Sequential Monte Carlo approximation technique. Our experiments show that online-LLDA significantly outperforms batch algorithm (batch-LLDA) in time, while preserving equivalent quality.

**Multi-Label Emotion Tagging for Online News by Supervised Topic Model**, *Ying Zhang*(Nankai University, China)*, Lili Su*(Nankai University, China)*, Zhifan Yang*(Nankai University, China)*, Xue Zhao*(Nankai University, China)*, Xiaojie Yuan*(Nankai University, China)

Abstract: An enormous online news services provide users with interactive platforms where users can freely share their subjective emotions, such as sadness, surprise, and anger, towards the news articles. Such emotions can not only help understand the preferences and perspectives of individual users, but also benefit a number of online applications to provide users with more relevant services. While most of previous approaches are intended for recognizing a single emotion of the author, it has been observed that different emotions of the readers are more representative of the news articles. Therefore, this paper focuses on predicting readers' multiple emotions evoked by online news. To the best of our knowledge, this is the first research work for addressing the task. This paper proposes a novel supervised topic model which introduces an additional emotion layer to associate latent topics with evoked multiple emotions of readers. In particular, the model generates a set of latent topics from emotions, followed by generating words from each topic. The experiments on the real dataset from online news service demonstrate the effectiveness of the proposed approach in multi-label emotion tagging for online news.

**Distinguishing Specific and Daily topics**, *Tao Ge*(Peking University, China)*, Wenzhe Pei*(Peking University, China)*, Baobao Chang*(Peking University, China)*, Zhifang Sui*(Peking University, China)

Abstract: The task of distinguishing specific and daily topics is useful in many applications such as event chronicle and timeline generation, and cross-document event coreference resolution. In this paper, we investigate several numeric features that describe useful statistical information for this task, and propose a novel Bayesian model for distinguishing specific and daily topics from a collection of documents based on documents' content. The proposed Bayesian model exploits mixture of Poisson distributions for modeling probability distributions of the numeric features. The experimental results show that our approach is promising to solve this problem.

**A Supervised Parameter Estimation Method of LDA**, *Zhenyan Liu*(Institute of Computing Technology, Chinese Academy of Sciences,China; University of Chinese Academy of Sciences, China; Institute of Information Engineering, Chinese Academy of Sciences, China; Beijing Institute of Technology, China),*Dan Meng*(Institute of Information Engineering, Chinese Academy of Sciences, China), *Weiping Wang*(Institute of Information Engineering, Chinese Academy of Sciences, China), *Chunxia Zhang*(Beijing Institute of Technology, China)

Abstract: Latent Dirichlet Allocation (LDA) probabilistic topic model is a very effective dimension-reduction tool which can automatically extract latent topics and dedicate to text representation in a lower-dimensional semantic topic space. But the original LDA and its most variants are unsupervised without reference to category label of the documents in the training corpus. And most of them view the terms in vocabulary as equally important, but the weight of each term is different, especially for a skewed corpus in which there are many more samples of some categories than others. As a result, we propose a supervised parameter estimation method based on category and document information which can estimate the parameters of LDA according to term weight. The comparative experiments show that the proposed method is superior for the skewed text classification, which can largely improve the recall and precision of the minority category.

Research Session 6: Metric Learning

**Tree-Based Metric Learning for Distance Computation in Data Mining**, *Ming Yan* (Harbin Institute of Technology, China)*, Zhang Yan*(Harbin Institute of Technology, China)*, Hongzhi Wang*(Harbin Institute of Technology, China)

Abstract: Distance is an essential measurement of data mining. A good metric often leads to a good performance. Then how to obtain a proper metric systematically is critical. Distance metric learning is a classic method to learn distances between instances on data set with complex distributions. However, most researches on distance metric learning are based on Mahalanobis metric, which is equivalent to linear transformation on distance space that has limitation on complex data. To solve this problem, we propose a metric learning method based on non-linear transformation suitable for complex data. By using the tree model, we could address non-linearly separable data that rearrange input data and represent them to another forms, and tree model could be able to implicitly represent data to a new distance space with a non-linear activator function. Furthermore, single tree model will lead to overfit that has higher generalization errors. Therefore, we design a randomize algorithm to combining different tree models which could reduce the generalization errors in theory and practice. According to analysis, we prove the correctness and effectiveness of our algorithm in theory. Extensive experiments demonstrate that algorithm is stable and suitable for data mining.

**Learning Similarity Functions for Urban Events Detection by Mining Hotline Phone Records**, *Pengjie Ren*(Shandong University, China)*, Peng Liu*(Shandong University, China)*, Zhumin Chen*(Shandong University, China)*, Jun Ma*(Shandong University, China)*, Xiaomeng Song*(Shandong University, China)

Abstract: Many cities around the world have established a platform, entitled public service hotline, to allow citizens to tell about city issues, e.g. noise nuisance, or personal encountered problems, e.g. traffic accident, by making a phone call. As a result of “crowd sensing”, these records contain rich human intelligence that can help to detect urban events. In this paper, we present an event detection approach to detect urban events based on phone records. Specifically, given a set of phone records in a period of time, we first learn a similarity matrix. Each element of the matrix is estimated as the probability that the corresponding pair of records describe the same event. Then, we propose an Improved Affinity Propagation (IAP) clustering approach which takes the similarity matrix as input and generates clusters as output. Each cluster is an urban event composed of several records. Extensive experiments demonstrate the great improvement of IAP on three standard datasets for clustering and the effectiveness of our event detection approach on real data from a hotline.

**A New Similarity Measure Between Semantic Trajectories based on Road Networks**, *Wu Xia*(Wuhan University, China)*, Yuanyuan Zhu*(Wuhan University, China)*, Shengchao Xiong*(Wuhan University, China)*, Yuwei Peng*(Wuhan University, China)*, Zhiyong Peng*(Wuhan University, China)

Abstract: With the development of the positioning technology, studies on trajectories have been growing rapidly in the past decades. As a fundamental part involved in trajectory recommendation and prediction, trajectory similarity has attracted considerable attention from researchers. However, most existing works focus on raw trajectory similarity by comparing their shapes, very few works study semantic trajectory similarity and none of them take all the geographical, semantic, and timestamp information into consideration. In this paper, we model semantic trajectories based on road networks considering all these information, and propose a Constrained Time-based Common Parts (CTCP) approach to measure the similarity. Since the strict time constraint in CTCP may lead to many zero values, we further propose an improved Weighted Constrained Time-based Common Parts (WCTCP) method by relaxing the time constraint to measure the similarity more accurately. We conducted extensive performance studies on real datasets to confirm the effectiveness of our approaches.

**Multiple Attribute Aware Personalized Ranking**, *Weiyu Guo*(University of Chinese Academy, China; Institute of Automation, Chinese Academy of Sciences, China)*, Shu Wu* (Institute of Automation, Chinese Academy of Sciences, China)*, Liang Wang*(Institute of Automation, Chinese Academy of Sciences, China)*, Tieniu Tan*(Institute of Automation, Chinese Academy of Sciences, China)

Abstract: Personalized ranking is a typical task of recommender systems. It can provide a set of items for specific user and help recommender systems more correctly direct each item to its user. Recently, as the dramatically increasing social media, an entity, i.e., user and item, usually associates with multiple kinds of characterized information, e.g., explicit ratings, implicit feedbacks, and multi-type attributes (such as age, sex, occupation, or posts of user). Intuitively, comprehensively considering these information, we can obtain better personalized ranking results. However, most conventional methods only take collaborative information (explicit ratings or implicit feedbacks) or single type attributes into account. In this work, we investigate how to combine multiple attribute and collaborative information to learn the latent factors for entities and the attribute-aware mappings. As a result, we propose a novel Multiple-attribute aware Bayesian Personalized Ranking model, Maa-BPR, for personalized ranking, which can learn reliable latent factors for entities as well as effective mappings for multiple attribute. The experimental results show that, compared with the state-of -the-art methods, Maa-BPR not only provides better ranking performance, but also is robust to new entities and the incomplete attributes.

Research Session 7: Semantics

**An Ensemble Matchers based Rank Aggregation Method for Taxonomy Matching**, *Hailun Lin*(Institute of Computing Technology, Chinese Academy of Sciences, China)*, Yuanzhuo Wang*(Institute of Computing Technology, Chinese Academy of Sciences, China)*, Yantao Jia*(Institute of Computing Technology, Chinese Academy of Sciences, China)*, Jinhua Xiong*(Institute of Computing Technology, Chinese Academy of Sciences, China)*, Peng Zhang*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Xueqi Cheng*(Institute of Computing Technology, Chinese Academy of Sciences, China)

Abstract: Taxonomy matching is an important operation of knowledge base merging. Several matchers for automating taxonomy matching have been proposed and evaluated in the knowledge base community. Studies reveal that there is no single taxonomy matcher suitable for any domain-specific taxonomy mapping, therefore an ensemble of taxonomy matchers is essential. In this paper, we propose taxonomy metamatching, a distributed computing framework for assembling taxonomy matchers and generating an optimal taxonomy mapping. And we introduce TRA, a Threshold Rank Aggregation algorithm for this problem. Experimental results show that TRA outperforms state-of-the-art approaches regardless of domains and scales of taxonomies, which demonstrates that TRA performs good adaptability to taxonomy matching.

**Random-based Algorithm for Efficient Entity Matching**, *Pingfu Chao*(East China Normal University, China)*, Zhu Gao*(East China Normal University, China)*, Yuming Li*(East China Normal University, China)*, Junhua Fang*(East China Normal University, China)*, Rong Zhang*(East China Normal University, China)*, Aoying Zhou*(East China Normal University, China)

Abstract: Most of the state-of-the-art MapReduce-based entity matching methods inherit traditional Entity Resolution techniques on centralized system and focus on data blocking strategies for structured entities in order to solve the load balancing problem occurred in distributed environment. In this paper, we propose a MapReduce-based entity matching framework for Entity Matching on semi-structured and unstructured data. Each entity is represented by a high dimensional vector generated from description data. In order to reduce network transmission, we produce lower dimensional bit-vectors called signatures for those entity vectors based on Locality Sensitive Hash (LSH) function. Our LSH is required for promising cosine similarity. A series of random algorithms are designed to ensure the performance for entity matching. Moreover, our design contains a solution for reducing redundant computation by one round of additional MapReduce job. Experiments show that our approach has a huge advantages on both processing speed and accuracy compared to the other methods.

**Research on Semantic Disambiguation in Treebank**, *Lin Miao*(Beijing Information Science and Technology University, China), *Xueqiang Lv*(Beijing Information Science and Technology University, China)*, Yunfang Wu*(Peking University, China)*, Yue Wang*(Beijing Information Science and Technology University, China)

Abstract: The increasingly widespread application of natural language pro-cessing technology leads parsing to play a significant role. As a result, the size and quality of treebank have become the focus of relevant research. However, there exists data sparseness when we use the treebank to parse. With the help of Cilin semantic information and words contextual information, this paper proposes a context-based lexical semantics disambiguation method. After applying this method on CTB (Chinese Treebank) 5.0 and TCT (Tsinghua Chinese Treebank), using Berkeley Parser achieved relatively good results. In Penn Chinese Tree-bank，the precision and recall rates reached 85.35% and 84.34% respectively, and the F value reached 84.84%. Comparing with the parsing results of using the original corpus, the correct rate increased by 1.86% and the recall rate increased by 1.02 % and the comprehensive index F value increased by 1.35%. As conse-quence, the overall parsing error rate dropped by 8.17%.

**An Self-Learning Rule-based Approach for Sci-tech Compound Phrase Entity Recognition**, *Tingwen Liu*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Yang Zhang*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Yang Yan*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Jinqiao Shi*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Li Guo*(Institute of Information Engineering, Chinese Academy of Sciences, China)

Abstract: Sci-tech compound phrase entity (e.g., the names of projects, books and patents) recognition is a fundamental task of science and technology data processing and discovery. However, much less work have been done on the problem. In this paper, we first give the characteristics of sci-tech entities that are different from personal name and other traditional entities. Then we introduce a self-learning rule-based approach to address the problem. The approach consists of three stages, namely rule- based text truncation, BlackPOS-based text split and WhiteKey-based confirmation. Constructing the best WhiteKey list is a NP-hard problem. We further design dynamic programming and greedy algorithms to address the problem. Experimental results show that our approach achieves 94.78% precision rate, 89.19% recall rate and 91.9% F 1 measure in average. Moreover, our approach is universal and orthogonal to prior named entity recognition work.

Research Session 8: Query Processing

**Efficient Location-dependent Skyline Queries in Wireless Broadcast Environments**, *Yingyuan Xiao*(Tianjin University of Technology, China)*, Pengqiang Ai*(Tianjin University of Technology, China)*, Hongya Wang*(Donghua University, China)*, Ching-Hsien Hsu*(Chung Hua University, China)*, Wenxiang Cui*(Tianjin University of Technology, China)

Abstract: We study the problem of location-dependent skyline query (LDSQ) processing in wireless broadcast environments in this paper. Compared with answering the skyline queries in a conventional setting, two new issues arise while processing location-dependent skyline in wireless broadcast environments. First, the result of an LDSQ is closely related to the query point’s location; secondly, query processing strategies must take the linear property of wireless broadcast media and limited battery life of mobile devices into consideration. To address these new issues, this paper proposes an efficient solution for LDSQ processing in wireless broadcast environments. In particular, data objects to be disseminated are first divided into two parts via pre-computation in the broadcast server, and then a novel air data organization scheme is designed in the broadcast disk. In the mobile client end, an energy-efficient LDSQ processing algorithm is presented. To demonstrate the efficiency of our solution, extensive experiments are conducted along with detailed performance analysis.

**Efficient Algorithms for Distance-based Representative Skyline Computation in 2D Space**, *Taotao Cai*(Shenzhen University, China)*, Rong-Hua Li*(Shenzhen University, China)*, Jeffrey Xu Yu*(The Chinese University of Hong Kong, Hong Kong)*, Rui Mao*(Shenzhen University, China)*, Yadi Cai*(Shenzhen University, China)

Abstract: Representative skyline computation is a fundamental issue in database area, which has attracted much attention in recent years. A notable definition of representative skyline is the distance-based representative skyline (DBRS). Given an integer *k*, a DBRS includes *k* representative skyline points that aims at min- imizing the maximal distance between a non-representative skyline point and its nearest representative. In the 2D space, the state-of-the-art algorithm to compute the DBRS is based on dynamic programming (DP) which takes *O*(*km*2 ) time complexity, where *m* is the number of skyline points. Clearly, such a DP-based algorithm cannot be used for handling large scale dataset due to the quadratic time cost. To overcome this problem, in this paper, we propose a new approximate algorithm called ARS, and a new exact algorithm named PSRS, based on a carefully-designed parametric search technique. We show that the ARS algorithm can guarantee a solution that is at most *E* larger than the optimal solution. The proposed ARS and PSRS algorithms run in *O*(*k* log2 *m* log(*T /E*)) and *O*(*k*2 log3 *m*) time respectively, where *T* is no more than the maximal distance between any two skyline points. We conduct extensive experimental studies over both synthetic and real-world datasets, and the results demonstrate the efficiency and effectiveness of the proposed algorithms.

**A Compression-Based Filtering Mechanism in Content-Based Publish/Subscribe System**, *Qin Liu*(Tongji University, China)*, Yiwen Zheng*(Tongji University, China)*, Kaile Wang*(Tongji University, China)

Abstract: Recently publish/subscribe (pub/sub) has been a popular paradigm in Internet-scale applications to decouple information publishers and subscribers. The main task of pub/sub is to find those subscriptions that match a given publication. Typically a matching algorithm utilizes a subscription indexing structure for higher efficiency. Unfortunately, given the diverse interests of publishers and subscribers, the semantic space of pub/sub becomes high dimensional and very sparse, leading to nontrivial space cost and inefficient matching algorithm. Existing work paretically tackled the both issues but at cost of scarifying another one, and few can meet the both goals. To overcome this issue, in this paper, we propose a novel coding approach. The approach can not only reduce the space cost of subscription index, and meanwhile helps optimizing the matching efficiency. Our experiments successfully demonstrate the proposed algorithm can achieve not only the low space cost of our proposed subscription indexing but also low running time of matching algorithm.

**Hashing Multi-Instance Data from Bag and Instance Level**, *Yao Yang* (Shandong university, China)*, Xin-Shun Xu* (Shandong university, China)*, Xiaolin Wang* (Shandong university, China)*, Shanqing Guo* (Shandong university, China)*, Lizhen Cui* (Shandong university, China)

Abstract: In many scenarios, we need to do similarity search of multi-instance data. Although the traditional kernel methods can measure the similarity of bags in original feature space, the time and storage cost of these methods are so high which makes such methods cannot deal with large scale problems. Recently, hashing methods have been widely used for similarity search due to its fast search speed and low storage cost. However, few works consider how to hash multi-instance data. In this paper, we present two multi-instance hashing methods: (1) Bag-level Multi-Instance Hashing (BMIH); (2) Instance-level Multi-Instance Hashing (IMIH). BMIH first maps each bag to a new feature representation by a feature fusion method; then, supervised hashing method is used to convert new features to hash code. To utilize more instance information in each bag, IMIH regards instances in all bags as training data and apply two types of hash learning methods (unsupervised and supervised, respectively) to convert all instances to binary code; then, for a test bag, a similarity measure is proposed to search similar bags. Our experiments on four real-world datasets show that instance-level hashing with supervised information outperforms all proposed techniques.

Research Session 9: Recommendation System I

**MATAR: Keywords Enhanced Multi-Label Learning for Tag Recommendation**, *Licheng Li* (Nanjing University, China)*, Yuan Yao*(Nanjing University, China)*, Feng Xu*(Nanjing University, China)*, Jian Lu*(Nanjing University, China)

Abstract: Tagging is a popular way to categorize and search online content, and tag recommendation has been widely studied to better support automatic tagging. In this work, we focus on recommending tags for content-based applications such as blogs and question-answering sites. Our key observation is that many tags actually have appeared in the content in these applications. Based on this observation, we first model the tag recommendation problem as a multi-label learning problem and then further incorporate keyword extraction to improve recommendation accuracy. Moreover, we speedup the proposed method using a locality-sensitive hashing strategy. Experimental evaluations on two real data sets demonstrate the effectiveness and efficiency of our proposed methods.

**Simple is Beautiful: An Online Collaborative Filtering Recommendation Solution with Higher Accuracy**, *Feng Zhang*(China University of Geoscience, China)*, Ti Gong*(China University of Geoscience, China)*, Victor Lee*(John Carroll University, USA)*, Gansen Zhao*(South China Normal University. China)*, Guangzhi Qu*(Oakland University, USA)

Abstract: Matrix factorization has high computation complexity. It is unrealistic to directly adopt such techniques to online recommendation where users, items, and ratings grow constantly. Therefore, implementing an online version of recommendation based on incremental matrix factorization is a significant task. Though some results have been achieved in this realm, there is plenty of room to improve. This paper focuses on designing and implementing algorithms to perform online collaborative filtering recommendation based on incremental matrix factorization techniques. Specifically, for the new-user and new-item problems, Moore-Penrose pseudoinverse is used to perform incremental matrix factorization; and for the new-rating problem, iterative stochastic gradient descentlearning procedure is directly applied to get the updates. The solutions seem simple but efficient: extensive experiments show that they outperform the benchmark and the state-of-the-art in terms of incremental properties.

**PDMA: A Probabilistic Framework for Diversifying Recommendation Lists**, *Yang-Hui Yan*(Tsinghua University, China)*, Hai-Tao Zheng*(Tsinghua University, China)*, Ying-Min Zhou*(Tsinghua University, China)

Abstract: In this paper, we derive a probabilistic ranking framework for diversifying the recommendations of baseline methods. Unlike conventional approaches to balance relevance and diversity, we produce the diversified list by maximizing user's current marginal aspect preference, thus avoiding the hyperparameters in making the tradeoff. Before diversification, we adopt clustering to generate a much smaller set of candidate items based on three requirements: efficiency, relevance and diversity. As a result, it helps us not only reduce the search space greatly but also promote a slight increase in performance. Our framework is flexible to incorporate new preference aspects and apply new marginal aspect preference algorithms. Evaluation results show that our method can get better diversity than others and maintain comparable accuracy to baseline methods, thus a better balance between relevance and diversity.

**A Semi-Supervised Solution for Cold Start Issue on Recommender Systems**, *Zhifeng Hao*(Guangdong University of Technology, China)*, Yingchao Cheng*(Guangdong University of Technology, China)*, Ruichu Cai*(Guangdong University of Technology, China)*, Wen Wen*(Guangdong University of Technology, China)*, Lijuan Wang*(Guangdong University of Technology, China)

Abstract: The recommender system is the most competitive solution to solve information overload problem, and has been extensively applied. The current collaborative filtering based recommender systems explore users’ latent interest with their historical online behavior records. They are all facing the cold start issue. In this work, we proposed a background-based semi-supervised tri-training method named BSTM to tackle this problem. In detail, we capture fine-grained users’ background information by using a factorization model. By exploring these information, the performance of our recommendation can be significantly promoted. Besides, we proposed a semi-supervised ensemble algorithm, which got both labeled and unlabeled data involved. This algorithm assembled diverse weak prediction models which are generated by exploring samples with diverse background information and by tri-training tactic. The experimental results show that, with this solution, the accuracy of recommendation is significantly improved, and the cold-start issue is alleviated.

Research Session 10: Semantic Web and Knowledge Base

**On The Marriage of SPARQL and Keywords**, *Peng Peng*(Peking University, China)*, Lei Zou*(Peking University, China)*, Dongyan Zhao*(Peking University, China)

Abstract: To maximize the advantages of both SPARQL and keyword search, we introduce a novel paradigm that combines both of them and propose a hybrid query (called a SK query) that integrates SPARQL and keyword search in this paper. In order to answer SK queries efficiently, a structural index is devised and a novel integrated query algorithm is proposed. We evaluate our method in large real RDF graphs and experiments demonstrate both effectiveness and efficiency of our method.

**On Coherent Indented Tree Visualization of RDF Graphs**, *Qingxia Liu*(Nanjing University, China)*, Gong Cheng*(Nanjing University, China)*, Yuzhong Qu*(Nanjing University, China)

Abstract: Indented tree has been widely used to organize information and visualize graph-structured data like RDF graphs. Given a starting resource in a cyclic RDF graph as root, there are different ways of transforming the graph into a tree representation to be visualized as an indented tree. In this paper, we aim to smooth user's reading experience by visualizing an optimally coherent indented tree in the sense of featuring the fewest reversed edges, which often cause confusion and interrupt the user's cognitive process due to lack of effective way of presentation. To achieve this, we propose a two-step approach to generate such an optimal tree representation for a given RDF graph. We empirically show the difference in coherence between tree representations of real-world RDF graphs generated by different approaches. These differences lead to a significance difference in user experience in our user study, which reports a notable degree of dependence between coherence and user experience.

**Knowledge Base Completion Using Matrix Factorization**, *Wenqiang He*(Peking University, China)*, Yansong Feng*(Peking University, China)*, Lei Zou*(Peking University, China)*, Dongyan Zhao*(Peking University, China)

Abstract: With the development of Semantic Web, the automatic construction of large scale knowledge bases (KBs) has been receiving increasing attention in recent years. Although these KBs are very large, they are still often incomplete. Many existing approaches to KB completion focus on performing inference over a single KB and suffer from the feature sparsity problem. Moreover, traditional KB completion methods ignore complementarity which exists in various KBs implicitly. In this paper, we treat KBs completion as a large matrix completion task and integrate different KBs to infer new facts simultaneously. We present two improvements to the quality of inference over KBs. First, in order to reduce the data sparsity, we utilize the type consistency constraints between relations and entities to initialize negative data in the matrix. Secondly, we incorporate the similarity of relations between different KBs into matrix factorization model to take full advantage of the complementarity of various KBs. Experimental results show that our approach performs better than methods that consider only existing facts or only a single knowledge base, achieving significant accuracy improvements in binary relation prediction.

Research Session 11: Text Analysis II

**Matching Reviews to Object Based on 2-Stage CRF**, *Yongxin Zhang*(Shandong Normal University, China)*, Qingzhong Li*(Shandong University, China)*, Dequan Wang*(Fudan University, China)*, Yanhui Ding*(Shandong Normal University, China)*, Congli Liu*(Department of Information Technology, QILU Bank, China)*, Zhongmin Yan*(Shandong University, China)

Abstract: With the development of web data integration, it poses a new challenge how to match relevant reviews to integrated database objects and provide users the more complete holistic views of entities. According to the features of web data integration and reviews from web, we proposed a method based on 2-layer Conditional Random Fields(CRF) to match reviews to database objects. On the one hand, our method leverages the integrated structured entity and significantly reduces the dependence on manually labeled training data. On the other hand, we employ semi-Markov CRF to recognize the structured entities and exploit a variety of entity-level and pattern-level recognition clues available in a database of entities and labeled reviews, thereby effectively resolving the entity variety and improving the accuracy of the entity recognition. Experiments in multiple domains show that our method can substantially superior to traditional tf-idf based methods as well as a recent language model-based method for the review matching problem.

**Sentiment Word Identification with Sentiment Contextual Factors**, *Jiguang Liang*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Xiaofei Zhou*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Yue Hu*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Li Guo* (Institute of Information Engineering, Chinese Academy of Sciences, China)*, Shuo Bai*(Institute of Information Engineering, Chinese Academy of Sciences, China)

Abstract: Sentiment word identification (SWI) refers to the task of automatically identifying whether a given word expresses positive or negative opinion. SWI is a critical component of sentiment analysis technologies. Traditional sentiment word identification techniques become unqualified because they need seed sentiment words which leads to low robustness. In this paper, we consider SWI as a matrix factorization problem and propose three models for it. Instead of seed words, we exploit sentiment matching and sentiment consistency for modeling. Extensive experimental studies on three real-world datasets demonstrate that our models outperform the-state-of-art approaches.

**Sentiment Classification for Chinese Product Reviews Based on Semantic Relevance of Phrase**, *Heng Chen*(Huazhong University of Science and Technology, China)*, Hai Jin*(Huazhong University of Science and Technology, China)*, Pingpeng Yuan*(Huazhong University of Science and Technology, China)*, Lei Zhu*(Huazhong University of Science and Technology, China)*, Hang Zhu*(Huazhong University of Science and Technology, China)

Abstract: The emotional tendencies of product reviews on web have an important influence. Analysis of the sentiment of reviews on the Internet became very necessary. In this paper, a new sentiment analysis algorithm is utilized to analyze sentiment of Chinese product reviews. On training stage, a model based on skip-gram is proposed to train phrase vectors respectively on positive and negative reviews, which represent the semantic relationship of phrases. The predication of emotional tendencies of reviews based on the phrase vectors. The model does not need any modeling and feature extraction for the review data, thus it is applicable for massive data. Experimental results show that when dealing with massive data, the algorithm is better than traditional algorithms on both accuracy and learning time.

Research Session 12: MapReduce

**Distributed XML Twig Query Processing using MapReduce**, *Xin Bi*(Northeastern University, China)*, Guoren Wang*(Northeastern University, China)*, Xiangguo Zhao*(Northeastern University, China)*, Zhen Zhang*(Northeastern University, China)*, Shuang Chen*(Northeastern University, China)

Abstract: Twig query processing is one of the core operations of XML queries. Centralized holistic twig algorithms suffer great efficiency losses when large-scale XML documents are partitioned and stored in the cloud. Previous work on distributed twig query processing have some limitations, e.g., utter dependence on priori knowledge of query patterns, iteration of MapReduce jobs, etc. In this paper, our arbitrary XML partitioning and storage strategy require no knowledge of query pattern; twig queries can be efficiently processed in a single-round MapReduce job with good scalability. Extensive experiments are conducted to verify the efficiency and scalability of our algorithms.

**Large-Scale Graph Classification based on Evolutionary Computation with MapReduce**, *Zhanghui Wang*(Northeastern University, China)*, Yuhai Zhao*(Northeastern University, China)*, Guoren Wang*(Northeastern University, China)*, Yurong Cheng*(Northeastern University, China)

Abstract: Discriminative subgraph mining from a large collection of graph objects is a crucial problem for graph classification. Several main memory-based approaches have been proposed to mine discriminative subgraphs, but they always lack scalability and are not suitable for large-scale graph databases. Based on the MapReduce model, we propose an efficient method, MRGAGC, to process discriminative subgraph mining. MRGAGC employs the iterative MapReduce framework to mine discriminative subgraphs. Each map step applies the evolutionary computation and three evolutionary strategies to generate a set of locally optimal discriminative subgraphs, and the reduce step aggregates all the discriminative subgraphs and outputs the result. The iteration loop terminates until the stopping condition threshold is met. In the end, we employ subgraph coverage rules to build graph classifiers using the discriminative subgraphs mined by MRGAGC. Extensive experimental results on both real and synthetic datasets show that MRGAGC obviously outperforms the other approaches in terms of both classification accuracy and runtime efficiency.

**A Lightweight Evaluation Framework for Table Layouts in MapReduce Based Query Systems**, *Feng Zhu*(Institute of Software, Chinese Academy of Sciences, China)*, Jie Liu*(Institute of Software, Chinese Academy of Sciences, China)*, Lijie Xu*(Institute of Software, Chinese Academy of Sciences, China)*, Dan Ye*(Institute of Software, Chinese Academy of Sciences, China)*, Jun Wei*(Institute of Software, Chinese Academy of Sciences, China)*, Tao Huang*(Institute of Software, Chinese Academy of Sciences, China)

Abstract: Table layout determines the way how the relational row-column data values are organized and stored. In recent years, considerable candidates have been developed in MapReduce based query systems; they differ on storage space utilization, data loading time, query performance and so on. In most time, users are confronted with the problem of choosing the comprehensive optimum table layout given the workloads and the schema of tables. The straightforward way to run queries on generated data and compare the results is time consuming, and incurs the inaccuracy due to the MapReduce's nondeterministic execution runtime. In this paper, we propose a lightweight framework to evaluate table layouts without running the query. The framework adopts the black box method to test critical metrics, and the query aware strategy that extracts table-layout-related operations from query. Based on the metrics and operations, the framework makes suggestions to users. We conduct extensive experiments to empirically study the popular table layouts. Through the results illustration, we discover that column projection and compression are the most two prominent factors for general cases. Moreover, we discuss optimization chances for the intermediate tables produced in high level language systems.

Research Session 13: Social Network

**Sleep Quality Evaluation of Active Microblog Users**, *Kai Wu* (Shandong University, China)*, Jun Ma* (Shandong University, China)*, Zhumin Chen* (Shandong University, China)*, Pengjie Ren* (Shandong University, China)

Abstract: In this paper, we propose a novel method to evaluate the sleep quality of Active Microblog Users(AMUs) based on Sina Microblog data, where Sina Microblog is the largest microblog platform with 500 million registered users in China. A microblog user is called AMU if s/he posts more than 100 microblogs during a year. Our study is meaningful because the amount of AMUs is huge in China and the results can reflect the lifestyle of these people. The primary works of this paper are as follows: First we successfully obtained 700 million microblogs from 0.55 million microblog users as our dataset. Then we detected the possible start and end sleep time of each AMU by a novel pattern and algorithm. Finally we designed an evaluation system to give the score of each AMU's sleep quality. In the experiment, we compared the sleep quality of AMUs in different cities of China and found the difference in topics between high and low score groups by LDA method.

**Analysis of Subjective City Happiness Index Based on Large Scale Microblog Data**, *Kai Wu*(Shandong University, China)*, Jun Ma*(Shandong University, China)*, Zhumin Chen*(Shandong University, China)*, Pengjie Ren*(Shandong University, China)

Abstract: City Happiness Index(CHI) is an important societal metric to evaluate the living status of the people in a city. In traditional method, CHI is usually calculated by the combination of other objective city indicators including economy, environment, technology and education. However, happiness is a kind of subjective feeling of people rather than the measurement on how much material wealth people have gotten. Therefore we propose a novel method to evaluate Subjective City Happiness Index(SCHI) by the analysis of public sentiment on microblogs. We carried out the analysis by mining the word distribution of the microblogs in Sina Weibo, which is the largest microblog platform in China. As an application, we used the model to calculate the SCHIs of 36 major cities of China based on 55 million microblogs posted by 0.9 million unique users. Furthermore, we investigated the variety of SCHI with time in different granularities(month, day, hour) in the year 2013.

**Location Sensitive Friend Recommendation in Social Network**, *Xueqin Sui*(Shandong University, China)*, Zhumin Chen*(Shandong University, China)*, Jun Ma*(Shandong University, China)

Abstract: How to recommend friends in social network has attracted many research efforts. Most current friend recommendation methods are just based on the assumption that people will become friends if they have common interests which are usually estimated with the contents of their published posts and following relationships. However, friends recommended by these methods are only suitable for virtual social space instead of the real world. In this paper, we propose a new method to recommend friends in social network from the perspective of not just common interests, but also real-life needs. That is, we focus on finding friends that they can communicate with each other by social network and participate in some real-life activities face to face. The central idea of our approach is that we suppose people are more likely to be friends if their lives share more location overlaps besides the common interests. Currently, most people publish posts containing their real-time location information at any time, which makes it possible to detect and use the location information to recommend friends. Thus, our method combines users’ published posts, their location sequences detected from the posts and how active they are in Sina Weibo to estimate whether they can become friends in not only social network but also the real world. Experiments on Sina Weibo dataset demonstrate that our method can significantly outperform the traditional friend recommendation methods in terms of Precision, Recall and F1 measures.

**A Secure and Efficient Framework for Privacy Preserving Social Recommendation**, *Shushu Liu*(Soochow University, China)*, An Liu, Zhixu Li*(Soochow University, China)*, Lei Zhao*(Soochow University, China)*, Guanfeng Liu*(Soochow University, China)*, Pengpeng Zhao*(Soochow University, China)*, Jiajie Xu*(Soochow University, China)

Abstract: The well-known cold start problem in traditional collaborative filtering based recommender systems can be effectively addressed by social recommendation, which has been witnessed by a number of researches recently in many application domains. The social graph used in social recommendation is typically owned by a third party such as Facebook and Twitter, and should be hidden from recommender systems for obvious reasons of commercial benefits, as well as due to privacy legislation. In this paper, we present a secure and efficient framework for privacy preserving social recommendation. Our framework is built on mature cryptographic building blocks, including Paillier cryptosystem and Yao' protocol, which lays a solid foundation for the security of our framework. Using our framework, the owner of sales data and the owner of social graph can cooperatively compute social recommendation, without revealing their private data to each other. We theoretically prove the security and analyze the complexity of our framework. Empirical study shows our framework has a linear complexity with respect to the number of users and items in recommender systems and is practical in real applications.

Research Session 14: Recommendation System II

**Effective Citation Recommendation by Unbiased Reference Priority Recognition**, *Wenyang Lu* (Nanjing University, China)*, Yubin Yang*(Nanjing University, China)*, XiaoJiao Mao*(Nanjing University, China)*, Qihai Zhu*(Nanjing University, China)

Abstract: Citation recommendation is a meaningful and challenging research problem nowadays. Most of prior researches make a simplified assumption that the citations are more preferential for the papers to cite than the uncited ones. Consequently, the unreasonable priority assertion between the cited and uncited papers derived from the above assumption makes citation recommendation prone to be biased. To address this issue, we firstly propose an instinctive assumption that the more preferential a reference is, the easier it can be recognized as a citation. Based on this assumption, we propose two methods CR and CR+C aiming to find more unbiased priority between the cited and uncited papers with c-SVC. Then, a improved RankSVM model is trained for citation ranking purpose. Experimental results demonstrate that, comparing with the RankSVM model, our methods achieve 5.27% improvement on Recall@50 and 8.28% improvement on MRR. Moreover, CR+C achieves advantage on efficiency by taking only 18.9% time it needs.

**Trustworthy Collaborative Filtering through Downweighting Noise and Redundancy**, *Qiuxiang Dong*(Peking University, China)*, Zhi Guan*(Peking University, China)*, Zhong Chen*(Peking University, China)

Abstract: Proliferation of Electronic Commerce (EC) has revolutionized the way people purchase online. Web-based technologies enable people to more actively interact with merchants and service providers. Such purchasing logs and comments further lead to proliferation of recommender systems. Existing recommendation algorithms exploit either pri- or transactions or customer reviews to predict user interests towards certain items. Vast noise may be introduced into such information by fake raters, and information redundancy also makes recommender sys- tem entangled. In this work, we first examine user reviews and prior transactions to estimate user credibility and item importance to reduce effect from content polluters. Then we propose to alleviate the redundant information from homogeneous users based on link analysis. A unified framework is finally proposed to incorporate them in a mathematical formulation, which can be efficiently optimized. Experimental results on real world data reveal that our model can significantly outperform other baselines.

**User Behavioral Context-Aware Service Recommendation for Personalized Mashups in Pervasive Environments**, *Wei He*(Shandong University, China)*, Guozhen Ren*(Shandong University, China)*, Lizhen Cui*(Shandong University, China)*, Hui Li*(Shandong University, China)

Abstract: With the rapid development of mobile Internet and increasing amount of smart devices, Internet services have been integrated into peoples’ daily lives. Due to the features of end-user-oriented mashups in pervasive environments, new challenges have been presented to conventional mashup approaches, including the complexity of user behavior process, personalized service provisions, difficulty of predicting real-time user preference and other dynamic contexts. In this paper, we propose a new paradigm for behavioral context-based personalized mashup provision in pervasive environments by combing user behavior and mashup instance execution into an integrated process. In the proposed paradigm, users with similar behavior patterns are computed and then probability distributions of potential behavior selection for user clusters are discovered from historical mashup logs, which provide supports for predicting and recommending user activities for future mashup constructions. Analysis and experiments indicate that our approach can effectively simplify personalized mashup composition without depending on end-users’ professional knowledge, as well as improve the quality of mashup composition and recommendation based on behavioral contexts and personalization in pervasive environments.

**Online Personalized Recommendation Based on Streaming Implicit User Feedback**, *Zhisheng Wang*(Sun Yat-sen University, China)*, Qi Li*(Sun Yat-sen University, China)*, Ye Liu*(Sun Yat-sen University, China)*, Wei Liu*(Sun Yat-sen University, China)*, Jian Yin*(Sun Yat-sen University, China)

Abstract: Since user preference is drifting over time, modeling temporary dynamic recommender system has been proven to be valuable for accurate recommendation performance. However, user feedback is continuously updating while the traditional recommender system is trained off-line in batch mode so that it can’t capture user taste change in time. In this paper, we build a dynamic real-time recommendation model based on implicit user feedback stream to improve both the recommendation accuracy and training efficiency. Moreover, our model has obvious advantages over the traditional approaches in diversity, interpretability, and strong robustness to hostile attack. Finally, we conduct experiments on two real world datasets to validate the effectiveness of our proposed method and demonstrate the superior performance when compared with state-of-the-art approach.

Research Session 15: Location-Based Service

**Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks**, *Yang Zhang*(University of Luxembourg, Luxembourg)*, Jun Pang*(University of Luxembourg, Luxembourg)

Abstract: With the emerging of location-based social networks, study on the relationship between human mobility and social relationships becomes quantitatively achievable. Understanding it correctly could result in appealing applications, such as targeted advertising and friends recommendation. In this paper, we focus on mining users' relationship based on their mobility information. More specifically, we propose to use distance between two users to predict whether they are friends. We first demonstrate that distance is a useful metric to separate friends and strangers. By considering location popularity together with distance, the difference between friends and strangers gets even larger. Next, we show that distance can be used to perform an effective link prediction. In addition, we discover that certain periods of the day are more social than others. In the end, we use a machine learning classifier to further improve the prediction performance. Extensive experiments on a Twitter dataset collected by ourselves show that our model outperforms the state-of-the-art solution by 30%.

**Hybrid-LSH for Spatio-Textual Similarity Queries**, *Mingdong Zhu*(Northeastern University, China)*, Derong Shen*(Northeastern University, China)*, Ling Liu*(Georgia Institute of Technology, USA)*, Ge Yu*(Northeastern University, China)

Abstract: Locality Sensitive Hashing (LSH) is a popular method for high dimensional indexing and search over large datasets. However, little efforts have put forward to utilizing LSH in mobile applications for processing spatio-textual similarity queries, such as find nearby shopping centers that have a top ranked hair salon. In this paper, we present hybrid-LSH, a new LSH method for indexing data objects according to both their spatial location and their keyword similarity. Our hybrid-LSH approach has two salient features: First our hybrid-LSH carefully combines the spatial location based LSH and textual similarity based LSH to ensure the correctness of the spatial and textual similarity based NN queries. Second, we present an adaptive query-processing model to address the fixed range problem of traditional LSH and to handle queries with varying ranges effectively. Extensive experiments conducted on both synthetic and real datasets validate the efficiency of our hybrid LSH method.

**Answering Spatial Approximate Keyword Queries on Disks**, *Jinbao Wang*(Harbin Institute of Technology, China)*, Donghua Yang*(Harbin Institute of Technology, China)*, Yuhong Wei*(ZTE Co. Ltd., Shenzhen, China)*, Hong Gao*(Harbin Institute of Technology, China)*, Jianzhong Li*(Harbin Institute of Technology, China)*, Ye Yuan*(Harbin Institute of Technology, China)

Abstract: Spatial approximate keyword queries consist of a spatial condition and a set of keywords as the fuzzy textual conditions, and they return objects labeled with a set of keywords similar to queried keywords while satisfying the spatial condition. Such queries enable users to find objects of interest in a spatial database, and make mismatches between user query keywords and object keywords tolerant. With the rapid growth of data, spatial databases storing objects from diverse geographical regions can be no longer held in main memories. Thus, it is essential to answer spatial approximate keyword queries over disk resident datasets. Existing works present methods either returns incomplete answers or indexes in main memory, and effective solutions in disks are in demand. This paper presents a novel disk resident index RMB-tree to support spatial approximate keyword queries. We study the principle of augmenting R-tree with capacity of approximate keyword searching based on existing solutions, and store multiple bitmaps in R-tree nodes to build an RMB-tree. RMB-tree supports spatial conditions such as range constraint, combined with keyword similarity metrics such as edit distance, dice etc. Experimental results against R-tree on two real world datasets demonstrate the efficiency of our solution.

**Reverse Direction-based Surrounder Queries**, *Xi Guo*(University of Science and Technology Beijing, China)*, Yoshiharu Ishikawa*(Nagoya University, Japan)*, Aziguli Wulamu*(University of Science and Technology Beijing, China)*, Yonghong Xie*(University of Science and Technology Beijing, China)

Abstract: This paper proposes a new spatial query called the reverse direction-based surrounder (RDBS) query, which retrieves a user who is seeing a point of interest (POI) as one of their direction-based surrounders (DBSs). According to a user, one POI can be dominated by a second POI if the POIs are directionally close and the first POI is farther from the user than the second is. Two POIs are directionally close if their included angle with respect to the user is smaller than an angular threshold. If a POI cannot be dominated by another POI, it is a DBS of the user. We also propose an extended query called the competitor RDBS query. POIs that share the same RDBSs with another POI are defined as competitors of that POI. We design algorithms to answer the RDBS queries and competitor queries. The experimental results show that the proposed algorithms can answer the queries efficiently.

Research Session 16: Graph Processing

**GSCS - Graph Stream Classification with Side Information**, *Amit Mandal*(University of Dhaka, Bangladesh)*, Mehedi Hasan*(University of Dhaka, Bangladesh)*, Anna Fariha*(University of Dhaka, Bangladesh)*, Chowdhury Farhan Ahmed*(University of Strasbourg, France)

Abstract: With the popularity of applications like Internet, sensor network and social network, which generate graph data in stream form, graph stream classification has become an important problem. Many applications are generating side information associated with graph stream, such as terms and keywords in authorship graph of research papers or IP addresses and time spent on browsing in web click graph of Internet users. Although side information associated with each graph object contains semantically relevant information to the graph structure and can contribute much to improve the accuracy of graph classification process, none of the existing graph stream classification techniques consider side information. In this paper, we have proposed an approach, Graph Stream Classification with Side information (GSCS), which incorporates side information along with graph structure by increasing the dimension of the feature space of the data for building a better graph stream classification model. Empirical analysis by experimentation on two real life datasets is provided to depict the advantage of incorporating side information in the graph stream classification process to outperform the state of the art approaches. It is also evident from the experimental results that GSCS is robust enough to be used in classifying graphs in form of stream.

**An Extended Graph-based Label Propagation Method for Readability Assessment**, *Zhiwei Jiang*(Nanjing University, China)*, Gang Sun*(Nanjing University, China)*, Qing Gu*(Nanjing University, China)*, Lixia Yu*(Nanjing University, China)*, Daoxu Chen*(Nanjing University, China)

Abstract: Readability assessment is to evaluate the reading difficulty of a document, which can be quantified as reading levels. In this paper, we propose an extended graph-based label propagation method for readability assessment. We employ three vector space models (VSMs) to compute edges and weights for the graphs, along with three graph sparsification techniques. By incorporating the pre-classification results, we develop four strategies to reinforce the graphs before label propagation to capture the ordinal relation among the reading levels. The reinforcement includes recomputing weights for the edges, and filtering out edges linking nodes with big level difference. Experiments are conducted systematically on datasets of both English and Chinese. The results demonstrate both effectiveness and potential of the proposed method.

**AILabel: A Fast Interval Labeling Approach for Reachability Query on Very Large Graphs**, *Feng Shuo*(Northeastern university, China)*, Derong Shen*(Northeastern university, China)

Abstract: Recently, reachability queries on large graphs have attracted much attention. Many state-of-the-art approaches leverage spanning tree to construct indexes. However, almost all of these work require indexes and original graph in memory simultaneously, which will limit the scalability. Thus, a new interval labeling ap-proach called AILabel (Augmented Interval Label) is proposed in this paper. AILabel labels each node with quadruples. Index construction time of AILabel is O(m+n), which requires only one traversal through the graph. Besides, AILabel only needs index to answer the queries. We further proposed an approach D-AILabel based on AILabel to handle reachability queries on dynamic graph. Fi-nally, experiments on real and synthetic datasets are conducted and prove that AILabel can efficiently scale to large graph and reachability queries on dynamic graph can be effectively handled by D-AILabel.

**Graph-based Hybrid Recommendation Using Random Walk and Topic Modeling**, *Yang-Hui Yan*(Tsinghua University, China)*, Hai-Tao Zheng*(Tsinghua University, China)*, Ying-Min Zhou*(Tsinghua University, China)

Abstract: In this paper, we propose a graph-based method for hybrid recommendation. Unlike a simple linear combination of several factors, our method integrates user-based, item-based and content-based techniques more fully. The interaction among different factors are not done once, but by iterative updates. The graph model is composed of target user's similar-minded neighbors, candidate items, target user's historical items and the topics extracted from items' contents using topic modeling. By constructing the concise graph, we filter out irrelevant noise and only retain useful information which is highly related to the target user. Top-N recommendation list is finally generated by using personalized random walk. We conduct a series of experiments on two datasets: movielen and lastfm. Evaluation results show that our proposed approach achieves good quality and outperforms existing recommendation methods in terms of accuracy, coverage and novelty.

Research Session 17: Data Mining

**Mining Weighted Frequent Itemsets with the Recency Constraint**, *Jerry Chun-Wei Lin*(Harbin Institute of Technology Shenzhen Graduate School, China)*, Wensheng Gan*(Harbin Institute of Technology Shenzhen Graduate School, China)*, Philippe Fournier-Viger*(University of Moncton, Canada)*, Tzung-Pei Hong*(National University of Kaohsiung, Taiwan; National Sun Yat-sen University, Taiwan)

Abstract: Weighted Frequent Itemset Mining (WFIM) has been proposed as an alternative to frequent itemset mining that considers not only the frequency of items but also their relative importance. However, an important limitation of WFIM is that it does not consider how recent the patterns are. To address this issue, we extend WFIM to consider the recency of patterns, and thus present the Recent Weighted Frequent Itemset Mining (RWFIM). A projection-based algorithm named RWFIM-P is designed to mine Recent Weighted Frequent Itemsets (RWFIs) based on a novel upper-bound downward closure property. Moreover, an improved algorithm named RWFIM-PE is also proposed, which introduces a new pruning strategy named Estimated Weight of 2-itemset Pruning (EW2P) to prune unpromising candidate of RWFIs early. An experimental evaluation against a state-of-the-art WFIM algorithm on several real-world and synthetic datasets show that the proposed algorithms are highly efficient.

**Boosting Explicit Semantic Analysis by Clustering Paragraph Vectors of Wikipedia Articles**, *Hai-Tao Zheng* (Tsinghua University, China)*, Wenzhen Wu* (Tsinghua University, China)

Abstract: Explicit Semantic Analysis (ESA) is an effective method that utilizes Wikipedia entries (articles) to represent text and compute semantic relatedness (SR) for text pairs. Analogous to ordinary web search techniques, ESA also suffers from the redundancy issues due to the ongoing expansion of the amount of Wikipedia entries. Entries redundancy could lead to biased representation that lay particular emphasis on semantics from a large number of similar entries. On the other hand, original ESA for SR has a weak point that it does not consider the correlations or similarities between the Wikipedia articles of the text representations. To tackle these problems, We develop a novel method to cluster the redundant or similar entries by similarity measurement based on Paragraph Vector (PV), a neural network language model. Results of experiments on four datasets show that our framework could gain better performance in relatedness accuracy against ESA.

**Probabilistic Frequent Pattern Mining by PUH-Mine**, *Wenzhu Tong*(University of Manitoba, Canada; Wuhan University, China; University of Illinois at Urbana-Champaign,USA)*, Carson Leung*(University of Manitoba, Canada)*, Dacheng Liu*(University of Manitoba, Canada; Wuhan University, China; Carnegie Mellon University, USA)*, Jialiang Yu, Fan Jiang*(University of Manitoba, Canada)

Abstract: To mine frequent itemsets from uncertain data, many existing algorithms rely on expected support based mining. An alternative approach relies on probabilistic based mining, which captures the frequentness probability. While the possible world semantics are widely used, the exponential growth of possible worlds makes the probabilistic based mining computationally challenging when compared to the expected support based mining. In this paper, we propose two efficient approximate hyperlinked structure based algorithms, which generate a collection of all potentially probabilistic frequent itemsets with a novel upper bound and verify if they are truly probabilistic frequent. Experimental results show the efficiency of our algorithms in mining probabilistic frequent itemsets from uncertain data.

**Learning to Hash for Recommendation with Tensor Data**, *Qiyue Yin*(Institute of Automation, Chinese Academy of Sciences, China)*, Shu Wu*(Institute of Automation, Chinese Academy of Sciences, China)*, Liang Wang*(Institute of Automation, Chinese Academy of Sciences, China)

Abstract: Recommender systems usually need to compare user interests and item characteristics in the context of large user and item space, making hashing based algorithms a promising strategy to speed up recommendation. Existing hashing based recommendation methods only model the users and items and dealing with the matrix data, e.g., user-item rating matrix. In practice, recommendation scenarios can be rather complex, e.g., collaborative retrieval and personalized tag recommendation. The above scenarios generally need fast search for one type of entities (target entities) using multiple types of entities (source entities). The resulting three or higher order tensor data makes conventional hashing algorithms fail for the above scenarios. In this paper, a novel hashing method is accordingly proposed to solve the above problem, where the tensor data is approached by properly designing the similarities between the source entities and target entities in Hamming space. Besides, operator matrices are further developed to explore the relationship between different types of source entities, resulting in auxiliary codes for source entities. Extensive experiments on two tasks, i.e., personalized tag recommendation and collaborative retrieval, demonstrate that the proposed method performs well for tensor data.

Research Session 18: Influence Maximization

**A Co-Ranking Framework to Select Optimal Seed Set for Influence Maximization in Heterogeneous Network**, *Yashen Wang*(Beijing Institute of Technology, China)*, Heyan Huang*(Beijing Institute of Technology, China)*, Chong Feng*(Beijing Institute of Technology, China)*, Xianxiang Yang*(Beijing Institute of Technology, China)

Abstract: The rising popularity of social media presents new opportunities for one of the enterprise’s most important needs—selecting most influential individuals in viral marketing, which has attracted increasing attention in both academia and industry. Most recent algorithms of influence maximization have demonstrated remarkable successes, however their applications are limited to homogeneous networks. In this paper, we formulate the problem of influence maximization in heterogeneous network, and propose a co-ranking framework to simultaneously select seed sets with different types. This framework is flexible and could adequately takes ad-vantage of additional information implicit in the heterogeneous structure. We conduct extensive experiments using the data collected from ACM Digital Li-brary, and the experimental results show that both the quality and the running time of the proposed algorithm rival the existing algorithms.

**UserGreedy : Exploiting the Activation Set to Solve Influence Maximization Problem**, *Wenxin Liang*(Dalian University of Technology, China)*, Chengguang Shen*(Dalian University of Technology, China)*, Xianchao Zhang*(Dalian University of Technology, China)

Abstract: solutions are mainly divided into two directions. The one is greedy-based methods and the other is heuristics-based methods. The greedy-based methods can effectively estimate influence spread using thousands of Monte-Carlo simulations. However, the computational cost of simulation is extremely expensive so that they are not scalable to large networks. The heuristics-based methods, estimating influence spread according to heuristic strategies, have low computational cost but without theoretical guarantees. In order to improve both performance and effectiveness, in this paper we propose a greedy-based algorithm, named UserGreedy. In UserGreedy, we first propose a novel concept called Activation Set, which is defined as a set of users that can be activated by a seed user with a certain probability under the most standard and popular independent cascade (IC) model. Based on the computation of such probabilities, we can directly estimate the influence spread without the expensive simulation process. We then design an influence spread function based on the the Activation Set and mathematically prove that it has the property of monotonicity and submodularity, which provides theoretical guarantee for the UserGreedy algorithm. Besides, we also propose an efficient method to obtain the Activation Set and hence implement the UserGreedy algorithm. Experiments on real-world social networks demonstrate that our algorithm is much faster than existing greedy-based algorithms and outperforms the state-of-art heuristics-based algorithms in terms of effectiveness.

**Minimizing the Cost to Win Competition in Social Network**, *Ziyan Liu*(Shandong University, China)*, Xiaguang Hong*(Shandong University, China)*, Zhaohui Peng*(Shandong University, China)*, Zhiyong Chen*(Shandong University, China)*, Weibo Wang*(Shandong University, China)*, Tianhang Song*(Shandong University, China)

Abstract: In social network, influences are propagating among users. Influence maximizing is an important problem which has been studied by many research in recent years. However,in the real world,sometimes users need to make their influence maximization through social network and defeat their competitors under minimum cost, such as online voting, expanding market share,etc. In this paper, we consider the problem of select a seed set with the minimum cost to influence more people than other competitors. We show this problem is NP-hard and propose a cost-effective greedy algorithm to approximately solve the problem and improve the efficiency based on the submodularity. Furthermore, a new Degree Adjust heuristics is proposed to get high efficiency. Experimental results show that our cost-effective greedy algorithm achieves better effectiveness than other algorithms and cost-effective Degree Adjust heuristic algorithm achieves high efficiency and get better effectiveness than Degree and Random heuristics.

**Ad Dissemination Game in Ephemeral Networks**, *Lihua Yin*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Yunchuan Guo*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Yanwei Sun*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Junyan Qian*(Institute of Information Engineering, Chinese Academy of Sciences, China)*, Thanos Vasilakos*(University of Western Macedonia, Greece)

Abstract: The dissemination of ads in an ephemeral network has become an important research topic. A challenge of ad dissemination is the guarantee of robustness such that rational nodes in the ephemeral network have sufficient impetus to forward ads, despite facing the limitation of resources and the risk of privacy leakage. This paper proposes a strategy for ad dissemination in an ephemeral network. Acknowledging the assumption of incomplete information, we propose a bargaining-based game GD that can be used by nodes to decide whether to forward an ad. There exists Bayesian Nash equilibrium in GD with which the proposed approach provides nodes a strong impetus to disseminate the ads with higher dissemination accuracy.

Industrial Session 1: Cloud and Real-Time Computation

**Hybrid Cloud Deployment of an ERP-based Student Administration System**, *Simon K.S. Cheung*(The Open University of Hong Kong, Hong Kong)

Abstract: Conceptually, a hybrid cloud approach to deploying an ERP system can effectively achieve a cost-performance balance, especially during seasonal peaks of concurrent accesses. This paper describes an actual implementation of an ERP-based student administration system, where the web tier and application tier are deployed on hyrbid cloud with the databases resided on premises. The key challenge of this implementation is to set up the web tier and application tier for the right cost-performance balance while attaining an appropriate level of resilience. In this paper, technical evaluation is conducted to demonstrate the cost efficiency of the hybrid cloud deployment, and the benefits are discussed. These provide a useful reference on hybrid-cloud deployment of ERP systems.

**A Benchmark Evaluation of Enterprise Cloud Infrastructure**, *Yong Chen*(Guangzhou Research Institute, China Telecom Co. Ltd., China)*, Xuanzhong Xie*(Guangzhou Research Institute, China Telecom Co. Ltd., China)*, Jiayin Wu*(Guangzhou Research Institute, China Telecom Co. Ltd., China)

Abstract: As cloud computing matures in enterprise IT system, agile and scalable evaluation of cloud infrastructure becomes critical. A lightweight benchmark approach is introduced, to meet the requirement of low cost, scalability and relevance of typical enterprise environment. Benchmark results are analyzed and compared to industry benchmark data, conclusion is given on implementing a scientific evaluation of enterprise cloud infrastructure.

**A Fast Data Ingestion and Indexing Scheme for Real-time Log Analytics**, *Haoqiong Bian*(Renmin University of China, China)*, Yueguo Chen*(Renmin University of China, China)*, Xiongpai Qin*(Renmin University of China, China)*, Xiaoyong Du*(Renmin University of China, China)

Abstract: Structured log data is a kind of append-only time-series data which grows rapidly as new entries are continuously generated and captured. It has become very popular in application domains such as Internet, sensor networks and telecommunications. In recent years, many systems have been developed to support batch analysis of such structured log data. But they often fail to meet the high throughput requirements of real-time log data ingestion and analytics. An efficient index is very important to accelerate log data analytics, and at the meanwhile to support high throughput data loading. This paper focuses on designing a specialized indexing scheme for real-time log data analytics. The solution adopts a dynamic global hash index to partition the tuples into hash buckets. Then the tuples in the hash buckets are sorted and buffered in the sorted buffer queue. When the amount of data in the queue reaches a threshold, the data is packed into segments before spilling to the disks. Moreover, an intra-segment index is maintained by meta database. With such an indexing scheme, the database system achieves high throughput and real-time data loading and query performance. As shown in the experiments, the data loading throughput reaches 5 million tuples per second per node. The delay of data loading does not exceed 10 seconds, and a sub-second query performance is achieved for the given queries.

Demonstration Session:

**A Fair Data Market System with Data Quality Evaluation and Repairing Recommendation**, *Xiaoou Ding*(Harbin Institue of Technology, China)*, Hongzhi Wang*(Harbin Institue of Technology, China)*, Dan Zhang*(Harbin Institue of Technology, China)*, Jianzhong Li*(Harbin Institue of Technology, China)*, Hong Gao*(Harbin Institue of Technology, China)

Abstract: With the development of data market, the data resource is playing a key role as the part of business resources. However, existing conventional data markets are too simple to reveal the real data values in practical application. Taking the data quality into deep consideration, we design a fair data price evaluation mechanism, which aims at meeting the needs of both supply and demand. For the data quality issues in the data market, several critical factors, namely accuracy, complete-ness, consistency, and currency, affecting data quality are integrated in order to show comprehensive assessment of the data. Moreover, our system can also provide data repairing recommendation based on data quality evaluation. Our system demonstrates the essential attributes of the data market in terms of efficiency, fairness, effectiveness and security.

**HouseIn: A Housing Rental Platform with Non-Redundant Information Integrated from Multiple Sources**, *Jian Zhou*(Soochow University, China)*, Zhixu Li*(Soochow University, China)*, Qiang Yang*(Soochow University, China)*, Jun Jiang*(Soochow University, China)*, An Liu*(Soochow University, China)*, Guanfeng Liu*(Soochow University, China)*, Lei Zhao*(Soochow University, China)*, Jia Zhu*(South China Normal University, China)

Abstract: Housing Rental Platforms (HRPs) such as rentalhouses.com, 58.com, ganji.com provide convenient ways to find accommodations, but the redundancy of the rental advertising information within and across platforms brings unpleasant user experience. Besides, rental advertisements are usually presented in the form of a list, which can hardly give users a straightforward big picture about the housing rental market of a particular interested area. In this demonstration, we introduce HouseIn, a novel HRP that expect to: 1) provide users a clear big picture in several aspects about the housing rental market of a particular interested area; and 2) detect those advertisements referring to the same property, such that to give users a price comparison between various platforms for the same apartment for rental. The core challenge in implementing HouseIn lies on the Record Matching problem between these advertisements, given that there is no exact Unit/Building Number of the apartment for the sake of privacy issue, and the detailed information about the house can be various given by different agencies. To handle the Record Matching(RM) problem between these advertisements, we employ several state-of-the-art RM methods plus Information Extraction (IE) techniques to use both structured and unstructured information in the advertisements. We show the advantages of HouseIn with several demonstration scenarios.

**ONCAPS: An Ontology-based Car Purchase Guiding System**, *Jianfeng Du*(Guangdong University of Foreign Studies, China)*, Jun Zhao*(Lancaster University, United Kingdom)*, Jiayi Cheng*(Guangdong University of Foreign Studies, China)*, Qingchao Su*(Guangdong University of Foreign Studies, China)*, Jiacheng Liang*(Guangdong University of Foreign Studies, China)

Abstract: In this paper, an ontology-based car purchase guiding system, ONCAPS, is presented. The system integrates Semantic Web technologies to provide car information to people who want to buy cars. It employs ontology-based data integration technology to gather data from multiple car selling websites in China. It captures personalized weights on car features by an analytic hierarchy process based approach. It computes the personalized distance between a car and a given requirement by a weighted sum model where ontology reasoning is involved. For purchase guiding, ONCAPS suggests the user all cars that have the minimum personalized distance to the given requirement. The suggestions are shown with explanations on why these cars are recommended.

**A Multiple Trust Paths Selection Tool in Contextual Online Social Networks**, *Linlin Ma*(Soochow University, China)*, Guanfeng Liu*(Soochow University, China)*, Guohao Sun*(Soochow University, China)*, Lei Li*(Heifei University of Technology, China)*, Zhixu Li*(Soochow University, China)*, An Liu*(Soochow University, China)*, Lei Zhao*(Soochow University, China)

Abstract: Online Social Network (OSN) is becoming popular, where trust is one of the most important factors for participants’ decision making. This requires to efficiently and effectively select those social trust paths that can yield the most trustworthy trust evalu- ation results between two unknown participants to establish reasonable trust relationships between them. Thus, we develop a Multiple Trust Paths (MTP) selection tool based on the state-of-the-art trust paths selection method proposed by us, which considers the so- cial contexts, like the social trust and social intimacy degree between participants, and the role impact factor of participants in trust paths selection. Our tool could help users evaluate the trustworthiness of unknown participants in a variety of OSN based applications, for in- stance, to help a retailer identify new trustworthy customers and introduce products to them, or help users select the trustworthy workers in OSN based crowdsourcing platforms.

**EPEMS: An Entity Matching System For E-Commerce Products**, *Lei Gao*(Soochow University, China)*, Pengpeng Zhao*(Soochow University, China)*, Victor S. Sheng*(University of Central Arkansas, USA)*, Zhixu Li*(Soochow University, China)*, An Liu*(Soochow University, China)*, Jian Wu*(Soochow University, China)*, Zhiming Cui*(Soochow University, China)

Abstract: Entity Matching is used to identify records representing the same entities in the real world. As e-commerce is developing rapidly, online products grow explosively in both amount and variety. Applying entity matching to e-commerce data and finding records representing the same products make customers convenient to compare prices. This paper proposes an entity matching system for e-commerce data, called EPEMS. Compared with existing systems, we improve an existing sorted neighborhood blocking method, which is used to reduce the number of comparisons. At the same time the similarity of product pictures is used to improve matching results.

**PPS-POI-Rec: A Privacy Preserving Social Point-of-Interest Recommender System**, *Xiao Liu*(Soochow University, China)*, An Liu*(Soochow University, China)*, Guanfeng Liu*(Soochow University, China)*, Zhixu Li*(Soochow University, China)*, Jiajie Xu*(Soochow University, China)*, Pengpeng Zhao*(Soochow University, China)*, Lei Zhao*(Soochow University, China)

Abstract: Point-of-Interest (POI) recommendation is an important task for location based service (LBS) providers. Social POI recommendation outperforms traditional, non-social approaches as social relations among users (a.k.a. social graph) could be used as a data source to calculate user similarities, which is generally hard to evaluate due to the lack of sufficient user check-in data. However, the social graph is typically owned by a social networking service (SNS) provider such as Facebook and should be hidden from LBS provider for obvious reasons of commercial benefits, as well as due to privacy legislation. In this paper, we present PPS-POI-Rec, a novel privacy preserving social POI recommender system that enables SNS provider and LBS provider to cooperatively recommend a set of POIs to a target user while keeping their private data secret. We will demonstrate step by step how a social POI recommendation can be jointly made by SNS provider and LBS provider, without revealing their private data to each other.

**Incorporating Contextual Information into a Mobile Advertisement Recommender System**, *Ke Zhu*(Tianjin University of Technology, China)*, Yingyuan Xiao*(Tianjin University of Technology, China)*, Pengqiang Ai*(Tianjin University of Technology, China)*, Hongya Wang*(Donghua University, China)*, Ching-Hsien Hsu*(Chung Hua University, Taiwan)

Abstract: The ever growing popularity of smart mobile devices and rapid advent of wireless technology have given rise to a new class of advertising system, i.e., mobile advertisement recommender system. The traditional internet advertising systems have largely ignored the fact that users interact with the system within a particular “context”. In this paper, we implemented a mobile advertisement recommender prototype system called MARS. MARS captures different user’s contextual information to improve recommendation results. The demonstration shows that MARS makes advertisement recommendation more effectively.

**Notes**