THE 4TH INTERNATIONAL WORKSHOP ON DATAOLOGY AND DATA SCIENCE (IWDD2013)

May 25 - 27, 2013

SHANGHAI & ZHENGZHOU, CHINA

http://iwdds.fudan.edu.cn/

Hosted By:

数据科学研究中心
http://www.dataology.fudan.edu.cn
Research Center for Dataology and DataScience

Sponsored By:

THE 4TH INTERNATIONAL WORKSHOP

ON

DATAOLOGY AND DATA SCIENCE

(IWDD2013)

May 25- 27, 2013

SHANGHAI&ZHENGZHOU, CHINA

http://iwdds.fudan.edu.cn/

Hosted by:

Research Center for Dataology and DataScience (RCDD)

Sponsored by:

School of Computer Science, Fudan University
Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences
School of Information Engineering, Zhengzhou University
InfinityData Investment Co., Ltd.

Table of Content

Welcome Message from General Chairs

On behalf of IWDD Committee, we would like to welcome you all to the 4thInternational Workshop on Dataology and Data Science (IWDD2013) held at Fudan University, Shanghai, and at Zhengzhou University, Zhengzhou, China.

It is a milestone for Dataology and Data Science in 2013 due to Big Data popularity, and more and more people are realizing the importance of data and analytics. We continue to grow and adapt, remaining always open to new ideas. Our organization is confronting a time of many changes and we are meeting these changes during a time of larger nation-wide and global change. We are exciting to see Dataology and Data Science to be an emerging discipline, and we will continue to meet and bring inspired people from interdisciplinary areas together in forums like this.

The workshop is divided into two parts. From May 25 to 26, the first part will be held in Shanghai with 3 sessions, namely, “Challenges of Big Data”, “Scientific research methods with data” and “Big Data query and mining”. Then, the remaining one will move to Zhengzhou on May 27, and the discussion topic is related to “Data Mining and Its Application”.

Compared to the last 3 years, we can see three big changes in this year. Firstly, the participants come from more disciplines, such as mathematics, astronomic, remote sensing, physics, finance. Secondly, the experimental methods of Dataology, i.e., data-driven scientific research methods, will be presented with successful applications, such as healthcare, astronomic, IPTV recommendation system. Finally, we believe that Dataology and Data Science is an emerging discipline, and thus, a disciplinary construction will be discussed in the last session held inShanghai.

The 4thInternational Workshop on Dataology and Data Science (IWDD2013) is hosted by Research Center for Dataology and DataScience (RCDD), and sponsored by School of Computer Science, Fudan University, Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, School of Information Engineering, Zhengzhou University, and InfinityData Investment Co., Ltd.

We all look forward to many excellent technical and social interactions during this 3-day workshop. We encourage all of you to fully participate in the technical and social events. We wish you an enjoyable and impressive meeting both in Shanghai and in Zhengzhou, and we thank you for attending!

General Chairs

Yangyong Zhu (Fudan University, China)

Xueming Si (Information Engineering University, China)

Organization

General Chairs:

Yangyong Zhu (Fudan University, China)
Xueming Si (Information Engineering University, China)

Program Chairs:

Yong Shi (Chinese Academy of Sciences, China)
Qinglei Zhou (Zhengzhou University, China)

Local Organizing Chairs:

Zhihong Zhang (Zhengzhou University, China)
Yitong Wang (Fudan University, China)
Gang Liu (Fudan University, China)

Venue

Workshop Venue: Lecture Hall 106, the Administration Building, Zhangjiang Campus, Fudan University
Address: No.825 Zhangheng Road, PudongNew Area District, Shanghai, China
(Place B in the map)

Hotel: Pudong International Talent City Hotel
Address: No. 1500, Keyuan Road, PudongNew Area District,Shanghai, China
(Place A in the map)

Remark: it takes about 15 minutes walking from the hotel to the venue.

Program at a Glance

<table>
<tr><th colspan="2">Day 0
Friday, May 24, 2013
Hotel Lobby*</th><th colspan="2">Day 1
Saturday, May 25, 2013
Fudan University**</th><th colspan="2">Day 2
Sunday, May 26, 2013
Fudan University**</th><th colspan="2">Day 3
Monday, May 27, 2013
Zhengzhou University***</th></tr>
<tr><td>09:00</td><td rowspan="10">Registration</td><td>08:00</td><td>Welcome Desk /Registration</td><td>08:00</td><td>Welcome Desk /Registration</td><td>08:30</td><td>Welcome Desk /Registration</td></tr>
<tr><td></td><td>08:30</td><td>Welcome Remarks</td><td rowspan="2">08:30</td><td rowspan="3">Session 3</td><td>09:00</td><td>Welcome Remarks</td></tr>
<tr><td></td><td>08:45</td><td rowspan="3">Session 1</td><td rowspan="2">09:15</td><td rowspan="3">Session 4</td></tr>
<tr><td></td><td></td><td></td><td>10:00</td><td>Break</td></tr>
<tr><td></td><td>10:15</td><td>Break</td><td>10:15</td><td rowspan="2">Big Data Query and Mining</td><td>10:45</td><td>Break</td></tr>
<tr><td></td><td>10:30</td><td>Challenge of Big Data</td><td>11:00</td><td rowspan="2">Data Mining and Its Application</td></tr>
<tr><td></td><td>12:00</td><td>Lunch</td><td>12:30</td><td>Lunch</td><td>12:00</td><td>Lunch</td></tr>
<tr><td></td><td>13:00</td><td rowspan="3">Session 2</td><td rowspan="4">13:30</td><td rowspan="4">End of Workshop in Shanghai</td><td rowspan="4">13:00</td><td rowspan="4">End of Workshop in Zhengzhou</td></tr>
<tr><td></td><td>15:00</td><td>Break</td></tr>
<tr><td></td><td>15:15</td><td>Scientific Research Method with Data</td></tr>
<tr><td></td><td>17:00</td><td>Panel
Big data, huge data, any data, what is the most important thing behind these?</td></tr>
<tr><td>21:00</td><td></td><td>18:00</td><td>Banquet</td><td></td><td></td><td></td><td></td></tr>
</table>

***** *Registration desk at lobby of Pudong International Talent CityHotel, Shanghai, China*

****** *Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University, Shanghai, China*

******* *Venue in Zhengzhou Universitywill be setup by the local Organizing team of Zhengzhou University.*

Shuttle Bus Time Table

May 25, 2013, **08:15** **Hotel** → **Venue** (Zhangjiang Campus)
18:00 **Venue** → **Hotel** (for Banquet)****

May 26, 2013, **08:15** **Hotel** → **Venue** (Zhangjiang Campus)
13:30 **Restaurant** → **Hotel****

******** *Estimated departure time*

Detail Program

Time	Agenda
Day 0	**Friday, May 24, 2013** ***Pudong International Talent City Hotel***
09:00-21:00	Registration
Day 1	**Saturday, May 25, 2013** ***Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University***
08:30-08:45	Welcome Remarks
	Session 1: Challenge of Big Data (8:45-12:00) Chair: Chengqi Zhang (University of Technology Sydney, Australia)
08:45-09:15	On Mining Big Data Philip S.Yu, University of Illinois at Chicago, USA
09:15-09:45	Projected Gradient Method for Sparse Principal Component Analysis Weiguo Gao, Fudan Univesity, China
09:45-10:15	Data Driven Semantic Computing Haixun Wang, Microsoft Research Asia
10:15-10:30	Break
10:30-11:00	Catch the Wind: Graph Workload Balancing on Cloud Jeffrey XuYu, the Chinese University of Hong Kong, HK
11:00-11:30	Big Data-enabled Management Decision Making Lihua Huang, Fudan University, China
11:30-12:00	Big Data and Decision Making Yong Shi, Chinese Academy of Sciences, China
12:00-13:00	Lunch
	Session 2: Scientific Research Method with Data (13:00-17:00) Chair: Wei Wang (Fudan University, China)
13:00-13:30	Research Issues and Challenges on Brain Informatics Ning Zhong, Maebashi Institute of Technology,Japan
13:30-14:00	Data Challenge in High Energy Physics Gang Chen, Institute of High Energy Physics, Chinese Academy of Science, China
14:00-14:30	Data Blogging: A New Implication for Science Impact Weigang Yan, Natural Environmental Research Council, UK
14:30-15:00	What Make Data Behind Healthcare Business Big and How Is Data Mining Challenged? Jian Pei, Simon Fraser University, Canada
15:00-15:15	Break
15:15-15:45	Data Mining for Astronomical Data Ali Luo, National Astronomical Observatory, Chinese Academy of Sciences, China
15:45-16:15	Data Storage and Database for Large Survey Telescope - LAMOST Yanxin Guo, National Astronomical Observatory, Chinese Academy of Sciences, China
16:15-16:45	Data Mining in Astronomy Yanxia Zhang, National Astronomical Observatory, Chinese Academy of Sciences, China
17:00-18:00	**Panel: Big data, huge data, any data, what is the most important thing behind these?** Chair: Hui Xiong (Rutgers, the State University of New Jersey, USA) Panelist: Philip S. Yu, Wei Wang(UCLA), Chengqi Zhang, Xiaoyang Wang, Zhengyuan Wang, Yangyong Zhu
18:00-20:00	**Banquet** *Address: Xin'an Hall, Pudong International TalentCity Hotel*

Day 2	**Sunday, May 26, 2013** ***Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University***
	Session 3:Big Data Query and Mining (8:30-12:15) Chair: Liang Zhang (Fudan University, China)
08:30-09:00	A Learning Approach to Ranking and Navigating SQL Query Results Zhiyuan Chen, University of Maryland Baltimore County, USA
09:00-09:30	A Proportional Association Matrix for Categorical Data Wenxue Huang, Guangzhou University, China
09:30-10:00	The Analysis and Fusion Technology for Big Data-Construction System for Knowledge Network Wei Wang, Fudan University, China
10:00-10:15	Break
10:15-10:45	Learning Hash Codes for Efficient Content Reuse Detection Xuanjing Huang,Fudan University, China
10:45-11:15	Query Result Organization with Attribute Association Weifeng Su, United International College, HK
11:15-11:45	Understanding Query Interfaces by Statistical Parsing Jiang Zhao, United International College, HK
11:45-12:15	Outlook of Dataology Yangyong Zhu, Yun Xiong and Mingmin Chi, Fudan University, China
12:30-13:30	Lunch
	End of Workshop in Shanghai

Day 3	**Monday, May 27, 2013** *Zhengzhou University*
09:00-09:15	Welcome Remarks
	Session 4: Data Mining and Its Application (9:15-12:00) Chair: Qinglei Zhou (Zhengzhou University, China)
09:15-09:45	Some Activities of Big Data Research in Australia Chengqi Zhang, Centre for Quantum Computation & Intelligent Systems, UTS, Australia
09:45-10:15	Mining Genetic Interactions in Genome-Wide Association Study Wei Wang, University of California, Los Angeles, USA
10:15-10:45	Ranking Fraud Detection for Mobile Apps: A Holistic View Hui xiong,Rutgers, the State University of New Jersey, USA
10:45-11:00	Break
11:00-11:30	Pseudo Gradient Search and Its Applications in Data Mining Zhenyuan Wang, University of Nebraska at Omaha, USA
11:30-12:00	Mining IPTV User Behaviors with a Coupled LDA Model Ya Zhang, Shanghai Jiaotong University, China
12:00-13:00	Lunch
	End of Workshop in Zhengzhou

Panel

Saturday, May 25, 2013

17:00-18:00

Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University

Title:

Big data, huge data, any data, what is the most important thing behind these?

Panel Chair:

Hui Xiong (Rutgers, the State University of New Jersey, USA)

Panelist:

Philip S. Yu (University of Illinois at Chicago, USA)

Wei Wang (University of California, Los Angeles, USA)

Chengqi Zhang (University of Technology Sydney, Australia)

Xiaoyang Wang (Fudan University, China)

Zhengyuan Wang (University of Nebraska at Omaha, USA)

Yangyong Zhu (Fudan University, China)

Talk Abstracts and Speakers Resume

Session 1: Challenge of Big Data (May 25, 2013, 8:45-12:00)

Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University

Chair: Chengqi Zhang (University of Technology Sydney, Australia)

Session 2: Scientific Research Method with Data (May 25, 2013, 13:00-17:00)

Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University

Chair: Wei Wang (Fudan University, China)

Session 3: Big Data Query and Mining (May 26, 2013, 8:30-12:15)

Lecture Hall 106, the Administration Building, Zhangjiang Compus, Fudan University

Chair: Liang Zhang (Fudan University, China)

Session 4: Data Mining and Its Application (May 27, 2013, 9:15-12:00)

Zhengzhou University

Chair: Qinglei Zhou (Zhengzhou University, China)

On Mining Big Data

Philip S.Yu
psyu@uic.edu
Computer Science Department, University of Illinois at Chicago, USA

ABSTRACT: The problem of big data has become increasingly importance in recent years. On the one hand, the big data is an asset that potentially can offer tremendous value or reward to the data owner. On the other hand, it poses tremendous challenges to realize the value out of the big data. The very nature of the big data poses challenges not only due to its volume, and velocity of being generated, but also its variety and veracity. Here variety means the data collected from various sources can have different formats from structured data to text to network/graph data to image, etc. Veracity concerns the trustworthiness of the data as the various data sources can have different reliability. In this talk, we will discuss these issues and approaches to address them.

Philip S. Yu is a Professor in the Department of Computer Science at UIC and also holds the Wexler Chair in Information and Technology. He spent most of his career at IBM Thomas J. Watson Research Center, where he was manager of the Software Tools and Techniques group. He has published more than 700 papers in refereed journals and conferences. He holds or has applied for more than 300 US patents. His main research interests include data mining (especially on graph/network mining), social network, privacy preserving data publishing, data stream, database systems, and Internet applications and technologies.

Projected Gradient Method for Sparse Principal Component Analysis

Weiguo Gao
wggao@fudan.edu.cn
Fudan Univesity, China

ABSTRACT: In this talk, we will review some formulations of sparse principal component analysis with l0 and l1 penalties. Theoretical aspects and various existing algorithms will be discussed. New relationship and equivalence are established. Then we introduce the gradient flow equation based on l0 penalty and propose a projected gradient method.

We also prove the convergence. Experiments with benchmarking data sets confirm the efficiency of our proposed algorithm.

Weiguo Gao is currently a Professor at the School of Mathematical Sciences at Fudan University and one of members of the Key Laboratory of Computational Physical Sciences, Ministry of Education. His main research interests are numerical linear algebra and high performance computing, including linear and nonlinear eigenvalue problems, large-scale science and parallel computing, numerical algebra problems in the control system, electronic structure and saddle point calculation. He has published many papers in refereed journals and conferences, such as SIAM J. Sci. Computer, IEEE Trans. Automat. Control, ACM Tran. Math. Software, Numer. Math., Internat. J. Numer. Methods Engrg., SIAM J. Matrix Anal.Appl.

Data Driven Semantic Computing

Haixun Wang
Haixun.Wang@microsoft.com
Microsoft Research Asia

ABSTRACT: Representing and reasoning over human knowledge is a computational grand challenge for the 21st century. This talk introduces Probase, a web scale, data driven, probabilistic knowledgebase developed for text processing. Probase has a large concept space that enables it to "understand" most concepts (about worldly facts) used by human beings. Furthermore, knowledge in Probase is not black and white. We develop a set of scores to quantify its uncertainty. These quantifications serve as the probabilistic priors and likelihoods that become the foundations of Probase's concept learning mechanism for understanding short text. Finally, we address one problem of knowledge-based approaches: They are mostly one-way forward propagation approaches that do not know how to adapt themselves for improving the performance on the given tasks. We bring text processing into the DNN framework by using knowledge as both an input and an output. Early results on real life search experiments show our approach is promising.

Haixun Wang is a Senior Researcher, who joined Microsoft Research Asia in Beijing, China in 2009.Before joining Microsoft, he had been a research staff member at IBM T. J. Watson Research Center for 9 years. He was Technical Assistant to Stuart Feldman (Vice President of Computer Science of IBM Research) from 2006 to 2007, and Technical Assistant to Mark Wegman (Head of Computer Science of IBM Research) from 2007 to 2009. He has published more than 120 research papers in referred international journals and conference proceedings. His research interests include data management, graph systems, data mining, knowledgebase, semantic network and text analytics.

Catch the Wind: Graph Workload Balancing on Cloud

Jeffrey Xu Yu
yu@se.cuhk.edu.hk
The Chinese University of Hong Kong, HK

ABSTRACT: Due to the large number of new applications need to deal with massive graphs, several graph database processing systems are developed. As one of the representatives, Google has developed Pregel as its internal graph processing platform where Pregel takes a vertex-centric approach and computes in a sequence of super-steps based on bulk-synchronous parallel (BSP) model.

In this talk, we will discuss graph partitioning, which is a key issue in graph database processing systems based on BSP for achieving high efficiency, on Cloud. However, the balanced graph partitioning itself is difficult because it is known to be NP-complete. We also observe that a static graph partitioning cannot make all graph algorithms efficient in parallel on Cloud, because the workload balancing in different iterations for different graph algorithms are all possible different. In this talk, we discuss graph behaviors by exploring the working window (we call it wind) changes, where a working window is a set of active vertices that a graph algorithm really needs to access in parallel computing. We investigated nine classic graph algorithms using real datasets, and propose simple yet effective policies that can achieve both high graph workload balancing and efficient partition on Cloud.

Jeffrey Xu Yuis a Professor in the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong. He served/serves in over 280 organization committees and program committees in international conferences/workshops. Dr. Yu served as an Information Director and a member in ACM SIGMOD executive committee (2007-2011), and an associate editor of IEEE Transactions on Knowledge and Data Engineering (2004-2008). His current main research interests include keywords search in relational databases, graph mining, graph query processing, and graph pattern matching. He has published over 210 papers including papers published in reputed journals and major international conferences.

Big Data-enabled Management Decision Making

Lihua Huang
lhhuang@fudan.edu.cn
Fudan University, China

ABSTRACT: Big Data is more than data gathering from various sources, it provides us with a new way to analyze problems. Specifically, Big Data enables us to predict based on correlation analysis and to have new insight on the business regular pattern, hereby enhances our understanding of new things and their rules. In Big Data era, business data in a company comes from both inside and outside of the organization, which provides a perfect opportunity for scientific management decision making. Management decision aims to improve an organization's efficiency and effectiveness. Both structured and unstructured decision making requires decision makers to judge and discern the definition of the problem, to create possible alternative solutions and make a selection, to deploy resources and implement the decision program, as well as continue to evaluate and adjust the implementation results. Researchers have pointed out that, in order to implement data-driven decisions, we not only need to use the new technology, but also to change the decision-making process. Therefore, to make management decisions enabled by Big Data, we need to make use of new technologies such as information collection, knowledge discovery, human-computer symbiotic intellectual technology and decision support technologies. More ever, we need to change the current decision process, especially to change the decision makers' cognition and behavior. This report will analyze management decision methodology and process enabled by Big Data.

Lihua Huang is currently a Professor at the School of Management at Fudan University. Her main research interests include management information system, electronic commerce, enterprise resource planning and business process reengineering. She has published many papers in refereed journals and conferences.

Big Data and Decision Making

Yong Shi
yshi@gucas.ac.cn
Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, China

ABSTRACT: Nowadays, Big Data becomes reality that no one can ignore. Big Data is our environment whenever we need to make a decision. Big Data is a buzz word that makes everyone understands how important it is. Big Data shows a big opportunity for academia, industry and government. Big Data then is a big challenge for all parties. This paper discusses some fundamental issues of Big Data problems, such as data heterogeneity vs. decision heterogeneity, data stream research and data-driven decision management. In the conclusion, the paper suggests a number of open research problems in Data Science, which is a growing field beyond Big Data.

Yong Shi currently is a full professor and executive deputy director of CAS Research Center on Fictitious Economy & Data Science. During 1991-2004, he has been the associate professor and visiting professor in the College of Information Science and Technology, University of Nebraska at Omaha. His research interests include: data mining, optimization and knowledge management. He has published 17 monographs, more than 200 papers on over 60 international journals. He is now served the chief-editors of many international journals, such as International Journal of Information Technology & Decision Making (SCI, CompuMath, ISI Alerting, Cabell), International Journal of Operations & Quantitative Management and International Journal of Business Intelligence & Data Mining, Information, International Journal of Service Science etc.

Research Issues and Challenges on Brain Informatics

Ning Zhong
zhong@maebashi-it.ac.jp
Maebashi Institute of Technology, Japan

ABSTRACT: Brain Informatics (BI) is a new interdisciplinary and multidisciplinary field that focuses on studying the mechanisms underlying the human information processing system. It brings together researchers and practitioners from diverse fields to explore the main research problems that lie in the interplay between the studies of human brain and the research of informatics, by using powerful equipment, including functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), positron emission tomography (PET), and eye-tracking. The systematic BI methodology has resulted in the big BI data, including various raw brain data, data-related information, extracted data features, found domain knowledge related to human intelligence, and so forth. In this talk, I demonstrate a systematic approach to an integrated understanding of macroscopic and microscopic level working principles of the brain by means of experimental, computational, and cognitive neuroscience studies, as well as utilizing advanced Web intelligence centric information technologies. I discuss research issues and challenges from three aspects of Brain Informatics studies that deserve closer attention: systematic investigations for complex brain science problems, new information technologies for supporting systematic brain science studies, and Brain Informatics studies based on Web intelligence research needs. These three aspects offer different ways to study traditional cognitive science, neuroscience, mental health, and artificial intelligence.

Ning Zhong is currently head of Knowledge Information Systems Laboratory, and a professor in Department of Life Science and Informatics at Maebashi Institute of Technology, Japan. He is also director and an adjunct professor in the International Web Intelligence Consortium (WIC) Institute, Beijing University of Technology, China. His research interests include Web intelligence (WI), brain informatics, knowledge discovery and data mining, rough sets and granular-soft computing, intelligent agents and databases, intelligent information systems, with more than 200 journal and conference publications, as well as more than 20 books.

Data Challenge in High Energy Physics

Gang Chen
Gang.Chen@ihep.ac.cn
Institute of High Energy Physics, Chinese Academy of Sciences, China

ABSTRACT: Large-scale scientific research plays key role in many fields of modern sciences, particularly in high energy physics (HEP), genomics, meteorology, biology, astronomy and environmental research, etc. To handle the gigantic data generated from HEP research, including acquisition, storage, sharing, analysis and visualization, are great challenges. This talk presents some examples of HEP projects and their computing models are presented. The details of data challenges are discussed. The activities of data intensive computing in China are introduced.

Gang Chen is a researcher in the Institute of High Energy Physics Chinese Academy of Science. He is currently responsible for constructing the center for the High Energy Physics Data Grid.

Data Blogging: ANew Implication for Science Impact

Weigang Yan
weig@ceh.ac.uk
Natural Environmental Research Council, UK

ABSTRACT: Data visualization integrates streams of data and transforms them into information via graphical means. Recently, data journalism combining with visualization analytics is burgeoning in the social and economic regimes. It effectively and clearly delivers the insights about the data to non-technical audients and broadcasts the social, political and commercial impacts of data. However, impacts by scientific data are under-exploited through data journalism. As we are increasingly exposed to high quality scientific data and data deluge, how to interpret such abstract data and increase the audient's understanding towards science is challenging. Elements of infographics, maps and charts are fundamentals to construct stories about scientific data. By infusing these elements, we created a data blog LULUCF datacosm based on the data produced from UK greenhouse gas inventory in land use sector as a contemporary example here. Data are analyzed, transformed and visualized as a story in the blog. Via the visualized data, we interpret and communicate complex issues covering climate change, biodiversity and carbon management and forest management. LULUCF datacosm is one of the unique digital platforms of open data focusing on data visualization and dissemination. By using this example, I will outline features and insights that help discover how data blogging creates ways to increase the impacts of science.

Weigang Yan joined NERC's Centre for Ecology & Hydrology since 2009. She is a dynamic and creative nformatics specialist with a background in environmental science and ecology. With her experience in large dataset analysis and skills in database management system, GIS and web technology, she has developed databases for various scientific projects and a web portal for NanoFATE project(www.nanofate.eu). Currently, she is developing a digital platform for the UK's greenhouse gas inventory in Land Use and Forest, as well as the novel concept of datacosm for greenhouse gas, biodiversity and land use management. She has been involved in developing and implementing data management procedure wthin the organization, provding data management solutions for science projects and data documentation and archiving. In addition, she is keen to develop Linked Open Data technologies and big data technologies for the discplines of ecology and environmental science and advocate for open data and open annotation.

What Make Data Behind Healthcare Business Big, and How Is Data Mining Challenged?

Jian Pei
jpei@cs.sfu.ca
Simon Fraser University, Canada

ABSTRACT: In this talk, I will deliberate a few important factors that lead to big data behind healthcare business with a focus on the business side, such as integration of multiple systems, legacy issues, organizations, people, culture, and historical evolvement. Using some real life examples, I will demonstrate how the existing data mining techniques are challenged. Moreover, I will discuss a few critical opportunities of data mining in healthcare business, such as data annotation and integration, validation, hypothesis generation, visualization and presentation, and business implementation.

Jian Pei is currently a Professor of Computing Science at the School of Computing Science at Simon Fraser University, Canada. Since 2000, he has published one monograph and over 140 research papers in refereed journals and conferences. His research interests can be summarized as developing effective and efficient data analysis techniques for novel data intensive applications. Particularly, he is currently interested in various techniques of data mining, Web search, information retrieval, data warehousing, online analytical processing, and database systems, as well as their applications in social networks, health-informatics, business and bioinformatics.

Data Mining for Astronomical Data

Ali Luo
lal@nao.ac.cn
National Astronomical Observatories, Chinese Academy of Sciences, China

ABSTRACT: With the development of big telescopes, large quantities of astronomical datasets are collected, which include images, spectrum, time serials data etc. Mining these datasets by using statistical analysis algorithms is an important task for astronomy. In this talk, I will introduce the astronomical data, the aim of data mining, some successful example of data mining, current status of astronomical data mining and challenges in the era of the big data.

Ali Luo is a professor at National Astronomical Observatories(NAOC), Chinese Academy of Sciences(CAS). His research interests include Astronomical data reduction and analysis, statistical analysis and automated measurement for large sample spectral data, including spectral classification and identification; redshift measurement for galaxies and QSOs; stellar parameter measurement; data mining and searching for rare object etc. Other research interests are including automatically observation of telescopes, astronomical software development, and integration of computer and network.

Data Storage and Database for Large Survey Telescope - LAMOST

Yanxin Guo
xin305@163.com
National Astronomical Observatories, Chinese Academy of Sciences, China

ABSTRACT: LAMOST, Large Sky Area Multi-Object Fibre Spectroscopy Telescope, as the world's highest spectral acquisition rate telescope, will get tens of thousands of spectral data each night. It has already designed complete set of automated observing and processing systems which has been running through the pilot and general survey. Off-line data processing system backed by kernel databases is introduced here, including data storage solution and dataflow, as well as spectral quality control workflow.

Data Mining in Astronomy

Yanxia Zhang
zyx@lamost.org
National Astronomical Observatories, Chinese Academy of Sciences, China

ABSTRACT: With the construction and development of ground- and space-based observatories, astronomical data amount to Terascale, even Petascale. Facing data avalanche in astronomy, knowledge discovery in databases (KDD) shows its superiority. How to extract knowledge from so massive data volume by automated methods is a big challenge for astronomers. Under this situation, many researchers have studied various approaches and developed different software to solve this issue. According to the special task of data mining, we need to select an appropriate technique suiting the requirement of data characteristics. Moreover all algorithms have their own pros and cons. We introduce the characteristics of astronomical data, present knowledge discovery issues and challenges faced in astronomy and the usually adopted methods of knowledge discovery in astronomy are touched upon. Finally the successful applications of data mining techniques in astronomy are summarized.

A Learning Approach to Ranking and Navigating SQL Query Results

Zhiyuan Chen
zhchen@umbc.edu
Department of Information Systems, University of Maryland, Baltimore County, USA

ABSTRACT: Users often find that their queries against a large database return too many answers, many of them irrelevant. This problem has become more common as both the sizes of databases as well as the number of online databases grow dramatically. Two solutions have been proposed to address this problem: ranking and navigation. However, the success of both approaches largely depends on capturing users' preferences. Existing work either assumes that every user has the same preference or a user's current preference is the same as his or her previous preference. However in practice users' preferences are often diverse, heterogeneous (some preferences are static and some are dynamic), and incomplete.

In this talk we will introduce a learning approach that addresses these issues in ranking and navigating SQL query results. For the problem of navigating SQL results, we propose a two-step approach [1]: 1) it first analyzes query history of all users in the system offline and generates a set of clusters over the data, each corresponding to one type of user preferences; 2) when user asks a query, the second step presents to the user a navigational tree over clusters generated in the first step such that the user can easily select the subset of clusters matching his needs.

For the problem of ranking SQL results, we propose a local ranking method [2]: 1) it addresses the diverse and heterogeneous issue by using skyline to capture users' static and common preferences and using users' current navigational behavior to capture users' dynamic and diverse preferences; 2) it addresses the incompleteness issue by using machine learning technique to learn a ranking function based on training examples constructed from the above two types of information. This method also builds upon the first method by applying ranking over clusters selected by users.

We have conducted user studies as well as simulations to compare proposed methods with existing methods and the results have verified the effectiveness of our methods.

Zhiyuan Chen is an Associate Professor in the Information Systems Department at the University of Maryland Baltimore County, USA. His research interests include privacy preserving data mining and data management, data exploration and navigation, and semantic-based search and data integration using semantic networks.

[1] Z. Chen and T. Li, "Addressing diverse user preferences in SQL-query-result navigation," presented at the SIGMOD, 2007, pp. 641–652.

[2] Z. Chen, T. Li, and Y. Sun, "A Learning Approach to SQL Query Results Ranking Using Skyline and Users' Current Navigational Behavior," *IEEE Trans. Knowl. Data Eng.*, In Press, 2012.

A Proportional Association Matrix for Categorical Data

Wenxue Huang
whuang123@yahoo.com
Guangzhou University, China
(Based on joint work with Yong Shi and Xiaogang Wang)

ABSTRACT: We introduce an intrinsic and informative local-to-global association matrix to measure the association of categories of a variable with another categorical variable. Towards a proportional prediction or a probabilistic averaging effect, the association matrix determines the expected confusion matrix of a multinomial response variable. The normalization of the diagonal of the matrix gives rise to an association vector, which provides the expected category accuracy lift rate distribution. A general scheme of global-to-global association measures associated with flexible weight vectors is derived from the diagonal. A hierarchy of equivalence relations defined by the association matrix and vector is shown. Applications to financial and survey data together with simulations results are presented. Extended Abstract: Nominal data are quite common in scientific and engineering research related to biomedical research, consumer behavior analysis, net-work analysis and search engine marketing optimization. When the population is cross-classified and there is no natural ordering for observed outcomes, association analysis can be described as nominal association measures. Even if categorical variables collected in these studies are ordinal, they are often treated as nominal if the ordering is not of interest, especially when the conditional distribution or expectation is of central concern. When the response variable is multinomial, the principle of optimal (condition-al mode based) or proportional (conditional Monte-Carlo based) prediction can be used to construct nonparametric nominal association measures. For example, Goodman and Kruskal in 1954, 1996, 2000 and others proposed some local-to-global association measures towards optimal predictions. Both the Monte Carlo and finite Markov chain methods are logically or conceptually based on the proportional associations. These methods are broadly used in asset pricing practice, financial risk management, inventory management and network analysis. The proportional associations between variables are probabilistically and statistically intrinsic. It reflects the probabilistically averaging effects of input on output distributions.

There are quite a few proportional association measures proposed in the literature. However, all these measures focus either on local-to-local (e. g. , contingent table analysis) or on global-to-global associations between two categorical variables. For some statistical inferences, commonly used local-to-local (category-to-category) or global-to-global (variable-to-variable) association measures can be sufficiently informative. When the population is cross-classified, effective local-to-global (category-to-variable) association measures are necessary to provide a more detailed description or evaluation of the intrinsic dependence structure. For example, these kinds of measures are of fundamental importance in targeted or group specific risk analysis such as credit risk migration analysis, clinical diagnosis, or public health management. Pattern recognition and feature selections are very

important in real-world 1applications. Although some very effective methods have been developed, there exist many challenging problems especially for handling categorical data. In order to gain important insights of the underlying dependence structure, we propose a nominal association matrix to measure the probabilistic proportional associations of every category of a response variable with an explanatory variable. We will show that this matrix estimates expected confusion matrix for multinomial response variables when a proportional (conditional Monte-Carlo) prediction is employed. The diagonal of the matrix also induces our association vector introduced in this article. Each component of the association vector represents the proportional dependence degree of a given category of a response variable with a given explanatory variable. Furthermore, combined with various sets of admissible weight vectors, the diagonal of the propose matrix provides a general scheme of generating different global-to-global association measures which embraces the seminal Goodman-Kruskal _ (the GK-_). The weighting scheme in the general class of association measure is now explicit when compared with that of the GK-_. Furthermore, one can now design various association measures according to different objectives. Consequently, scientists are no longer constrained by the limited number of association measures in the literature that might or might not be optimal for their inferential needs. For example, association measures designed to capture rare events should be different from those aimed at describing the entire population. More importantly, any prior knowledge or relevant expert opinions can be formally incorporated into this framework. A hierarchy of equivalence relations induced by the association matrix and vectors are also presented to help understand the strengths of proposed association measures. This hierarchy is of fundamental interest to the structural analysis for multivariate categorical data. We also show that the association vector, the association matrix and any global association measure determined by a weight vector, particularly the GK-_, are equivalent only when the response variable is binary. For high dimensional data, the proposed global association measures are also capable of evaluating the corresponding collective effect of a set of explanatory variables on a given response variable in order to remove redundancy. We prove that a dimensionality reduction for high dimensional categorical data with a response variable is theoretically possible by using the proposed association measures.

Wenxue Huang is a Full Professor in the School of Mathematics and Information Sciences at Guangzhou University, and an adjunct professor at both York University, Canada and Nanchang University. He received his Ph.D. in Mathematics at the University of Western Ontario in 1995. Before joining Guangzhou University, he had served as a full professor at Shantou University. He has over 11 years of data mining research and development experience in industry and served as the Chief Scientist at Generation 5 Math Tech. Inc. for around 10 years.

He is the author or co-author of 20 plus peer-reviewed publications in the areas of algebraic groups and monoids, associative algebras, data mining, and differential geometry and topology. He has delivered more than 30 invited talks on algebra and data mining at international academic conferences and at Canadian and Chinese research institutes and universities.

The Analysis and Fusion Technology for Big Data-Construction System for Knowledge Network

Wei Wang
Weiwang1@fudan.edu.cn
Fudan University, China

ABSTRACT: Knowledge Network is a new technology used in many areas, such as search engine, content management, semantic web, etc. The foundation the knowledge network is to extract the concepts/entities and the relationships between them from the massive WEB documents and the documents from specific area. The construct system of knowledge network system will deal with massive WEB data. The data changes every day. The data from different source have different quality and format. So it is a typical big data analysis and fusion problem.

In this present, we will describe the problem of big data analysis and fusion and their solution from the point view of the construct system from knowledge network. The detail is:

- The data fusion technology for heterogeneous information resources. In our system, the data is described with network. It can represent the relationship between the data. For the fusion technique, we will consider more about the confidence and coverage of the data, which are the nature feature of the data set. Compare with tradition data fusion approach, the data sources in our system will be more complex and diverse.
- Stream based system architecture. Different to tradition data mining and machine learning approach, the knowledge graph is built on many data set not just one data set. We proposed an incremental learning approach and an evidence base confidence evaluation approach.
- Big graph management. A large knowledge network system is a large graph with more than 1010 nodes and edges. So it need a scalable platform to manage and analysis these data.

Wei Wang is currently a professor at the School of Computer Science at Fudan University. His research interests include database system, data mining, and WEB data process. His research has been supported by NSFC, The Ministry of Science and Technology of the P.R.China, Science and Technology Commission of Shanghai Municipality, SAP, EMC. He has published more than 70 papers in refereed journals and conferences, including the major database conference and journals, such as SIGKDD, SIGMOD, VLDB, WWW, ICDE, IEEE TKDE, JIIS. He has served in the organization committees and the program committees of many international conferences, such as ICDM, ICDE, SIAM DM, WAIM.

Learning Hash Codes for Efficient Content Reuse Detection

Xuanjing Huang
Xuanjing.huang@gmail.com
Fudan University, China

ABSTRACT: There is a quick expansion in the popularity of user generated content in forums, microblogging sites, blogs, and other medius in recent years. While the increasing of UGC, content reuse, which is the practice of using existing content components, occurs frequently in these mediums. It contains various forms including duplicate, near-duplicate, and partial-duplicate. Since, content reuse is extremely common in user generated mediums, reuse detection serves as be the basis for many applications. However, along with the explosion of Internet and continuously growing uses of user generated mediums, the task becomes more critical and difficult. In this talk, we present a novel efficient and scalable approach to detect content reuse. We investigate a novel approach to detect sentence level content reuse by mapping sentence to a signature space. Signature of a sentence is created by taking the bitwise-or of all signatures of words occurs in the sentence. Rather than using traditional hash functions, which do not consider statistics of words or characters, to assign hash code for each word/character, we analyze the requirements of what the good codes should satisfy and formalize it as a constraint optimization problem. In order to deal with tens of billions of documents, we implement the detection approach on graphical processing units (GPUs). The experimental comparison in this paper involves studies of efficiency and effectiveness of the proposed approach in different types of document collections, including ClueWeb09, Tweets2011, and so on. Experimental results show that the proposed approach can achieve the same detection rates with state-of-the-art systems while uses significantly less execution time than them (from 400X to 1500X speedup).

Xuanjing Huang is a Professor and the deputy director of the Laboratory of Intelligent Media Computing now. Her research interests include text retrieval, text filtering, natural language processing, and digital library. She has published some papers in several major conferences including ICML, SIGIR, ACL, CIKM, ISWC, EMNLP, IJCNLP and AIRS.

Query Result Organization with Attribute Association

Weifeng Su
wfsu@uic.edu.hk
United International College, HK

ABSTRACT: QROAA (Query Result Organization with Attribute Association) is a novel approach for addressing the problem of too many records returned from a database in response to a user query. QROAA effectively clusters the query result records using two components: a preprocessing component and a query processing component. In the preprocessing component, several parameters used for clustering the query results, including attribute importance, attribute value distance, and attribute variance and covariance, are calculated according to the workload and the records in the database. In the query processing component, a clustering algorithm is performed on the query results according to the parameters. Thereafter, a record is selected out of each cluster to represent the cluster and the representative records are ranked according to the size of their corresponding clusters. Furthermore, each cluster is equipped with a concise description, which helps the user to browse its content.

Compared with existing work, our approach has the following advantages. First, attribute association is fully considered to prevent overestimating the importance of some attributes. Second, the record difference is fully counted so that diverse records are presented at the top of the recommendation. Finally, a concise description is provided for each cluster that holds similar records. Preliminary experiments show that QROAA effectively captures the user's preferences and provides a reasonable and intuitive organization of the query result records.

Weifeng Su is an associate professor at the Computer Science Program, BNU-HKBU United International College (UIC). He has published over 10 papers in the most important journals and conference proceedings, including ACM TRANSACTION ON DATABASE SYSTEM (TODS), IEEE Transaction on Knowledge and Data Engineering (TKDE), his papers has been cited over 200 times. His research interests include deep Web, data mining, machine learning, word sense disambiguation, and natural language processing.

Understanding Query Interfaces by Statistical Parsing

Jing Zhao
jzhao@uic.edu.hk
United International College, HK

ABSTRACT: Users submit queries to an online database via its query interface. Query interface parsing, which is important for many applications, understands the query capabilities of a query interface. Since most query interfaces are organized hierarchically, we present a novel query interface parsing method, StatParser (Statistical Parser), to automatically extract the hierarchical query capabilities of query interfaces. StatParser automatically learns from a set of parse d query interfaces and parses new query interfaces. StatParser starts from a small grammar and enhances the grammar with a set of probabilities learned from parsed query interfaces under the maximum-entropy principle. Given a new query interface, the Probability-enhanced-grammar identifies the parse tree with the largest global probability to be the query capabilities of the query interface. Experimental results show that StatParser very accurately extracts the query capabilities and can effectively overcome the problems of existing query interface parsers.

Jing Zhao is an assistant professor at the Computer Science Department, United Inernational College (UIC). Her research interests include web information discovery, management and integration, especially in a distributed environment.

Outlook of Dataology and Data Science

Yangyong Zhu, Yun Xiong, and Mingmin Chi
{yyzhu, yunx,mmchi}@fudan.edu.cn
Fudan University, China

ABSTRACT: Nowadays, almost all disciplines generate huge amount of data for the applications such as business analytics, healthcare, hazard monitoring. This leads to the data deluge. Even worse, those data come from interdisciplinary sources with fast velocity such that no existing theory or method can be utilized for processing the big data. Therefore, it is highly urgent to boost research and education in the new theories and methods for the emerging discipline, i.e., Dataology or interchangeable name Data Science.

Fortunately, Dataology and Data Science have attracted more and more attention from both the academic and industrial areas in recent years. Different kinds of data science research centers or institutes have been built in the universities over the world, such as China, USA, Australia, Japan, and Poland. Meanwhile, new international conferences related to Data Science are emerging per year, which are organized by peoples from different research areas from universities, research institutes, and industries. Furthermore, "the sexiest job" of the 21st century - data scientist teams have been built up by almost all of data gigants to strength their competition in the market for data and analytics.

Unfortunately, Dataology and Data Science is still in its infancy. So far, although there is no unified name and definition for data science, researchers over the world realized the importance of data and analytics. We believe that the rapid progress in Dataology and Data Science will be seen in the next few years, including the gradual development of: (1) the fundamental theories; (2) data-driven scientific research methods; (3) disciplinary construction with the courses for Dataology. In the talk, we will discuss key issues on Dataology and Data Science, including the new definition, the fundamental theories, the research methods, and the disciplinary construction of Dataology and Data Science.

Yangyong Zhu is a Professor of Computer Science atFudan University and the director of Dataology and DataScience Research Center. His research interests include Dataology and Data Science and big data analytics. He has published two monographs and 100+ research papers in refereed journals and conferences.

Yun Xiong is an Associate Professor of Computer Science atFudan University, Shanghai, China. Her research interests include Dataology and Data Science and big data analysis. She has published two monographs and 20+ research papers in refereed journals and conferences.

Mingmin Chi is an Associate Professor in the School of Computer Science at FudanUniversity, Shanghai, China. Dr. Chi has published 20+ papers in refereed journals and conferences (such as ICML) with 560+ citations. Her research interests include machine learning and data mining especially on financial analysis, network analysis, image analysis, text mining, and data stream mining.

Some Activities of Big Data Research in Australia

Chengqi Zhang
Chengqi.Zhang@uts.edu.au
Centre for Quantum Computation & Intelligent Systems, UTS, Australia

ABSTRACT: Big Data Research is a hot topic in recent years world-wide. It attracted many people from universities, research organizations, and government agencies to devote their time and energy to this area. This talk will briefly describe the activities, events, research centres, and future plans about Big Data Research in Australia.

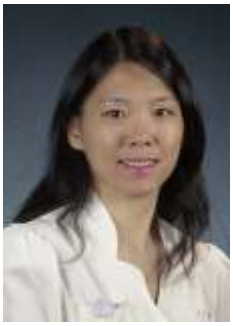
Chengqi Zhang has been a Research Professor in Information Technology at The University of Technology, Sydney (UTS) since December 2001. He has been the Director of the UTS Priority Investment Research Centre for Quantum Computation and Intelligent Systems. He has published more than 200 refereed research papers, including several in first-class international journals, such as Artificial Intelligence (he is the first author who published a paper in this world renowned Journal in distributed Artificial Intelligence research area world-wide in 1992), IEEE and ACM Transactions. He has published six monographs and edited 16 books. His research interests mainly focus on Distributed artificial intelligence, data mining and its applications.

Mining Genetic Interactions in Genome-Wide Association Study

Wei Wang
weiwang@cs.unc.edu
University of California, Los Angeles, USA

ABSTRACT: Advanced biotechnologies have rendered feasible high-throughput data collecting in human and other model organisms. The availability of such data holds promise for dissecting complex biological processes. Making sense of the flood of biological data poses great statistical and computational challenges.

I will discuss the problem of mining gene-gene interactions in high-throughput genetic data. Finding genetic interactions is an important biological problem since many common diseases are caused by joint effects of genes. Previously, it was considered intractable to find genetic interactions in the whole-genome scale due to the enormous search space. The problem was commonly addressed using heuristics which do not guarantee the optimality of the solution. I will show that by utilizing the upper bound of the test statistic and effectively indexing the data, we can dramatically prune the search space and reduce computational burden. Moreover, our algorithms guarantee to find the optimal solution. In addition to handling specific statistical tests, our algorithms can be applied to a wide range of study types by utilizing convexity, a common property of many commonly used statistics.



Wei Wang is a professor in the Department of Computer Science at University of California at Los Angeles. Her research interests include data mining, bioinformatics and computational biology, and databases. She has filed 7 patents, and has published one monograph and more than 100 research papers in international journals and major peer-reviewed conference proceedings.

Ranking Fraud Detection for Mobile Apps: A Holistic View

Hui Xiong
xionghui@gmail.com
Rutgers, the State University of New Jersey, USA

ABSTRACT: Ranking fraud in the mobile App market refers to fraudulent or deceptive activities which have a purpose of bumping up the Apps in the popularity list. Indeed, it becomes more and more frequent for App develops to use shady means, such as inflating their App's sales or posting phony App ratings, to commit ranking fraud. While the importance of preventing ranking fraud has been widely recognized, there is limited understanding and research in this area. To this end, in this talk, we introduce a holistic view of ranking fraud and propose a ranking fraud detection system for mobile Apps. Specifically, we investigate two types of evidences, ranking based evidences and rating based evidences, by modeling Apps' ranking and rating behaviors through statistical hypotheses tests. In addition, we propose an optimization based aggregation method to integrate all the evidences for fraud detection. Finally, we evaluate the proposed system with real-world App data collected from the iOS App Store for a long time period. In the experiments, we validate the effectiveness of the proposed system, and show the scalability of the detection algorithm as well as some regularity of ranking fraud activities.

Hui Xiong is currently an Associate Professor in the Management Science and Information Systems Departmentat Rutgers, the State University of New Jersey, USA. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. His research has been supported in part by the National Science Foundation (NSF), IBM Research, SAP Corporation, Panasonic USA Inc., and Rutgers, the State University of New Jersey. He has published prolifically in refereed journals and conference proceedings (3 books, 20+ journal papers, and 40+ conference papers), such as JOC, TKDE, VLDBJ, JDMKD, KDD, CCS, etc.

Pseudo Gradient Search and Its Applications in Data Mining

[1]Zhenyuan Wang and [2]Bo Guo
[1]Department of Mathematics, University of Nebraska at Omaha, USA
[2]College of Information Science and Technology, University of Nebraska at Omaha, USA

ABSTRACT: The pseudo gradient search [1] is one of effective soft computing techniques that can be used for solving optimization problems numerically. It only requires that the objective function is continuous with respect to the unknown parameters. At a given initial point in the search space, through a multi-direction learning procedure, we may determine a pseudo gradient at this point based on the required precision for the numerical solution and the given data set. As a special case when the gradient of the objective function exists at this point, the find pseudo gradient almost coincides with the gradient. Then, a one-dimensional search along the pseudo gradient direction is performed to find a new point that is much "better" than the initial point in the search space and is used to replace the original initial point. In this search procedure, a so-called "forward-double-backward-half" strategy is adopted to guarantee the convergence of the search. Repeating this procedure for several times, a satisfactory approximate solution can usually be found for the given optimization problem.

The pseudo gradient search is a local search. Similar to the gradient search, its search speed is high. However, like the other local search methods, it can only find a local extremer (or a saddle point, a singular point) for the objective function. That is, there exists a risk of "falling" in some local extremer but not in the global extremer. So, usually, it should be used with some global search method, such as a genetic algorithm, together. For example, a genetic algorithm is used first and then the pseudo gradient search is followed once a premature occurs in the procedure of the genetic algorithm [1].

Zhenyuan Wang is a Professor in the Department of Mathematics at the University of Nebraska at Omaha. His main research interests include nonadditive measure and nonlinear integration, probability and statistics, nonlinear optimization, soft computing, and applications to information fusion and data mining.

[1] Z. Wang, R. Yang, and K. S. Leung, *Nonlinear Integrals and Their Applications in Data Mining*, World Scientific, Singapore, 2010.

Mining IPTV User Behaviors with a Coupled LDA Model

Ya Zhang
ya_zhang@sjtu.edu.cn
Shanghai Jiaotng University, China

ABSTRACT: Cloud computing as a new technology in the IT industry has undergone a major change, the smart grid informatization construction cannot leave a cloud computing platform. For this reason, gives the power of cloud computing platform architecture. Electric cloud application environment of wide distribution, large amount of data, heterogeneous, real-time characteristics on power cloud computing platform safety put forward higher requirements, so the design of the electric cloud computing platform security measures, to ensure the safe, reliable operation of the smart grid.

Ya Zhang is a Researcher at the School of Electronic Information and Electrical Engineering at the Shanghai Jiao Tong University. Her research interests include information retrieval, data mining, machine learning and pattern recognition. In the recent years, she has published more than 30 papers, and her papers have been cited over 500 times, has applied for 5 US patents (holds 2 patents).

Research Center for Dataology and DataScience (RCDD) was founded by Prof. Yangyong Zhu at Fudan University, Shanghai, China. It is renamed from “Data Mining Group (DMGroup)”, one of the earliest data mining research teams in China (constructed in 1999). Both are directed by Prof. Yangyong Zhu.

The RCDD aims to study the theories, methods and technologies of Dataology and Data Sciencewith interdisciplinary applications such as finance, insurance, stock market, intelligent transportation, biomedical area.The center seeks to forge closer relationships between researchers over the world, to attract interdisciplinary graduate students interested in problems relating to Dataology and Data Science, and to build strong and mutually beneficial relationships with industry partners. The center seeks to attract external funding from both governmental and industrial sources to support its research and educational mission.

For more information, please visit:
http://www.datascience.fudan.edu.cn/
http://www.dataology.fudan.edu.cn/