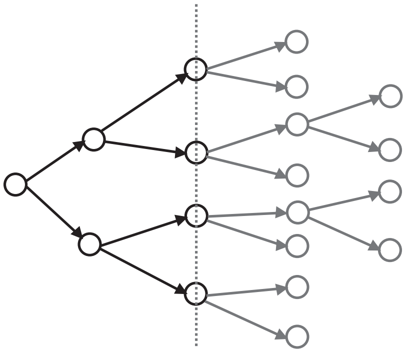
Final Program and Abstracts

2012 SIAM
International Conference
on DATA MINING

April 26-28, 2012



Disney's Paradise Pier Hotel
Anaheim, California, USA

Sponsored by the SIAM Activity Group
on Data Mining and Analytics (SIAG/DMA)

The purpose of the SIAM Activity Group on Data Mining and Analytics (SIAG/DMA) is to advance the mathematics of data mining, to highlight the importance and benefits of the application of data mining, and to identify and explore the connections between data mining and other applied sciences.

This conference is held in cooperation with the American Statistical Association.

Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 USA
Telephone: +1-215-382-9800 Fax: +1-215-386-7999
Conference E-mail: meetings@siam.org
Conference Web: www.siam.org/meetings/
Membership and Customer Service: (800) 447-7426 (US & Canada) or +1-215-382-9800 (worldwide)

www.siam.org/meetings/sdm12

Table of Contents

Organizing Committee

Steering Committee Chair

Chandrika Kamath
Lawrence Livermore National Laboratory, USA

Conference Chairs

Joydeep Ghosh
University of Texas - Austin, USA

Huan Liu
Arizona State University, USA

Program Chairs

Ian Davidson
University of California - Davis, USA

Carlotta Domeniconi
George Mason University, USA

Workshop Chairs

Rui Kuang
University of Minnesota, USA

Chandan Reddy
Wayne State University, USA

Tutorial Chair

Nitesh Chawla
University of Notre Dame, USA

Publicity Chairs

Nitin Agarwal
University of Arkansas at Little Rock, USA

Sponsorship Chairs

Jennifer Dy
Northeastern University, USA

Jimeng Sun
IBM T.J. Watson Research Center, USA

Local Chair

Yan Liu
University of Southern California, USA

Doctoral Forum Organizers

Jennifer Dy
Northeastern University, USA

Yan Liu
University of Southern California, USA

Jimeng Sun
IBM T.J. Watson Research Center, USA

Senior Program Committee Members

Charu Aggarwal
IBM T. J. Watson Research Center, USA

Arindam Banerjee
University of Minnesota, USA

Sugato Basu
Google Research, USA

Roberto Bayardo
Google Research, USA

Francesco Bonchi,
Yahoo! Research Barcelona, Spain

Varun Chandola
Oak Ridge National Laboratory, USA

Wei Fan
IBM T. J. Watson Research Center and IBM CRL, USA

Dimitrios Gunopulos
University of Athens, Greece

Jiawei Han
University of Illinois at Urbana-Champaign, USA

George Karypis
University of Minnesota, USA

Eamonn Keogh
University of California – Riverside, USA

Tamara Kolda
Sandia National Laboratories, USA

Jure Leskovec
Stanford University, USA

Olfa Nasraoui
University of Louisville, USA

Alexander Smola
Yahoo! Research, USA

Myra Spiliopoulou
Otto-von-Guericke-Universitaet Magdeburg, Germany

Parthasarathy Srinivasan
Ohio State University,USA

Hannu Toivonen
University of Helsinki, Finland

Wei Wang
University of North Carolina at Chapel Hill, USA

Qiang Yang
Hong Kong University of Technology, Hong Kong

Jieping Ye
Arizona State University, USA

Bianca Zadrozny
IBM T.J. Watson Research Center, USA

Mohammed Zaki
Rensselaer Polytechnic Institute, USA

Program Committee Members

Naoki Abe
IBM Research, USA

Foto Afrati
National Technical University of Athens, Greece

Nitin Agarwal
University of Arkansas at Little Rock, USA

Mohammad Al Hasan
Indiana University-Purdue University, USA

Aijun An
York University, Canada

Tony Bagnall
University of East Anglia, United Kingdom

James Bailey
University of Melbourne, Australia

Jose Luis Balcazar
Universidad de Cantabria, Spain

Daniel Barbara
George Mason University, USA

Gustavo Batista
University of California at Riverside, USA

Tanya Berger-Wolf
University of Illinois at Chicago, USA

Kanishka Bhaduri
NASA Ames Research Center, USA

Mustafa Bilgic
Illinois Institute of Technology, USA

Jean-Francois Boulicaut
Universite de Lyon, France

Wray Buntine
Canberra Research Laboratory, NICTA, Australia

Toon Calders
Eindhoven University of Technology, The Netherlands

Michelangelo Ceci
University of Bari "Aldo Moro", Italy

Nicolo Cesa-Bianchi
University of Milan, Italy

Vineet Chaoji
Yahoo! Labs, Bangalore, India

Nitesh Chawla
University of Notre Dame, USA

Ling Chen
University of Technology, Sydney, Australia

Hong Cheng
Chinese University of Hong Kong, Hong Kong

Diane Cook
Washington State University, USA

Alfredo Cuzzocrea
Italian National Research Council, Italy

Florence D'Alche-Buc
Universite d'Evry Val d'Essonne, France

Kamalika Das
NASA Ames Research Center, USA

Christian Desrosiers
Ecole de technologie superieure, Montreal, Canada

Chris Ding
University of Texas at Arlington, USA

Wei Ding
University of Massachusetts Boston, USA

Daniel Dunlavy
Sandia National Laboratories, USA

Haimonti Dutta
Columbia University, USA

Saso Dzeroski
Jozef Stefan Institute, Slovenia

Tina Eliassi-Rad
Rutgers University, USA

Ya Ju Fan
Lawrence Livermore National Laboratory, USA

Yi Fang
Purdue University, USA

Xiaoli Fern
Oregon State University, USA

Maurizio Filippone
University of Glasgow, United Kingdom

George Forman
Hewlett-Packard Labs, USA

Ana Fred
Technical University of Lisbon, Portugal

Johannes Furnkranz
Technical University Darmstadt, Germany

Joao Gama
University of Porto, Portugal

Byron Gao
Texas State University, USA

Jing Gao
SUNY Buffalo, USA

Claudio Gentile
University of Insubria, Italy

Amol Ghoting
IBM T. J. Watson Research Center, USA

Fosca Giannotti
ISTI-CNR, Italy

Aris Gkoulalas-Divanis
IBM Research, Zurich, Switzerland

David Gleich
Purdue University, USA

Bart Goethals
University of Antwerp, Belgium

Francesco Gullo
Yahoo! Barcelona, Spain

Tias Guns
Katholieke Universiteit Leuven, Belgium

Maria Halkidi
University of Pireaus, Greece

Shen-Shyang Ho
University of Maryland College Park, USA

Vasant Honavar
Iowa State University, USA

Ming Hua
Facebook, USA

Jun Huan
University of Kansas, USA

Eyke Huellermeier
University of Marburg, Germany

Shuiwang Ji
Old Dominion University, USA

Rouoming Jin
Kent State University, USA

Alipio Jorge
University of Porto, Portugal

Alexandros Kalousis
University of Geneva, Switzerland

Kristian Kersting
Fraunhofer IAIS, University of Bonn, Germany

George Kollios
Boston University, USA

Vipin Kumar
University of Minnesota, USA

Terran Lane
University of New Mexico, USA

Tao Li
Florida International University, USA

Jessica Lin
George Mason University, USA

Jinze Liu
University of Kentucky, USA

Jun Liu
Siemens Corporate Research, USA

Nathan Nan Liu
Hong Kong University of Science and Technology, Hong Kong

Stefano Lonardi
University of California at Riverside, USA

Chang-Tien Lu
Virginia Polytechnic Institute and State University, USA

Claudio Lucchese
Institute of the National Research Council, Italy

Sheng Ma
DoubleClick, USA

Donato Malerba
University of Bari, Italy

Rosa Meo
University of Torino, Italy

Fabian Moerchen
Siemens, USA

Alessandro Moschitti
University of Trento, Italy

Emmanuel Muller
Karlsruhe Institute of Technology, Germany

Art Munson
Sandia National Laboratories, USA

Mirco Nanni
ISTI CNR, Italy

Tim Oates
University of Maryland, Baltimore, USA

Zoran Obradovic
Temple University, USA

Nikunj Oza
NASA Ames Research Center, USA

Themis Palpanas
University of Trento, Italy

Sinno Jialin Pan
Institute for Infocomm Research, Singapore

Spiros Papadimitriou
Google Research, USA

Stelios Paparizos
Microsoft, USA

Rajesh Parekh
Yahoo! Research, USA

Dino Pedreschi
University of Pisa, Italy

Jian Pei
Simon Fraser University, Canada

Jing Peng
Montclair State University, USA

Ruggero Pensa
Consiglio Nazionale delle Ricerche, Italy

Ali Pinar
Sandia National Laboratories, USA

Kunal Punera
Yahoo! Research, USA

Huzefa Rangwala
George Mason University, USA

S.S. Ravi
Suny Albany, USA

Chandan K. Reddy
Wayne State University, USA

Saeed Salem
North Dakota State University, USA

Guna Seetharaman
Air Force Research Laboratory, USA

Arno Siebes
Universiteit Utrecht, The Netherlands

Michael Steinbach
University of Minnesota, USA

Masashi Sugiyama
Tokyo Institute of Technology, Japan

Einoshin Suzuki
Kyushu University, Japan

Andrea Tagarelli
University of Calabria, Italy

Evimaria Terzi
Boston University, USA

Volker Tresp
Siemens, Germany

Duygu Ucar
Stanford University, USA

Antti Ukkonen
Yahoo! Research Barcelona, Spain

Giorgio Valentini
University of Milan, Italy

Matthijs van Leeuwen
Universiteit Utrecht, The Netherlands

Michalis Vazirgiannis
Athens University of Economics & Business, Greece

Ricardo Vilalta
University of Houston, USA

Michail Vlachos
IBM Research Zurich, Switzerland

Slobodan Vucetic
Temple University, USA

Fei Wang
IBM Research, USA

Jianyong Wang
Tsinghua University, Beijing, China

Pu Wang
StumbleUpon, USA

Tony Wirth
University of Melbourne, Australia

Xindong Wu
University of Vermont, USA

Hui Xiong
Rutgers University, USA

Dit-Yan Yeung
Hong Kong University of Science and Technology, Hong Kong

Philip Yu
University of Illinois at Chicago, USA

Kun Zhang
Xavier University of Lousiana, USA

Xiang Zhang
Case Western Reserve University, USA

Ying Zhao
Tsinghua University, Beijing China

Hill Zhu
University of Technology Sydney, Australia

SIAM Registration Desk

The SIAM registration desk is located in the Pacific Ballroom Foyer – 1st Floor. It is open during the following times:

Wednesday, April 25
5:00 PM – 7:00 PM

Thursday, April 26
7:00 AM – 7:30 PM

Friday, April 27
7:30 AM – 3:30 PM

Saturday, April 28
7:15 AM – 4:00 PM

Hotel Address

Disney's Paradise Pier Hotel
1717 S. Disneyland Drive
Anaheim, CA 92802
Direct Telephone: +1-714-999-0990
Direct Reservations Number:
+714-520-5005
Hotel Fax: +1-714-520-7097
http://disneyland.disney.go.com/paradise-pier-hotel/

Hotel Telephone Number

To reach an attendee or to leave a message, call +1-714-999-0990. The hotel operator can either connect you with the SIAM registration desk or to the attendee's room. Messages taken at the SIAM registration desk will be posted to the message board located in the registration area.

Hotel Check-in and Check-out Times

Check-in time is 3:00 PM and check-out time is 11:00 PM.

Child Care

The Grand Californian Hotel & Spa offers Pinocchio's Workshop for guests of the Resort. Available from 5:00pm – 12:00am every night for children ages 5 – 12, with a cost of $13.00 per hour per child with a 2 hour minimum. Reservations are not required – but highly suggested. Reservations can be made at any of the three hotels Guest Services Desks.

Another childcare option is Baby Sitters Unlimited. Contact Lisa Bramlett at 951-780-7100 for pricing and a description of what they offer.

Corporate Members and Affiliates

SIAM corporate members provide their employees with knowledge about, access to, and contacts in the applied mathematics and computational sciences community through their membership benefits. Corporate membership is more than just a bundle of tangible products and services; it is an expression of support for SIAM and its programs. SIAM is pleased to acknowledge its corporate members and sponsors. In recognition of their support, non-member attendees who are employed by the following organizations are entitled to the SIAM member registration rate.

Corporate Institutional Members

The Aerospace Corporation

Air Force Office of Scientific Research

AT&T Laboratories - Research

Bechtel Marine Propulsion Laboratory

The Boeing Company

CEA/DAM

DSTO- Defence Science and Technology Organisation

Hewlett-Packard

IBM Corporation

IDA Center for Communications Research, La Jolla

IDA Center for Communications Research, Princeton

Institute for Defense Analyses, Center for Computing Sciences

Lawrence Berkeley National Laboratory

Lockheed Martin

Mathematical Sciences Research Institute

Max-Planck-Institute for Dynamics of Complex Technical Systems

Mentor Graphics

The MITRE Corporation

National Institute of Standards and Technology (NIST)

National Security Agency (DIRNSA)

Naval Surface Warfare Center, Dahlgren Division

NEC Laboratories America, Inc.

Oak Ridge National Laboratory, managed by UT-Battelle for the Department of Energy

PETROLEO BRASILEIRO S.A. – PETROBRAS

Philips Research

Sandia National Laboratories

Schlumberger-Doll Research

Tech X Corporation

Texas Instruments Incorporated

U.S. Army Corps of Engineers, Engineer Research and Development Center

United States Department of Energy

**List current March 2012*

Funding Agency

SIAM and the Conference Organizing Committee wish to extend their thanks and appreciation to the U.S. National Science Foundation for its support of this conference.

Leading the applied mathematics community . . .

Join SIAM and save!

SIAM members save up to $130 on full registration for the SIAM International Conference on Data Mining (SDM12)! Join your peers in supporting the premier professional society for applied mathematicians and computational scientists. SIAM members receive subscriptions to *SIAM Review* and *SIAM News* and enjoy substantial discounts on SIAM books, journal subscriptions, and conference registrations.

If you are not a SIAM member and paid the Non-Member rate to attend the conference, you can apply the difference between what you paid and what a member would have paid ($130) towards a SIAM membership. Contact SIAM Customer Service for details or join at the conference registration desk.

If you are a SIAM member, it only costs $10 to join the SIAM Activity Group on Data Mining and Analytics (SIAG/DMA). As a SIAG/DMA member, you are eligible for an additional $10 discount on this conference, so if you paid the SIAM member rate to attend the conference, you might be eligible for a free SIAG/DMA membership. Check at the registration desk.

Free Student Memberships are available to students who attend an institution that is an Academic Member of SIAM, are members of Student Chapters of SIAM, or are nominated by a Regular Member of SIAM.

Join onsite at the registration desk, go to *www.siam.org/joinsiam* to join online or download an application form, or contact SIAM Customer Service

Telephone: +1-215-382-9800 (worldwide); or 800-447-7426 (U.S. and Canada only)

Fax: +1-215-386-7999

E-mail: *membership@siam.org*

Postal mail:
Society for Industrial and Applied Mathematics,
3600 Market Street, 6th floor,
Philadelphia, PA 19104-2688 USA

Standard Audio/Visual Set-Up in Meeting Rooms

SIAM does not provide computers for any speaker. When giving an electronic presentation, speakers must provide their own computers. SIAM is not responsible for the safety and security of speakers' computers.

The Plenary Session Room will have two (2) overhead projectors, two (2) screens and one (1) data projector. Cables or adaptors for Apple computers are not supplied, as they vary for each model. Please bring your own cable/adaptor if using a Mac computer.

All other concurrent/breakout rooms will have one (1) screen and one (1) data projector. Cables or adaptors for Apple computers are not supplied, as they vary for each model. Please bring your own cable/adaptor if using a Mac computer. Overhead projectors will be provided only when requested.

If you have questions regarding availability of equipment in the meeting room of your presentation, or to request an overhead projector for your session, please see a SIAM staff member at the registration desk.

E-mail Access

Attendees booked in the SIAM room block will have access to complimentary wireless Internet access in the hotel sleeping rooms. SIAM will also provide a limited number of email stations.

Registration Fee Includes

- Admission to all technical sessions
- Admission to tutorial sessions
- Admission to a workshop
- CD of conference proceedings, workshop and tutorial notes
- Coffee breaks daily
- Continental Breakfast daily
- Room set-ups and audio/visual equipment
- Welcome Reception and Poster Session

Job Postings

Please check with the SIAM registration desk regarding the availability of job postings or visit *http://jobs.siam.org.*

Important Notice to Poster Presenters

The poster session is scheduled for Thursday, April 26 at 6:00 PM. Poster presenters are requested to set up their poster material on the provided 4' x 6' poster boards in the Crystal Cove Room beginning Thursday, April 26 at 7:00 AM. All materials must be posted by Thursday, April 26 by 6:00 PM, the official start time of the session. Poster displays must be removed by Friday, April 27 by 10:00 AM. Posters remaining after this time will be discarded. SIAM is not responsible for discarded posters.

SIAM Books and Journals

Display copies of books and complimentary copies of journals are available on site. SIAM books are available at a discounted price during the conference. If a SIAM books representative is not available, completed order forms and payment (credit cards are preferred) may be taken to the SIAM registration desk. The books table will close at 12:30 PM on Saturday, April 28.

Table Top Displays

SIAM

Springer

Conference Sponsors

Google

IBM Research

Data Mining and Knowledge Discovery

Data Mining and Knowledge Discovery is pleased to sponsor the SIAM SDM Best Student Paper Award. This premier technical journal in data mining is published by Springer Science & Business Media, a leading publisher of scholarly books and journals.

Partial Sponsorship of the Doctoral Forum Reception and Poster Session:

School of Computing, Informatics, and Decision Systems Engineering

Ira A. Fulton Schools of Engineering, Arizona State University

Name Badges

A space for emergency contact information is provided on the back of your name badge. Help us help you in the event of an emergency!

Comments?

Comments about SIAM meetings are encouraged! Please send to:

Sven Leyffer, SIAM Vice President for Programs (*vpp@siam.org*)

Get-togethers

- Welcome Reception and Poster Session, Thursday, April 26,
  6:00 PM – 9:00 PM
- Business Meeting
  (open to SIAG/DMA members)
  Friday, April 27,
  5:15 PM – 5:45 PM
  Complimentary wine and beer will be served.
- Local Reception, Doctoral Forum and Student Posters
  Friday, April 27
  7:30 PM - 9:30 PM

Please Note

SIAM is not responsible for the safety and security of attendees' computers. Do not leave your laptop computers unattended. Please remember to turn off your cell phones, pagers, etc. during sessions.

Recording of Presentations

Audio and video recording of presentations at SIAM meetings is prohibited without the written permission of the presenter and SIAM.

Social Media

SIAM is promoting the use of social media, such as Facebook and Twitter, in order to enhance scientific discussion at its meetings and enable attendees to connect with each other prior to, during and after conferences. If you are tweeting about a conference, please use the designated hashtag to enable other attendees to keep up with the Twitter conversation and to allow better archiving of our conference discussions. The hashtag for this meeting is #SIAMSDM12.

Invited Plenary Speakers

* *All Invited Plenary Presentations will take place in Pacific Ballroom AB **

Thursday, April 26

8:15 AM – 9:30 AM

IP1 Rapid Learning Systems to Improve Patient Outcomes and Control Health Costs

Bharat Rao, *SIEMENS Healthcare - Health Services, USA*

1:30 PM – 2:45 PM

IP2 Some Assembly Required: Organizing in the 21st Century

Noshir Contractor, *Northwestern University, USA*

Friday, April 27

8:15 AM – 9:30 AM

IP3 Cross-Domain Knowledge Transfer in Data Mining

Qiang Yang, *Hong Kong University of Science and Technology, Hong Kong*

1:30 PM – 2:45 PM

IP4 Temporal Dynamics and Information Retrieval

Susan Dumais, *Microsoft Research, USA*

Tutorials

* *All Tutorials will take place in Santa Monica - 2nd Floor **

Thursday, April 26

10:00 AM – 12:05 PM

TS1: Tutorial Session: Distance Metric Learning in Data Mining

Jimeng Sun, *IBM T.J. Watson Research Center, USA*

Fei Wang, *IBM T.J. Watson Research Center, USA*

3:00 PM – 5:05 PM

TS2: Tutorial Session: Discovering Roles and Anomalies in Graphs:Theory and Applications

Tina Eliassi-Rad, *Rutgers University, USA*

Christos Faloutsos, *Carnegie Mellon University, USA*

Friday, April 27

10:00 AM – 12:05 PM

TS3: Tutorial Session: Multi-Task Learning: Theory, Algorithms, and Applications

Jieping Ye, *Arizona State University, USA*

Jiayu Zhou, *Arizona State University, USA*

Saturday, April 28

8:30 AM – 12:00 PM

TS4: Tutorial Session: Privacy-Preserving Medical Data Sharing

Aris Gkoulalas-Divanis, *IBM Research-Zurich, Switzerland*

Grigorios Loukides, *Cardiff University, United Kingdom*

1:30 PM – 5:00 PM

TS5: Tutorial Session:

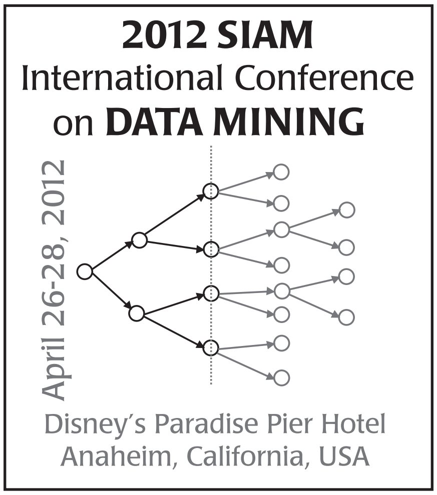
How to do Good Research and Get it Published in Top Venues

Eamonn Keogh, *University of California, Riverside, USA*

SIAM Activity Group on Data Mining

NEW! (SIAG/DMA)

www.siam.org/activity/dma



2012 SIAM International Conference on **DATA MINING**

April 26-28, 2012

Disney's Paradise Pier Hotel
Anaheim, California, USA

A GREAT WAY TO GET INVOLVED!

Collaborate and interact with mathematicians and applied scientists whose work involves data mining.

ACTIVITES INCLUDE:

- Annual conference and proceedings
- Special sessions at SIAM Annual Meetings
- Website

BENEFITS OF SIAG/DMA MEMBERSHIP:

- Listing in the SIAG's online-only membership directory
- Additional $10 discount on registration at the SIAM International Conference on Data Mining (excludes student)
- Electronic communications about recent developments in your specialty
- Eligibility for candidacy for SIAG/DMA office
- Participation in the selection of SIAG/DMA officers

ELIGIBILITY:

- Be a current SIAM member

COST:

- $10 per year
- Student SIAM members can join 2 activity groups for free!

2012-2013 SIAG/DMA OFFICERS:

- Chair: Chandrika Kamath, Lawrence Livermore National Laboratory
- Vice Chair: Charu Aggarwal, IBM T. J. Watson Research Center
- Program Director: Huan Liu, Arizona State University
- Secretary: Michael Mahoney, Stanford University

TO JOIN:

SIAG/DMA: ***my.siam.org/forms/join_siag.htm***

SIAM: ***www.siam.org/joinsiam***

Full Day Workshops

Saturday, April 28

Broadening Participation in Data Mining

7:45 AM – 5:30 PM

Co-Chairs:

Caio V. Soares, *Robert Bosch Research and Technology Center, USA*

Brandeis Marshall, *Purdue University, USA*

Soundararajan Srinivasan, *Robert Bosch Research and Technology Center, USA*

Shaun Gittens, *Booz Allen Hamilton INC., USA*

Pacific Ballroom A - 1st Floor

See page 14 for schedule

Text Mining 2012

8:50 AM – 4:30 PM

Co-Chairs:

Michael W. Berry, *University of Tennessee, USA*

Jacob Kogan, *University of Maryland, Baltimore County, USA*

Pacific Ballroom B - 1st Floor

See page 15 for schedule

Half Day Workshops

Saturday, April 28

Data Mining in Official Statistics Workshop

8:30 AM – 12:00 PM

Co-Chairs:

Anna Klimova, *University of Washington, Seattle, USA*

Tamas Rudas, *Eotvos Lorand University, Budapest, Hungary*

San Diego - 2nd Floor

See page 17 for schedule

Dynamic Network Analysis Workshop

8:30 AM – 12:00 PM

Co-Chairs:

Sitaram Asur, *Hewlett-Packard Labs, USA*

Duygu Ucar, *Stanford University, USA*

Oceanside - 2nd Floor

See page 18 for schedule

Half Day Workshops

Saturday, April 28

MultiClust Workshop:
Discovering, Summarizing and Using Multiple Clusterings
8:30 AM – 12:00 PM
Co-Chairs:
Emmanuel Müller, *Karlsruhe Institute of Technology (KIT), Germany*
Thomas Seidl, *RWTH Aachen University, Germany*
Suresh Venkatasubramanian, *University of Utah*
Arthur Zimek, *Ludwig-Maximilians-Universität München (LMU Munich), Germany*
Redondo - 2nd Floor
See page 18 for schedule

Analytics for Cyber-Physical Systems
1:30 PM – 5:00 PM
Co-Chairs:
Umeshwar Dayal, *Hewlett-Packard Labs, USA*
Chetan Gupta, *Hewlett-Packard Labs, USA*
Varun Chandola, *Oak Ridge National Laboratory, USA*
Ranga Raju Vatsavai, *Oak Ridge National Laboratory, USA*
Robert Grossman, *University of Chicago, USA*
Elke Rundensteiner, *Worcester Polytechnic Institute, USA*
San Diego - 2nd Floor
See page 19 for schedule

Broadening Participation in Data Mining

Full Day Workshop Schedule

Pacific Ballroom A

Saturday, April 28

7:45 AM - 9:00 AM

Feature Speaker

Vipin Kumar, *University of Minnesota, USA*

9:00 AM – 10:00 AM

Panel: Data Mining Research
(Academia vs. Industry)

10:00 AM –10:30 AM

Coffee Break and Networking

10:30 AM – 11:30 AM

Mentoring/Networking Activity

11:30 AM – 1:00 PM

Lunch Break

1:00 PM – 2:00 PM

Feature Speaker

Malu Castellanos, *Hewlett-Packard Labs, USA*

2:00 PM – 3:00 PM

Panel: Challenges of an Underrepresented Data Mining Researcher

3:00 PM – 3:30 PM

Coffee Break and Networking

3:30 PM – 4:30 PM

Panel: Preparing for a Career in Data Mining

4:30 PM – 5:30 PM

Wrap-up Session

Text Mining 2012

Full Day Workshop Schedule

Pacific Ballroom B - 1st Floor

Saturday, April 28

8:50 AM – 9:00 AM

Introduction
Michael W. Berry, *University of Tennessee, Knoxville, USA*

9:00 AM – 10:00 AM

Keynote Speaker

Tapping Social Media for Sentiments with Live Customer Intelligence (LCI)
Malu Castellanos, *Hewlett-Packard Labs, USA*

The explosion of Web opinion data that Web 2.0 and its increasingly popular social sites like Twitter, Facebook, blogs and review sites have brought about, has made essential the need for automatic tools to analyze and understand sentiments toward different topics. This has fueled the emerging field known as sentiment analysis whose goal is to translate the vagaries of human emotion into hard data. Live Customer Intelligence (LCI) is a system that taps into what is being said to understand the sentiment with the particular ability of doing so in near real-time. LCI integrates novel algorithms for sentiment analysis and a configurable dashboard with different kinds of charts including dynamic ones that change as new data is ingested. LCI has been researched and prototyped at HP Labs in close interaction with business divisions and a few selected customers. In this talk I give an overview of LCI, focusing in particular on challenging issues and illustrating its capabilities with selected use cases.

10:00 AM – 10:30 AM

Coffee Break

Session I: Document Ranking and Representation (Michael W. Berry, Chair)

10:30 AM – 11:00 AM

Finding Interesting Documents in a Corpus

Pradeep Chandrasekaran and *David Skillicorn*

11:00 AM – 11:30 AM

Granules of Words in Text Representation: An Approach Based on Fuzzy Relations

Patrícia Castro and Geraldo Xexéo

11:30 AM – 12:00 PM

Latent Semantic Indexing with Selective Query Expansion”

Andy Garron and April Kontostathis

12:00 PM – 1:30 PM

Lunch Break

Text Mining 2012

Full Day Workshop Schedule

Pacific Ballroom B - 1st Floor

Saturday, April 28

Session II: Document Classification and Clustering (Jacob Kogan, Chair)

1:30 PM – 2:00 PM

Incremental Clustering of News Reports

Joel Azzopardi and Christopher Staff

2:00 PM – 2:30 PM

The Effects of Tabular-based Content Extraction on Patent Document Clustering

Denise Koessler, Benjamin Martin, and Bruce Kiefer, and Michael W. Berry

2:30 PM – 3:00 PM

Extracting Hierarchies from Data Clusters for Better Classification"

German Sapozhnikov and *Alexander Ulanov*

3:00 PM – 3:30 PM

Coffee Break

Session III: Text Summarization and Anomaly Detection (Jacob Kogan, Chair)

3:30 PM – 4:00 PM

Better Metrics to Automatically Predict the Quality of a Text Summary"

Peter Rankel, John Conroy and Judith Schlesinger

4:00 PM – 4:30 PM

Contextual Anomaly Detection In Text Data"

*Amogh Mahapatr*a, Nisheeth Srivastava and Jaideep Srivastava

4:30 PM

Adjourn

Data Mining in Official Statistics Workshop

Half Day Workshop Schedule

San Diego - 2nd Floor

Saturday, April 28

8:30 AM – 8:55 AM

Model-Assisted Survey Estimation for Official Statistics Using Modern Regression Methods

Jay Breidt

8:55 AM – 9:15 AM

Analyzing Unemployment Data Using Symbolic Data Analysis

Paula Brito

9:15 AM – 9:35 AM

Shifting Paradigms in Official Statistics: From Design-based to Model-based to Algorithmic Inference

Bart Buelens

9:35 AM – 10:00 AM

Model Selection for Survey Data

Jean Opsomer

10:00 AM – 10:40 PM

Coffee Break and Posters

Poster Submissions

Mining Association in a Valued Network

Anna Klimova

On Evaluation Methods for Set-Based Bayesian Networks Structure Learning

Hoai-Tuong Nguyen

Estimating Demographic Parameters with Uncertainty from Fragmentary Data

Mark Wheldon

10:40 AM – 11:05

Mining Official Data: What’s New?

Gilbert Saporta

11:05 AM – 11:30 AM

Data Mining Analysis of Treatment Costs for Medical Tourists

Ram Shanmugam

11:30 AM – 11:50 AM

On Testing the Independence Assumption in Capture-recapture Studies

Tamás Rudas

11:50 AM – 12:00 PM

Closing Discussion

Dynamic Network Analysis Half Day Workshop Schedule

Oceanside - 2nd Floor

Saturday, April 28

8:30 AM – 9:30 AM

Invited Talk

Srinivasan Parthasarathy, Ohio State University, USA

9:30 AM – 10:00 AM

Full Paper - Swayed by Friends or by the Crowd?

10:00 AM – 10:30 AM

Coffee Break

10:30 AM – 11:00 AM

Full Paper - Algorithms for Offline Tracking of Connected Components in Large Evolving Networks

11:00 AM – 11:30 AM

Full Paper - Artificial Inflation: the Real Story of Trends in Sina Weibo

11:30 AM – 11:45 AM

Short Paper - Retrieval of Relevant and Non-redundant Nodes

11:45 AM – 12:00 PM

Short Paper - The Pulse of News in Social Media: Forecasting Popularity

MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings

Half Day Workshop Schedule

Redondo - 2nd Floor

Saturday, April 28

Invited Talk

Subspace Clustering Ensembles

Carlotta Domeniconi

Research Papers

Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering Approach

Shehroz Khan and Amir Ahmad

Co-RCA: Unsupervised Distance-Learning for Multi-View Clustering

Matthias Schubert and Hans-Peter Kriegel

New Subspace Clustering Problems in the Smartphone Era

Dimitrios Gunopulos and Vana Kalogeraki

Multiple Clustering Views via Constrained Projection

Xuan-Hong Dang, Ira Assent and James Bailey

Explorative Multi-View Clustering Using Frequent-Groupings

Martin Hahmann, Markus Dumat, Dirk Habich and Wolfgang Lehner

Analytics for Cyber-Physical Systems Workshop
Half Day Workshop Schedule

San Diego - 2nd Floor

Saturday, April 28

1:00 PM – 1:05 PM

Opening Remarks

Chetan Gupta, Hewlett-Packard Labs, USA

1:05 PM – 1:50 PM

Invited Talk

TransDec: A Data-Driven Framework for Decision-Making in Transportation Systems

Cyrus Shahabi, University of Southern California, USA

1:50 PM – 2:00 PM

Break

2:00 PM – 2:20 PM

Performance Monitoring via Nonparametric Adaptive Eventrate Analysis

Ron Maurer and Alina Maor

2:20 PM – 2:40 PM

QoX Driven Framework for Sensor Event Stream Analytics

Di Wang, Lei Cao, Qingyang Wang, Elke A.Rundensteiner

2:40 PM – 3:00 PM

Enabling Partial Data Cube Computations using the Bloom Filter

Ram Kosuru, Lakshminarayan Choudur

3:00 PM – 3:20 PM

Coffee Break

3:20 PM – 4:05 PM

Invited Talk

Streaming Models and Systems for Smarter Transportation Systems: Challenges and Solutions

Wei Fan, IBM T. J. Watson Research Center, USA

4:05 PM – 4:25 PM

Spatio-temporal Analysis for Identifying Grid Disruptions

Olufemi Omitaoumu, Steven J. Fernandez, Varun Chandola

4:25 PM – 4:45 PM

Reactive Process Monitoring via Semi-Automatic Decomposition of System-Wide Diagnostic Signals

Ayelet Pnueli, Ron Maurer, Ami Shiff, Avner Arnstein, TsafrirYedid-Am, Lior Katz

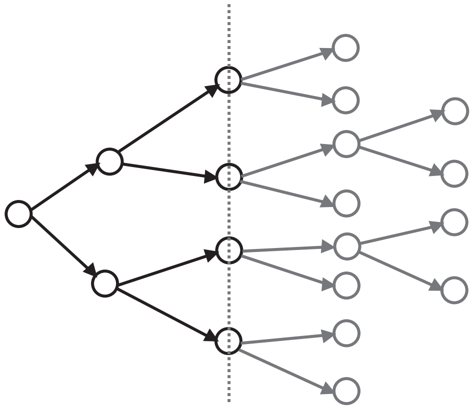
4:45 PM – 4:50 PM

Closing Remarks

Varun Chandola

SDM12 Program

2012 SIAM
International Conference
on **DATA MINING**



April 26-28, 2012

Disney's Paradise Pier Hotel
Anaheim, California, USA

Wednesday, April 25

Registration

5:00 PM-7:00 PM

Room:Pacific Ballroom Foyer - 1st Floor

Thursday, April 26

Registration

7:00 AM-7:30 PM

Room:Pacific Ballroom Foyer - 1st Floor

Announcements

8:00 AM-8:15 AM

Room:Pacific Ballroom AB - 1st Floor

Thursday, April 26

IP1

Rapid Learning Systems to Improve Patient Outcomes and Control Health Costs

8:15 AM-9:30 AM

Room:Pacific Ballroom AB - 1st Floor

Chair: Ian Davidson, University of California, Davis, USA

In this talk, I will briefly present data mining solutions that analyze millions of patient records, impacting three major areas in healthcare. These include automated quality measurement and decision-support from hospitals EMR's, computer-aided diagnosis systems to identify suspicious lesions on medical images, and "rapid learning systems" to develop predictive models for personalized medicine. The last is based on a first-of-kind rapid learning system: a Euro-US health IT network spanning cancer centers in 5 nations to learn personalized therapies for lung cancer. The majority of the talk will present case studies that illustrate some of the challenges unique to mining healthcare data, and identify a few promising areas for research. These include the breakdown of traditional assumptions inherent in most mining algorithms, learning from multi-source systems, and the development of predictive models for personalized medicine. We conclude with a glimpse of a more-efficient healthcare future, where treatment decisions are driven by evolving knowledge that is continuously mined from patient records collected in health systems all over the world.

Bharat Rao
SIEMENS Healthcare - Health Services, USA

Coffee Break

9:30 AM-10:00 AM

Room:Pacific Ballroom Foyer - 1st Floor

Thursday, April 26

CP1

Applications - Climate and Geography

10:00 AM-12:05 PM

Room:Pacific Ballroom A - 1st Floor

Chair: Arindam Banerjee, University of Minnesota, USA

10:00-10:20 Detecting and Tracking Coordinated Groups in Dense, Systematically Moving, Crowds

James C. Rosswog and Kanad Ghose, Binghamton University, USA

10:25-10:45 Large-Scale Nonparametric Estimation of Vehicle Travel Time Distributions

Rikiya Takahashi, Takayuki Osogami, and Tetsuro Morimura, IBM Research - Tokyo, Japan

10:50-11:10 Drought Detection of the Last Century: An Mrf-Based Approach

Qiang Fu, University of Minnesota, Twin Cities, USA; Arindam Banerjee, University of Minnesota, USA; Stefan Liess and Peter Snyder, University of Minnesota, Twin Cities, USA

11:15-11:35 Toward Data-Driven, Semi-Automatic Inference of Phenomenological Physical Models: Application to Eastern Sahel Rainfall

Saurabh V. Pendse, Isaac Tetteh, and Fredrick Semazzi, North Carolina State University, USA; Vipin Kumar, University of Minnesota, Minneapolis, USA; Nagiza Samatova, North Carolina State University, USA

11:40-12:00 Sparse Group Lasso: Consistency and Climate Applications

Soumyadeep Chatterjee and Karsten Steinhaeuser, University of Minnesota, Twin Cities, USA; Arindam Banerjee, University of Minnesota, USA; Snigdhansu Chatterjee, University of Minnesota, Twin Cities, USA; Auroop Ganguly, Northeastern University, USA

Thursday, April 26

CP2

Clustering

10:00 AM-12:05 PM

Room:Pacific Ballroom B - 1st Floor

Chair: Charu C. Aggarwal, IBM T.J. Watson Research Center, USA

10:00-10:20 The Multi-Set Stream Clustering Problem

Charu C. Aggarwal, IBM T.J. Watson Research Center, USA

10:25-10:45 Stratification Based Hierarchical Clustering Over a Deep Web Data Source

Tantan Liu and Gagan Agrawal, Ohio State University, USA

10:50-11:10 Cluster-Aware Compression with Provable K-Means Preservation

Nikolaos Freris and Michail Vlachos, IBM Research-Zurich, Switzerland; Deepak Turaga, IBM T.J. Watson Research Center, USA

11:15-11:35 Supervised Clustering of Label Ranking Data

Mihajlo Grbovic, Nemanja Djuric, and Slobodan Vucetic, Temple University, USA

11:40-12:00 Symmetric Nonnegative Matrix Factorization for Graph Clustering

Da Kuang, Georgia Institute of Technology, USA; Chris Ding, University of Texas at Arlington, USA; Haesun Park, Georgia Institute of Technology, USA

Thursday, April 26

CP3

Social Media

10:00 AM-12:05 PM

Room:San Diego - 2nd Floor

Chair: Huan Liu, Arizona State University, USA

10:00-10:20 Feature Selection with Linked Data in Social Media

Jiliang Tang and Huan Liu, Arizona State University, USA

10:25-10:45 Microscopic Social Influence

Ting Wang, Mudhakar Srivatsa, and Dakshi Agrawal, IBM T.J. Watson Research Center, USA; Ling Liu, Georgia Institute of Technology, USA

10:50-11:10 Har: Hub, Authority and Relevance Scores in Multi-Relational Data for Query Search

Xutao Li, Harbin Institute of Technology, China; Michael K. Ng, Hong Kong Baptist University, Hong Kong; YunmingYe, Harbin Institute of Technology, China

11:15-11:35 Evaluating Event Credibility on Twitter

Manish Gupta, Peixiang Zhao, and Jiawei Han, University of Illinois at Urbana-Champaign, USA

11:40-12:00 Multi-Skill Collaborative Teams Based on Densest Subgraphs

Amita Gajewar, Yahoo! Labs, Santa Clara, USA; Atish Das Sarma, Google Research, USA

Thursday, April 26

TS1

Tutorial Session: Distance Metric Learning in Data Mining

10:00 AM-12:05 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

In this tutorial, we will give an overview of the existing distance metric learning approaches and point out the challenges as well as future research directions. Specifically, this tutorial will cover two parts: (1) traditional distance metric learning algorithms, including supervised and unsupervised methods; (2) advanced topics, including online methods, distributed metric learning, active metric learning and transferred metric learning. Finally we will illustrate some applications of distance metric learning techniques and ummarize some challenges in this field and propose some future trends.

Jimeng Sun
Fei Wang
IBM T.J. Watson Research Center, USA

Lunch Break

12:05 PM-1:30 PM

Attendees on their own

Thursday, April 26

IP2

Some Assembly Required: Organizing in the 21st Century

1:30 PM-2:45 PM

Room:Pacific Ballroom AB - 1st Floor

Chair: Joydeep Ghosh, University of Texas at Austin, USA

Recent advances in Web Science provide comprehensive digital traces of social actions, interactions, and transactions. These data provide an unprecedented exploratorium to model the socio-technical motivations for creating, maintaining, dissolving, and reconstituting into teams – for research, business, or social causes. Using examples from research in team science and massively multiplayer online games, Contractor will argue that Network Science serves as the foundation for the development of social network theories and methods to help advance our ability to understand the emergence of effective teams. More importantly, he will argue that these insights will also enable effective teams by building a new generation of recommender systems that leverage our research insights on the socio-technical motivations for creating ties.

Noshir Contractor
Northwestern University, USA

Coffee Break

2:45 PM-3:00 PM

Room:Pacific Ballroom Foyer - 1st Floor

Thursday, April 26

CP4

Multi-Source and Multi-Task

3:00 PM-5:05 PM

Room:Pacific Ballroom A - 1st Floor

Chair: Jieping Ye, Arizona State University, USA

3:00-3:20 Heterogeneous Data Fusion Via Space Alignment Using Nonmetric Multidimensional Scaling
Jaegul Choo, Georgia Institute of Technology, USA; Shawn Bohn, Grant Nakamura, and Amanda White, Pacific Northwest National Laboratory, USA; Haesun Park, Georgia Institute of Technology, USA

3:25-3:45 Heterogeneous Datasets Representation and Learning using Diffusion Maps and Laplacian Pyramids
Neta Rabin, Yale University, USA

3:50-4:10 A Bayesian Nonparametric Joint Factor Model for Learning Shared and Individual Subspaces from Multiple Data Sources
Sunil K. Gupta, Dinh Phung, and Svetha Venkatesh, Deakin University, Australia

4:15-4:35 Adaptive Multi-Task Sparse Learning with An Application to Fmri Study
Xi Chen, Carnegie Mellon University, USA; *Jingrui He* and Rick Lawrence, IBM T.J. Watson Research Center, USA; Jaime Carbonell, Carnegie Mellon University, USA

4:40-5:00 Learning from Heterogeneous Sources Via Gradient Boosting Consensus
Xiaoxiao Shi, University of Ilinois at Chicago, USA; Jean-Francois Paiement and David Grangier, AT&T Labs - Research, USA; Philip Yu, University of Ilinois at Chicago, USA

Thursday, April 26

CP5

Pattern Mining

3:00 PM-5:05 PM

Room:Pacific Ballroom B - 1st Floor

Chair: Bart Goethals, University of Antwerp, Belgium

3:00-3:20 Slim: Directly Mining Descriptive Patterns

Koen Smets and Jilles Vreeken, Universiteit Antwerpen, Belgium

3:25-3:45 Marbles: Mining Association Rules Buried in Long Event Sequences

Boris Cule, Nikolaj Tatti, and Bart Goethals, University of Antwerp, Belgium

3:50-4:10 Mining Patterns in Networks Using Homomorphism

Anton Dries, Universitat Pompeu Fabra, Spain; *Siegfried Nijssen*, Katholieke Universiteit Leuven, Belgium

4:15-4:35 Scalable Induction of Probabilistic Real-Time Automata Using Maximum Frequent Pattern Based Clustering

Jana Schmidt and Sonja Ansorge, TU München, Germany; Stefan Kramer, Johannes Gutenberg-Universität, Mainz, Germany

4:40-5:00 Class Relevant Pattern Mining in Output-Polynomial Time

Henrik Grosskreutz, Fraunhofer Institute for Autonomous Intelligent Systems, Germany

Thursday, April 26

CP6

Time Series and Sequence Analysis

3:00 PM-5:05 PM

Room: San Diego - 2nd Floor

Chair: Varun Chandola, University of Minnesota, USA

3:00-3:20 Simplex Distributions for Embedding Data Matrices over Time

Kristian Kersting and *Mirwaes Wahabzada*, Fraunhofer Institute for Autonomous Intelligent Systems, Germany; Christoph Roemer, University of Bonn, Germany; Christian Thurau, Fraunhofer Institute for Autonomous Intelligent Systems, Germany; Agim Ballvora, University of Bonn, Germany; Uwe Rascher, Forschungszentrum Jülich, Germany; Jens Leon, University of Bonn, Germany; Christian Bauckhage, Fraunhofer Institute for Autonomous Intelligent Systems, Germany; Lutz Pluemer, University of Bonn, Germany

3:25-3:45 Transformation Based Ensembles for Time Series Classification

Anthony Bagnall, Luke Davis, Jon Hills, and *Jason Lines*, University of East Anglia, United Kingdom

3:50-4:10 Mining Compressing Sequential Patterns

Hoang Thanh Lam, Technische Universiteit Eindhoven, The Netherlands; Fabian Moerchen and Dmitriy Fradkin, Siemens Corporation Research, Germany; Toon Calders, Technische Universiteit Eindhoven, The Netherlands

4:15-4:35 Beam Methods for the Profile Hidden Markov Model

Samuel J. Blasiak, Huzefa Rangwala, and Kathryn Laskey, George Mason University, USA

4:40-5:00 Optimal Distance Estimation Between Compressed Data Series

Nikolaos Freris and Michail Vlachos, IBM Research-Zurich, Switzerland; Serdar Kozat, Koc University, Turkey

Thursday, April 26

TS2

Tutorial Session: Discovering Roles and Anomalies in Graphs: Theory and Applications

3:00 PM-5:05 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

The objective of this tutorial is to provide a concise and intuitive overview of the most important concepts and tools, which can detect roles (or functions) for nodes in both static and dynamic graphs. We review the state of the art in three related fields: (a) community discovery, (b) equivalences (from sociology), and (c) propositionalisation (from multi-relational data mining). The emphasis of this tutorial is to give the intuition behind these powerful mathematical concepts and tools, which are usually lost in the various technical literatures, as well as to give case studies that illustrate their practical use.

Tina Eliassi-Rad

Rutgers University, USA

Christos Faloutsos

Carnegie Mellon University, USA

Organizational Break

5:05 PM-5:15 PM

Thursday, April 26

PP1

Plenary Session: Short Presentations

5:15 PM-6:00 PM

Room:Pacific Ballroom AB - 1st Floor

To Be Determined

Event Detection in Social Streams

Charu C. Aggarwal, IBM T.J. Watson Research Center, USA; Karthik Subbian, University of Minnesota, USA

On Influential Node Discovery in Dynamic Social Networks

Charu C. Aggarwal, IBM T.J. Watson Research Center, USA; Shuyang Lin and Philip Yu, University of Ilinois at Chicago, USA

Query-based Biclustering using Formal Concept Analysis

Faris Alqadah, Johns Hopkins University, USA; Joel S. Bader, Johns Hopkins University, USA; Rajul Anand and Chandan Reddy, Wayne State University, USA

Granger Causality Analysis in Irregular Time Series

Mohammad Taha Bahadori and Yan Liu, University of Southern California, USA

Clustering Based on Yukawa Potential

Xue Bai, Zezhen Lin, Yun Xiong, and Yangyong Zhu, Fudan University, China

Deterministic Cur for Improved Large-Scale Data Analysis: An Empirical Study

Christian Thurau, Kristian Kersting,and Christian Bauckhage, Fraunhofer Institute for Autonomous Intelligent Systems, Germany

Combining Active Learning and Dynamic Dimensionality Reduction

Mustafa Bilgic, Illinois Institute of Technology, USA

Context-Aware Search for Personal Information Management Systems

Jidong Chen, EMC Corporation, China; Wentao Wu, Fudan University, China; Hang Guo, EMC Corporation, China; Wei Wang, Fudan University, China

continued in next column

Mining Social Dependencies in Dynamic Interaction Networks

Freddy Chua, Hady Lauw, and Ee-Peng Lim, Singapore Management University, Singapore

Detecting Irregularly Shaped Significant Spatial and Spatio-Temporal Clusters

Weishan Dong, Xin Zhang, Li Li, Changhua Sun, Lei Shi, and Wei Sun, IBM Research, China

Contextual Collaborative Filtering Via Hierarchical Matrix Factorization

Erheng Zhong, Sun Yat-Sen University, China; *Wei Fan*, IBM T.J. Watson Research Center, USA; Qiang Yang, Hong Kong University of Science and Technology, Hong Kong

Active Learning with Monotonicity Constraints

Ad Feelders and Nicola Barile, Universiteit Utrecht, The Netherlands

Pseudo Cold Start Link Prediction with Multiple Sources in Social Networks

Liang Ge, State University of New York, Buffalo, USA

Discovering Context-Aware Influential Objects

Huiping Cao, Yangpai Liu, and *Yifan Hao*, New Mexico State University, USA; Peng Han and Xinda Zeng, Chongqing Academy of Science and Technology, China

Monitoring and Mining Insect Sounds in Visual Space

Yuan Hao, Bilson J. Campana, and Eamonn Keogh, University of California, Riverside, USA

Image Mining of Historical Manuscripts to Establish Provenance

Bing Hu, University of California, Riverside, USA

RP-growth: Top-k Mining of Relevant Patterns with Minimum Support Raising

Yoshitaka Kameya and Taisuke Sato, Tokyo Institute of Technology, Japan

Fast Random Walk Graph Kernel

U Kang, Carnegie Mellon University, USA; Hanghang Tong and Jimeng Sun, IBM T.J. Watson Research Center, USA

continued in next column

Tracking Spatio-Temporal Diffusion in Climate Data

Jaya Kawale and Aditya Pal, University of Minnesota, USA; Rob Fatland, Microsoft Research, USA

Group Sparsity in Nonnegative Matrix Factorization

Jingu Kim, Renato C. Monteiro, and Haesun Park, Georgia Institute of Technology, USA

Global Linear Neighborhoods for Efficient Label Propagation

Ze Tian and *Rui Kuang*, University of Minnesota, USA

Generalized Similarity Kernels for Efficient Sequence Classification

Pavel P. Kuksa, Rutgers University, USA; Imdadullah Khan, Gulf University for Science and Technology, Kuwait; Vladimir Pavlovic, Rutgers University, USA

Detecting Extreme Rank Anomalous Collections

Hanbo Dai, Feida Zhu, Ee-Peng Lim, Hwee Hwa Pang, and *Hady Lauw*, Singapore Management University, Singapore

Visualizing Variable-Length Time Series Motifs

Yuan Li and Jessica Lin, George Mason University, USA; Tim Oates, University of Maryland, Baltimore County, USA

Which Distance Metric Is Right: An Evolutionary K-Means View

Chuanren Liu, Rutgers University, USA; Tianming Hu, Dongguan University of Technology, China; Yong Ge and Hui Xiong, Rutgers University, USA

Constructing Training Sets for Outlier Detection

Liping Liu and Xiaoli Z. Fern, Oregon State University, USA

A Flexible Open-Source Toolbox for Scalable Complex Graph Analysis

Adam Lugowski, University of California, Santa Barbara, USA; Aydin Buluc, Lawrence Berkeley National Laboratory, USA; David Alber, Microsoft Corporation, USA; John R. Gilbert, University of California, Santa Barbara, USA; Steve Reinhardt, Cray, Inc., USA; Yun Teng and Andrew Waranis, University of California, Santa Barbara, USA

continued on next page

Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection

Fragkiskos D. Malliaros, University of Patras, Greece; Vasileios Megalooikonomou, University of Patras, Greece, and Temple University, USA; Christos Faloutsos, Carnegie Mellon University, USA

On Finding Joint Subspace Boolean Matrix Factorizations

Pauli Miettinen, Max Planck Institute for Informatics, Germany

Generalized Optimization Framework for Graph-Based Semi-Supervised Learning

Marina M. Sokol and Konstantin Avrachenkov, INRIA Sophia Antipolis, France; Paulo Goncalves, INRIA Rhone, France; *Alexey Mishenin*, St. Petersburg State University, Russia

A Tree-Based Kernel for Graphs

Nicolo' Navarin, Giovanni Da San Martino, and Alessandro Sperduti, University of Padova, Italy

Density-Based Projected Clustering over High Dimensional Data Streams

Eirini C. Ntoutsi and Arthur Zimek, Ludwig-Maximilians-Universität München, Germany; Themis Palpanas, University of Trento, Italy; Peer Kröger and Hans-Peter Kriegel, Ludwig-Maximilians-Universität München, Germany

A Novel Approximation to Dynamic Time Warping Allows Anytime Clustering of Massive Time Series Datasets

Qiang Zhu, Gustavo E. Batista, *Thanawin Rakthanmanon*, and Emaonn Keogh, University of California, Riverside, USA

Nearest-Neighbor Search on a Time Budget via Max-Margin Trees

Parikshit Ram, Dongryeol Lee, and Alexander Gray, Georgia Institute of Technology, USA

Efficient Clustering of Metagenomic Sequences Using Locality Sensitive Hashing

Zeehasham Rasheed, Huzefa Rangwala, and Daniel Barbara, George Mason University, USA

continued in next column

Balancing Prediction and Recommendation Accuracy: Hierarchical Latent Factors for Preference Data

Ettore Ritacco, *Nicola Barbieri*, Giuseppe Manco, and Riccardo Ortale, ICAR-CNR, Italy

On Evaluation of Outlier Rankings and Outlier Scores

Arthur Zimek, *Erich Schubert*, Remigius Wojdanowski, and Hans-Peter Kriegel, Ludwig-Maximilians-Universität München, Germany

Regularized Structured Output Learning with Partial Labels

Sundararajan Sellamanickam, Charu Tiwari, and Sathiya Keerthi Selvaraj, Yahoo! Labs, Bangalore, India

The Similarity Between Stochastic Kronecker and Chung-Lu Graph Models

C. Seshadhri, Ali Pinar, and Tamara G. Kolda, Sandia National Laboratories, USA

Wigm: Discovery of Subgraph Patterns in a Large Weighted Graph

Wei Su and Jiong Yang, Case Western Reserve University, USA; Shirong Li, Aliyun Inc., China; Mehmet Dalkilicb, Indiana University, USA

Legislative Prediction Via Random Walks over a Heterogeneous Graph

Jun Wang, Kush Varshne, and Aleksandra Mojsilovic, IBM T.J. Watson Research Center, USA

An Iterative and Re-Weighting Framework for Rejection and Uncertainty Resolution in Crowdsourcing

Sihong Xie, University of Illinois, Chicago, USA; Wei Fan, IBM T.J. Watson Research Center, USA; Philip Yu, University of Ilinois at Chicago, USA

Citation Prediction in Heterogeneous Bibliographic Networks

Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han, University of Illinois at Urbana-Champaign, USA

Mining Multi-Label Data Streams Using Ensemble-Based Active Learning

Peng Wang, *Peng Zhang*, and Li Guo, Chinese Academy of Sciences, China

continued in next column

Feature Selection for High-Dimensional Integrated Data

Charles Y. Zheng, Texas A&M University, USA; Scott Schwartz, Texas Agrilife, USA; Robert Chapkin, Raymond Carroll, and Ivan Ivanov, Texas A&M University, USA

Welcome Reception and Poster Session

6:00 PM-9:00 PM

Room: Pacific Ballroom CD - 1st Floor

Friday, April 27

Registration

7:30 AM-3:30 PM

Room:Pacific Ballroom Foyer - 1st Floor

Announcements

8:00 AM-8:15 AM

Room:Pacific Ballroom AB - 1st Floor

IP3

Cross-Domain Knowledge Transfer in Data Mining

8:15 AM-9:30 AM

Room:Pacific Ballroom AB - 1st Floor

Chair: Huan Liu, Arizona State University, USA

In data mining, we often encounter situations where we have an insufficient amount of high-quality data in a target domain, but we may have plenty of auxiliary data in related domains. Transfer learning aims to exploit these additional data to improve the learning performance in the target domain. In this talk, I will give an overview on some recent advances in transfer learning for challenging data mining problems. I will present structural transfer-learning solutions under heterogeneous feature representations. I will also survey cross-domain transfer learning solutions in online recommendation, social media and social network mining. I will discuss some current limitations of cross-domain transfer learning and explore possible future directions.

Qiang Yang
Hong Kong University of Science and Technology, Hong Kong

Coffee Break

9:30 AM-10:00 AM

Room:Pacific Ballroom Foyer - 1st Floor

Friday, April 27

CP7

Kernels and Classification

10:00 AM-12:05 PM

Room:Pacific Ballroom A - 1st Floor

Chair: Huzefa Rangwala, George Mason University, USA

10:00-10:20 Multi-Objective Multi-Label Classification
Chuan Shi, Beijing University of Posts and Telecommunications, China; Xiangnan Kong and Philip Yu, University of Ilinois at Chicago, USA; Bai Wang, Beijing University of Posts and Telecommunications, China

10:25-10:45 Bayesian Supervised Multilabel Learning with Coupled Embedding and Classification
Mehmet Gönen, Aalto University, Finland

10:50-11:10 Subtree Replacement in Decision Tree Simplification
Salvatore Ruggieri, Universita' di Pisa, Italy

11:15-11:35 A Distributed Kernel Summation Framework for General-Dimension Machine Learning
Dongryeol Lee, Richard Vuduc, and Alexander Gray, Georgia Institute of Technology, USA

11:40-12:00 Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information
Tinghui Zhou, Carnegie Mellon University, USA; Hanhuai Shan, University of Minnesota, Twin Cities, USA; Arindam Banerjee, University of Minnesota, USA; Guillermo Sapiro, University of Minnesota, Minneapolis, USA

Friday, April 27

CP8

Social Network and Graphs

10:00 AM-12:05 PM

Room:Pacific Ballroom B - 1st Floor

Chair: Jimeng Sun, IBM T.J. Watson Research Center, USA

10:00-10:20 On Dynamic Link Inference in Heterogeneous Networks
Charu C. Aggarwal, IBM T.J. Watson Research Center, USA; Yan Xie and Philip Yu, University of Ilinois at Chicago, USA

10:25-10:45 A Framework for the Evaluation and Management of Network Centrality
Vatche Ishakian, Dora Erdos, Evimaria Terzi, and Azer Bestavros, Boston University, USA

10:50-11:10 Parameter-Free Identification of Cohesive Subgroups in Large Attributed Graphs
Leman Akoglu, Carnegie Mellon University, USA; Hanghang Tong, IBM T.J. Watson Research Center, USA; Brendan Meeder and Christos Faloutsos, Carnegie Mellon University, USA

11:15-11:35 Structural Analysis in Multi-Relational Social Networks
Bing Tian Dai, Freddy Chua, and Ee-Peng Lim, Singapore Management University, Singapore

11:40-12:00 Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model
Xinran He and Guojie Song, Peking University, China; Wei Chen, Microsoft Research Asia; Qingye Jiang, Columbia University, USA

Friday, April 27

CP9

Feature Selection, Networks and Predication

10:00 AM-12:05 PM

Room: San Diego - 2nd Floor

Chair: Jacob Kogan, University of Maryland, Baltimore County, USA

10:00-10:20 Feature Selection over Distributed Data Streams through Optimization

Jacob Kogan, University of Maryland, Baltimore County, USA

10:25-10:45 Feature Selection ``Tomography' --- Illustrating That Optimal Feature Filtering Is Hopelessly Ungeneralizable

George Forman, HP Labs, USA

10:50-11:10 Sampling Strategies to Evaluate the Performance of Unknown Predictors

Hamed Valizadegan, Saeed Amizadeh, and Milos Hauskrecht, University of Pittsburgh, USA

11:15-11:35 A Bayesian Markov-Switching Model for Sparse Dynamic Network Estimation

Huijing Jiang, Aurelie Lozano, and Fei Liu, IBM T.J. Watson Research Center, USA

11:40-12:00 Learning Hierarchical Relationships among Partially Ordered Objects with Heterogeneous Attributes and Links

Chi Wang and Jiawei Han, University of Illinois at Urbana-Champaign, USA; Qi Li, Xiang Li, Wen-Pin Lin, and Heng Ji, City University of New York, USA

Friday, April 27

TS3

Tutorial Session: Multi-Task Learning: Theory, Algorithms, and Applications

10:00 AM-12:05 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

This tutorial gives a comprehensive overview of theory, algorithms, and applications of multi-task learning. Many real-world applications involve multiple related classification/regression tasks. For example, in the prediction of therapy outcome, the tasks of predicting the effectiveness of several combinations of drugs are related. In the prediction of disease progression, the prediction of outcome at each time point can be considered as a task and these tasks are temporally related. Traditionally these tasks are solved independently, ignoring the task relatedness. In multi-task learning, we learn these related tasks simultaneously by extracting appropriate shared information across tasks. Multi-task learning is especially useful when the training sample size for each task is small.

Jieping Ye
Arizona State University USA

Jiayu Zhou
Arizona State University, USA

Lunch Break

12:05 PM-1:30 PM

Attendees on their own

Friday, April 27

IP4

Temporal Dynamics and Information Retrieval

1:30 PM-2:45 PM

Room:Pacific Ballroom AB - 1st Floor

Chair: Carlotta Domeniconi, George Mason University, USA

Many digital resources, like the Web, are dynamic and ever-changing collections of information. However, most information retrieval tools developed for interacting with Web content, such as browsers and search engines, focus on a single static snapshot of the information. In this talk, I will present analyses of how Web content changes over time, how people re-visit Web pages over time, and how re-visitation patterns are influenced by changes in user intent and content. These results have implications for many aspects of information retrieval and management including crawling policy, ranking and information extraction algorithms, result presentation, and systems evaluation. I will describe a prototype that supports people in understanding how the information they interact with changes over time, and new retrieval models that incorporate features about the temporal evolution of content to improve core ranking. Finally, I will conclude with an overview of some general challenges that need to be addressed to fully incorporate temporal dynamics in information retrieval and information management systems.

Susan Dumais
Microsoft Research, USA

Coffee Break

2:45 PM-3:00 PM

Room:Pacific Ballroom Foyer - 1st Floor

Friday, April 27

MS1

Data Mining, Uncertainty Quantification, and Opportunities for Collaboration

3:00 PM-5:05 PM

Room:Pacific Ballroom A - 1st Floor

The fields of data mining and uncertainty quantification have much in common. For example, both deal with issues related to data in high-dimensional spaces, consider statistical aspects of sampling, and build predictive models. In addition, the broad area of data analysis has had to deal with “reasoning under uncertainty”, and the term “Uncertainty in AI” has been around a long time. With the recent formation of the two SIAM Activity Groups (SIAGs) on Uncertainty Quantification (UQ) and Data Mining and Analytics (DMA), we see an opportunity for the two communities to come together and discuss topics of mutual interest.

Organizer: Roger Ghanem
University of Southern California, USA

Organizer: Habib N. Najm
Sandia National Laboratories, USA

Organizer: Chandrika Kamath
Lawrence Livermore National Laboratory, USA

3:05-3:30 Diffusion on Random Manifolds
Hadi Meidani and Roger Ghanem, University of Southern California, USA

continued in next column

3:35-4:00 Efficient Monte Carlo Computation of Fisher Information Matrix using Prior Information
Sonjoy Das, Massachusetts Institute of Technology, USA; James Spall, Johns Hopkins University, USA; Roger Ghanem, University of Southern California, USA

4:05-4:30 Probabilistic Models of Past Climate Change
Julien Emile-Geay and Dominique Guillot, University of Southern California, USA; Tapio Schneider, California Institute of Technology, USA; Bala Rajaratnam, Stanford University, USA

4:35-5:00 A Priori Testing of Adaptive Sampling and Sparse PC Representations for Ocean General Circulation Models
Justin Winokur, Johns Hopkins University, USA; Patrick R. Conrad, Massachusetts Institute of Technology, USA; Ihab Sraj and Alen Alexanderian, Johns Hopkins University, USA; Mohamed Iskandarani and Ashwanth Srinivasan, University of Miami, USA; Youssef M. Marzouk, Massachusetts Institute of Technology, USA; Omar Knio, Duke University, USA

Friday, April 27

MS2

Math Science, Data Mining and Emerging Applications

3:00 PM – 5:05 PM

Room: Santa Monica – 2nd Floor

Organizer: TBD

3:05 – 3:30 One New Algorithm for Ten New Applications
Charles Elkan, University of California, San Diego, USA

3:35 – 4:00 Computational Entomology: An Emerging Application of Data Mining
Eamonn Keogh, University of California, Riverside, USA

4:05 – 4:30 Algorithmic Primitives for Network Analysis: Through the Lens of the Laplacian Paradigm
Shang-Hua Teng, University of Southern California, USA

4:35 – 5:00 Modeling Social Network Event Data over Time
Padhraic Smyth, University of California, Irvine, USA

Friday, April 27

CP10

Transfer Learning

3:00 PM-4:40 PM

Room:Pacific Ballroom B - 1st Floor

Chair: Wei Fan, IBM T.J. Watson Research Center, USA

3:00-3:20 Transfer Learning of Distance Metrics by Cross-Domain Metric Sampling Across Heterogeneous Spaces

Guo-Jun Qi, University of Illinois at Urbana-Champaign, USA; Charu C. Aggarwal, IBM T.J. Watson Research Center, USA; Thomas Huang, University of Illinois at Urbana-Champaign, USA

3:25-3:45 Dual Transfer Learning

Mingsheng Long, Jianmin Wang, and Guiguang Ding, Tsinghua University, China; Wei Cheng, University of North Carolina at Chapel Hill, USA; Xiang Zhang, Case Western Reserve University, USA; Wei Wang, University of North Carolina at Chapel Hill, USA

3:50-4:10 Transfer Significant Subgraphs Across Graph Databases

Xiaoxiao Shi, Xiangnan Kong, and Philip Yu, University of Ilinois at Chicago, USA

4:15-4:35 Transfer Topic Modeling with Ease and Scalability

Jeon-Hyung Kang, Jun Ma, and Yan Liu, University of Southern California, USA

Friday, April 27

CP11

Applications - Healthcare and Networks

3:00 PM-4:40 PM

Room: San Diego - 2nd Floor

Chair: George Forman, HP Labs, USA

3:00-3:20 Sor: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and Its Healthcare Applications

Dijun Luo, University of Texas at Arlington, USA; Fei Wang, Jimeng Sun, Marianthi Marka, Jianying Hu, and Shahram Ebadollahi, IBM T.J. Watson Research Center, USA

3:25-3:45 Mining Massive Archives of Mice Sounds with Symbolized Representations

Jesin Zakaria, Sarah Rotschafer, Abdullah Mueen, Khaleel Razak, and Eamonn Keogh, University of California, Riverside, USA

3:50-4:10 IntruMine: Mining Intruders in Untrustworthy Data of Cyber-Physical Systems

Lu-An Tang, Quanquan Gu, Xiao Yu, and Jiawei Han, University of Illinois at Urbana-Champaign, USA; Thomas La Porta, Pennsylvania State University, USA; Alice Leung, BBN Technology, USA; Tarek Abdelzaher, University of Illinois at Urbana-Champaign, USA; Lance Kaplan, U.S. Army Research Laboratory, USA

4:15-4:35 Robust Reputation-Based Ranking on Bipartite Rating Networks

Rong-Hua Li, *Jeffery Xu Yu*, Xin Huang, and Hong Cheng, Chinese University of Hong Kong, Hong Kong

Organizational Break

5:05 PM-5:15 PM

SIAG/DMA Business Meeting

5:15 PM-5:45 PM

Room:Pacific Ballroom AB - 1st Floor

Complimentary wine and beer will be served.

Friday, April 27

Organizational Break

5:45 PM-6:00 PM

Funding Agency Panel

6:00 PM-7:30 PM

Room:Pacific Ballroom AB - 1st Floor

Local Reception, Doctoral Forum and Student Posters

7:30 PM-9:30 PM

Room: Pacific Ballroom CD - 1st Floor

Saturday, April 28

Registration

7:15 AM-4:00 PM

Room:Pacific Ballroom Foyer - 1st Floor

Broadening Participation in Data Mining Workshop

7:45 AM-10:00 AM

Room:Pacific Ballroom A - 1st Floor

For Schedule, see page 14

MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings

8:30 AM-10:00 AM

Room:Redondo - 2nd Floor

For Schedule, see page 18

Data Mining in Official Statistics Workshop

8:30 AM-10:00 AM

Room: San Diego - 2nd Floor

For Schedule, see page 17

Dynamic Network Analysis Workshop

8:30 AM-10:00 AM

Room:Oceanside - 2nd Floor

For Schedule, see page 18

Saturday, April 28

TS4

Tutorial Session: Privacy-Preserving Medical Data Sharing

8:30 AM-10:00 AM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

In this tutorial, we will first demonstrate the need for privacy-preserving medical data sharing by discussing analysis and mining tasks that disseminated data need to support, as well privacy threats that data sharing entails. Then, we will review privacy-preserving principles and algorithms that have been developed for sharing different types of medical data, including demographics, clinical and genomic data, and discuss a number of key issues, such as how data can be transformed to achieve both utility and privacy and how these two properties can be effectively balanced.

Aris Gkoulalas-Divanis
IBM Research, Zurich, Switzerland
Grigorios Loukides
Cardiff University, United Kingdom

Text Mining 2012 Workshop

8:50 AM-10:00 AM

Room:Pacific Ballroom B - 1st Floor

For Schedule, see page 15-16

Coffee Break

10:00 AM-10:30 AM

Room:Pacific Ballroom Foyer - 1st Floor

Saturday, April 28

Dynamic Network Analysis Workshop, continued

10:30 AM-12:00 PM

Room:Oceanside - 2nd Floor

For Schedule, see page 18

MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings, continued

10:30 AM-12:00 PM

Room:Redondo - 2nd Floor

For Schedule, see page 18

Broadening Participation in Data Mining Workshop, continued

10:30 AM-11:30 AM

Room:Pacific Ballroom A - 1st Floor

For Schedule, see page 14

Text Mining 2012 Workshop, continued

10:30 AM-12:00 PM

Room:Pacific Ballroom B - 1st Floor

For Schedule, see page 15-16

Saturday, April 28

TS4

Tutorial Session: Privacy-Preserving Medical Data Sharing, continued

10:30 AM-12:00 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

In this tutorial, we will first demonstrate the need for privacy-preserving medical data sharing by discussing analysis and mining tasks that disseminated data need to support, as well privacy threats that data sharing entails. Then, we will review privacy-preserving principles and algorithms that have been developed for sharing different types of medical data, including demographics, clinical and genomic data, and discuss a number of key issues, such as how data can be transformed to achieve both utility and privacy and how these two properties can be effectively balanced.

Aris Gkoulalas-Divanis
IBM Research, Zurich, Switzerland

Grigorios Loukides
Cardiff University University, United Kingdom

Saturday, April 28

Data Mining in Official Statistics Workshop, continued

10:40 AM-12:00 PM

Room: San Diego - 2nd Floor

For Schedule, see page 17

Lunch Break

12:00 PM-1:30 PM

Attendees on their own

Broadening Participation in Data Mining Workshop, continued

1:00 PM-3:00 PM

Room:Pacific Ballroom A - 1st Floor

For Schedule, see page 14

Analytics for Cyber-Physical Systems Workshop

1:00 PM-3:00 PM

Room: San Diego - 2nd Floor

For Schedule, see page 19

Text Mining 2012 Workshop, continued

1:30 PM-3:00 PM

Room:Pacific Ballroom B - 1st Floor

For Schedule, see page 15-16

Saturday, April 28

TS5

Tutorial Session: How to do Good Research and Get it Published in Top Venues

1:30 PM-3:00 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

In this tutorial Dr. Keogh will demonstrate some simple ideas to enhance the probability of success in getting your paper published in a top conference such as SDM , ICDM or SIGKDD. These tips and tricks are based on 13 years experience as a prolific author and reviewer, and wisdom solicited from many of the most prolific researchers/reviewers of the last decade.

Eamonn Keogh
University of California, Riverside, USA

Coffee Break

3:00 PM-3:30 PM

Room:Pacific Ballroom Foyer - 1st Floor

Saturday, April 28

Analytics for Cyber-Physical Systems Workshop,continued

3:20 PM-4:50 PM

Room: San Diego - 2nd Floor

For Schedule, see page 19

Broadening Participation in Data Mining Workshop, continued

3:30 PM-5:30 PM

Room:Pacific Ballroom A - 1st Floor

For Schedule, see page 14

Text Mining 2012 Workshop, continued

3:30 PM-4:30 PM

Room:Pacific Ballroom B - 1st Floor

For Schedule, see page 15-16

Saturday, April 28

TS5

Tutorial Session: How to do Good Research and Get it Published in Top Venues, continued

3:30 PM-5:00 PM

Room: Santa Monica - 2nd Floor

Chair: Nitesh Chawla, University of Notre Dame, USA

In this tutorial Dr. Keogh will demonstrate some simple ideas to enhance the probability of success in getting your paper published in a top conference such as SDM , ICDM or SIGKDD. These tips and tricks are based on 13 years experience as a prolific author and reviewer, and wisdom solicited from many of the most prolific researchers/ reviewers of the last decade.

Eamonn Keogh
University of California, Riverside, USA

Notes

SDM12 Abstracts

2012 SIAM
International Conference
on **DATA MINING**

April 26-28, 2012

Disney's Paradise Pier Hotel
Anaheim, California, USA

Abstracts are printed as submitted by the authors.

IP1

Rapid Learning Systems to Improve Patient Outcomes and Control Health Costs

In this talk, I will briefly present data mining solutions that analyze millions of patient records, impacting three major areas in healthcare. These include automated quality measurement and decision-support from hospitals EMRs, computer-aided diagnosis systems to identify suspicious lesions on medical images, and rapid learning systems to develop predictive models for personalized medicine. The last is based on a first-of-kind rapid learning system: a Euro-US health IT network spanning cancer centers in 5 nations to learn personalized therapies for lung cancer. The majority of the talk will present case studies that illustrate some of the challenges unique to mining healthcare data, and identify a few promising areas for research. These include the breakdown of traditional assumptions inherent in most mining algorithms, learning from multi-source systems, and the development of predictive models for personalized medicine. We conclude with a glimpse of a more-efficient healthcare future, where treatment decisions are driven by evolving knowledge that is continuously mined from patient records collected in health systems all over the world.

Bharat Rao
SIEMENS Healthcare - Health Services
bharat.rao@siemens.com

IP2

Some Assembly Required: Organizing in the 21st Century

Recent advances in Web Science provide comprehensive digital traces of social actions, interactions, and transactions. These data provide an unprecedented exploratorium to model the socio-technical motivations for creating, maintaining, dissolving, and reconstituting into teams for research, business, or social causes. Using examples from research in team science and massively multiplayer online games, Contractor will argue that Network Science serves as the foundation for the development of social network theories and methods to help advance our ability to understand the emergence of effective teams. More importantly, he will argue that these insights will also enable effective teams by building a new generation of recommender systems that leverage our research insights on the socio-technical motivations for creating ties.

Noshir Contractor
Northwestern University
nosh@northwestern.edu

IP3

Cross-Domain Knowledge Transfer in Data Mining

In data mining, we often encounter situations where we have an insufficient amount of high-quality data in a target domain, but we may have plenty of auxiliary data in related domains. Transfer learning aims to exploit these additional data to improve the learning performance in the target domain. In this talk, I will give an overview on some recent advances in transfer learning for challenging data mining problems. I will present structural transfer-learning solutions under heterogeneous feature representations. I will also survey cross-domain transfer learning solutions in online recommendation, social media and social network mining. I will discuss some current limitations of cross-domain transfer learning and explore possible future directions.

Qiang Yang
Department of Computer Science,
Hong Kong University of Science
qyang@cse.ust.hk

IP4

Temporal Dynamics and Information Retrieval

Many digital resources, like the Web, are dynamic and ever-changing collections of information. However, most information retrieval tools developed for interacting with Web content, such as browsers and search engines, focus on a single static snapshot of the information. In this talk, I will present analyses of how Web content changes over time, how people re-visit Web pages over time, and how re-visitation patterns are influenced by changes in user intent and content. These results have implications for many aspects of information retrieval and management including crawling policy, ranking and information extraction algorithms, result presentation, and systems evaluation. I will describe a prototype that supports people in understanding how the information they interact with changes over time, and new retrieval models that incorporate features about the temporal evolution of content to improve core ranking. Finally, I will conclude with an overview of some general challenges that need to be addressed to fully incorporate temporal dynamics in information retrieval and information management systems.

Susan Dumais
Microsoft Research
sdumais@microsoft.com

CP1

Sparse Group Lasso: Consistency and Climate Applications

We address the challenge of designing statistical predictive models for climate data that promote *structured sparsity*. We prove theoretical statistical consistency of estimators with *tree-structured* norm regularizers. We consider one particular model, the *Sparse Group Lasso* (SGL), to construct predictors of land climate using ocean climate variables. Our experimental results demonstrate that the SGL model provides better predictive performance than the current state-of-the-art, remains climatologically interpretable, and is robust in its variable selection.

Soumyadeep Chatterjee, Karsten Steinhaeuser
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
chat0129@umn.edu, ksteinha@umn.edu

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

Snigdhansu Chatterjee
School of Statistics
University of Minnesota, Twin Cities
chatterjee@stat.umn.edu

Auroop Ganguly
Northeastern University
Boston, MA
a.ganguly@neu.edu

CP1

Drought Detection of the Last Century: An Mrf-Based Approach

Droughts are one of the most damaging climate-related hazards. The late 1960s Sahel drought in Africa and the North American Dust Bowl of the 1930s are two examples of severe droughts that have an impact on society and the environment. Due to the importance of understanding droughts, we consider the problem of their detection based on gridded datasets of precipitation. We formulate the problem as the one of finding the most likely configuration of a Markov Random Field and propose an efficient inference algorithm. We apply this algorithm to the Climate Research Unit precipitation dataset spanning 106 years. The empirical results show that the algorithm successfully identifies the major droughts of the twentieth century in different regions of the world.

Qiang Fu
University of Minnesota, Twin Cities
qifu@cs.umn.edu

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

Stefan Liess, Peter Snyder
University of Minneosta, Twin Cities
liess@umn.edu, pksnyder@umn.edu

CP1

Toward Data-Driven, Semi-Automatic Inference of Phenomenological Physical Models: Application to Eastern Sahel Rainfall

First-principles based predictive understanding of complex, dynamic physical phenomena, such as regional precipitation, is quite limited due to the lack of complete phenomenological models underlying their physics. We propose a methodology for *data-driven*, *semi-automatic* inference of plausible phenomenological models and apply it to derive the model for eastern Sahel rainfall variability. To the best of our knowledge, this is the first model of this phenomenon; several of its components are consistent with the known evidence.

Saurabh V. Pendse
North Carolina State University
Oak Ridge National Laboratory
svpendse@ncsu.edu

Isaac Tetteh, Fredrick Semazzi
North Carolina State University
itetteh@ncsu.edu, fred_semazzi@ncsu.edu

Vipin Kumar
University of Minnesota
kumar@cs.umn.edu

Nagiza Samatova
North Carolina State University
Oak Ridge National Laboratory
samatova@csc.ncsu.edu

CP1

Detecting and Tracking Coordinated Groups in Dense, Systematically Moving, Crowds

We address the problem of detecting and tracking clusters of moving objects in very noisy environments. Monitoring a crowded football stadium for small groups of individuals acting suspiciously is an example instance of this problem. In this example the vast majority of individuals are not part of a suspicious group and are considered as noise. Existing spatio-temporal cluster algorithms are only capable of detecting small clusters in extreme noise when the noise objects are moving randomly. In reality, including the example cited, the noise objects move more systematically instead of moving randomly. The members of the suspicious groups attempt to mimic the behaviors of the crowd in order to blend in and avoid detection. This significantly exacerbates the problem of detecting the true clusters. We propose the use of Support Vector Machines (SVMs) to differentiate the true clusters and their members from the systematically moving noise objects. Our technique utilizes the relational history of the moving objects, implicitly tracked in a relationship graph, and a SVM to increase the accuracy of the clustering algorithm. A modified DB-SCAN algorithm is then used to discover clusters of highly related objects from the relationship graph. We evaluate our technique experimentally on several data sets of mobile objects. The experiments show that our technique is able to accurately and efficiently identify groups of suspicious individuals in dense crowds.

James C. Rosswog, Kanad Ghose
Binghamton University
jim.rosswog@binghamton.edu, ghose@cs.binghamton.edu

CP1

Large-Scale Nonparametric Estimation of Vehicle Travel Time Distributions

Fitting distributions of travel-time in vehicle traffic is an important application of spatio-temporal data mining. While regression methods to forecast the expected travel-time are standard approaches of travel-time prediction, we need to estimate distributions of the travel-time when using state-of-the-art risk-sensitive route recommendation systems. The authors introduce a novel nonparametric density estimator of travel-time for each road or link. The new estimator consists of basis functions modeled as mixtures of gamma or log-normal density functions, a sparse link similarity matrix given as an approximate diffusion kernel on a link connectivity graph, and importance weights for each link. Unlike the existing nonparametric methods that are computationally intensive, the new estimator is stably applicable to large datasets, because the basis functions and the importance weights are globally optimized with a fast convex clustering algorithm. Experimental results using real probe-car datasets show advantages of the new nonparametric estimator over parametric regression methods.

Rikiya Takahashi, Takayuki Osogami, Tetsuro Morimura
IBM Research - Tokyo
rikiya@jp.ibm.com, osogami@jp.ibm.com, tetsuro@jp.ibm.com

CP2

The Multi-Set Stream Clustering Problem

The problem of clustering has been widely studied by the data mining community because of its applications to a wide variety of problems in the context of customer segmentation, electronic commerce and learning. In general, the problem of clustering is generally presented as one of clustering *individual instances* of data records. In many applications, we have a collection of multiple *sets* of records. Each such set is essentially a database of records, and each database may possibly contain a different number of records. It is desirable to cluster these sets on the basis of the *similarity of underlying data distribution*. Thus, this problem may also be understood as that of clustering sets of data sets, as opposed to clustering sets of instances. The problem is especially challenging when the data sets are not available at one time, but are presented in the form of out-of-order and mixed streams, in which the records from different data sets do not arrive in any particular order, but are mixed with one another. In this paper, we present a first approach to the problem with the use of anchor-based summarization. We present experimental results for the effectiveness and efficiency of the approach on a number of real data sets.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

CP2
Cluster-Aware Compression with Provable K-Means Preservation

This work rigorously explores the design of cluster-preserving compression schemes for high-dimensional data. We focus on the K-means algorithm and identify conditions under which running the algorithm on the compressed data yields the same clustering outcome as on the original. The compression is performed using single and multi-bit minimum mean square error quantization schemes as well as a given clustering assignment of the original data. We provide theoretical guarantees on post-quantization cluster preservation under certain conditions on the cluster structure, and propose an additional data transformation that can ensure cluster preservation unconditionally; this transformation is invertible and thus induces virtually no distortion on the compressed data. In addition, we provide an efficient scheme for multi-bit allocation, per cluster and data dimension, which enables a trade-off between high compression efficiency and low data distortion. Our experimental studies highlight that the suggested scheme accurately preserved the clusters formed in all cases, while incurring minimal distortion on the data shapes. Our results can find many applications, e.g., in a) clustering, analysis and distribution of massive datasets, where the proposed data compression can boost performance while providing provable guarantees on the clustering result, as well as, in b) cloud computing services, as the optional transformation provides a data-hiding functionality in addition to preserving the K-means clustering outcome.

Nikolaos Freris, Michail Vlachos, Deepak Turaga
IBM Research
nif@zurich.ibm.com, michalis0@gmail.com,
turaga@us.ibm.com

CP2
Supervised Clustering of Label Ranking Data

This paper studies supervised clustering of label ranking data. Potential applications include target marketing, where the goal is to cluster customers in feature space by taking into consideration the assigned, potentially incomplete product preferences. We establish several heuristic baselines and propose a principled algorithm based on the Plackett-Luce ranking model specifically tailored for this type of clustering. Experimental evaluation on synthetic and real-life data showed that the PL-based method was superior to the baseline approaches.

Mihajlo Grbovic, Nemanja Djuric, Slobodan Vucetic
Temple University
mihajlo.grbovic@temple.edu, nemanja.djuric@temple.edu,
slobodan.vucetic@temple.edu

CP2
Symmetric Nonnegative Matrix Factorization for Graph Clustering

We offer conceptual understanding for the capabilities and shortcomings of nonnegative matrix factorization (NMF) as a clustering method, and propose Symmetric NMF (SymNMF) as a general framework for graph clustering. SymNMF finds a nonnegative symmetric factorization of a matrix containing pairwise similarity values. We then explain why SymNMF captures cluster structures in graph more naturally than spectral clustering. Promising experiment results with Newton-like algorithms are shown using artificial graph data, text data, and image data.

Da Kuang
Georgia Institute of Technology
da.kuang@cc.gatech.edu

Chris Ding
University of Texas at Arlington
chqding@uta.edu

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

CP2
Stratification Based Hierarchical Clustering Over a Deep Web Data Source

This paper focuses on the problem of clustering data from a *hidden* or a deep web data source. A key characteristics of deep web data sources is that data can only be accessed through the limited *query interface* they support. Because the underlying data set cannot be accessed directly, data mining must be performed based on sampling of the datasets. The samples, in turn, can only be obtained by querying the deep web databases with specific inputs. Unlike existing sampling based methods, sampling costs, and not the computation or memory costs, are the dominant consideration in designing the technique for sampling. We have developed a new methodology for addressing the clustering problem on the deep web. Our work includes three new ideas, which are a method for stratifying a deep web data source, an algorithm for hierarchical clustering based on stratified sampling, and a two phase technique for sampling, which includes a representative sampling in the first phase, and sampling focusing on the boundary points between the clusters in the second phase. We have evaluated our approach using two synthetic and one real data set. Our experiments show that each of the three ideas we have introduced leads to significant improvements in accuracy and efficiency of clustering a hidden data source. Specifically, we improve the accuracy of the clusters obtained (measured by average distance to centers) by up to 20% over the existing approach. Compared in another way, our method can achieve the same accuracy with up to 25%

fewer samples, thus reducing the sampling cost.

Tantan Liu, Gagan Agrawal
The Ohio State University
liut@cse.ohio-state.edu, agrawal@cse.ohio-state.edu

CP3
Multi-Skill Collaborative Teams Based on Densest Subgraphs

We consider the problem of identifying a team of skilled individuals for collaboration in the presence of a social network with the goal of maximizing collaborative compatibility of the team. The collaborative compatibility is measured as the density of the induced subgraph on selected nodes. We present a 3-approximation algorithm for the single-skill team formation problem and a special case of multiple skills. Our experiments show that these algorithms outperform previous work on several metrics.

Amita Gajewar
Yahoo! Labs, Yahoo! Inc., Santa Clara, CA
amitag@yahoo-inc.com

Atish Das Sarma
Google Research, Google Inc.,Mountain View, CA, USA
dassarma@google.com

CP3
Evaluating Event Credibility on Twitter

Given a set of popular Twitter events (with related users and tweets), we study the problem of automatically assessing credibility of such events. We propose a PageRank-like credibility analysis approach using a multi-typed network of events, tweets, and users. Further, event credibility scores are updated using event graph-based optimization, within each iteration. Experiments on events extracted from millions of tweets show that our methods perform significantly better ($\sim 86\%$) than classifier approach ($\sim 72\%$).

Manish Gupta, Peixiang Zhao
Univ of Illinois at Urbana-Champaign
manishg.iitb@gmail.com, pzhao4@illinois.edu

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@cs.uiuc.edu

CP3
Har: Hub, Authority and Relevance Scores in Multi-Relational Data for Query Search

In this paper, we propose a framework HAR to study the hub and authority scores of objects, and the relevance scores of relations in multi-relational data for query search. The basic idea of our framework is to consider a random walk in multi-relational data, and study in such random walk, limiting probabilities of relations for relevance scores, and of objects for hub scores and authority scores. The main contribution of this paper is to (i) propose a framework (HAR) that can compute the hub, authority and relevance scores by solving limiting probabilities arising from multi-relational data, and can incorporate input query vectors to handle query-specific search; (ii) show existence and uniqueness of such limiting probabilities so that they can be used for query search effectively; and (iii) develop an iterative algorithm to solve a set of tensor (multivariate polynomial) equations to obtain such probabilities. Extensive experimental results on TREC and DBLP data sets suggest that the proposed method is very effective in obtaining relevant results to the querying inputs. In the comparison, we find that the performance of HAR is better than those of HITS, SALSA and TOPHITS.

Xutao Li
Harbin Institute of Technology, China
xutaolee08@gmail.com

Michael K. Ng
Department of Mathematics, Hong Kong Baptist University
mng@math.hkbu.edu.hk

YUNMING Ye
Harbin Institute of Technology, China
yeyunming@hit.edu.cn

CP3
Feature Selection with Linked Data in Social Media

Feature selection is widely used in preparing high-dimensional data for effective data mining. Increasingly popular social media data presents new challenges to feature selection. Social media data consists of (1) traditional high-dimensional, attribute-value data such as posts, tweets, comments, and images, and (2) linked data that describes the relationships between social media users as well as who post the posts, etc. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the already challenging problem of feature selection. In this paper, we illustrate the differences between attribute-value data and social media data, investigate if linked data can be exploited in a new feature selection framework by taking advantage of social science theories, extensively evaluate the effects of user-user and user-post relationships manifested in linked data on feature selection, and discuss some research issues for future work.

Jiliang Tang
Arizona State University
ARIZONA STATE UNIVERISTY
Jiliang.Tang@asu.edu

Huan Liu
Arizona State University
huanliu@asu.edu

CP3
Microscopic Social Influence

Social influences, the phenomena that one individual's actions can induce similar behaviors among his/her friends via their social ties, have been observed prevailingly in socially networked systems. While most existing work focuses on studying general, macro-level influence (e.g., diffusion); equally important is to understand social influence at *microscopic* scales (i.e., at the granularity of single individuals, actions, and time-stamps), which may benefit a range of applications. We propose μSI, a microscopic social-influence model wherein: individuals' actions are modeled as temporary interactions between social network (formed by individuals) and object network (formed by targets of actions); one individual's actions influence his/her friends in a dynamic, network-wise manner (i.e., dependent on both social and object networks). We develop for μSI a suite of novel inference tools that enable to answer questions of the form: How may an occurred interaction trigger another? More importantly, when and where may a new interaction be observed? We carefully address the computational challenges for inferencing over such semantically rich models by dynamically identifying sub-domains of interest and varying the precision of solutions over different sub-domains. We demonstrate the breadth and generality of μSI using two seemingly disparate applications. In the context of social tagging service, we show how it can help improve the accuracy and freshness of resource recommendation; in the context of mobile phone call service, we show how it can help improve the efficiency of paging operation.

Ting Wang
IBM T.J. Watson Research Center
tingwang@us.ibm.com

Mudhakar Srivatsa, Dakshi Agrawal
IBM Research
msrivats@us.ibm.com, agrawal@us.ibm.com

Ling Liu
Georgia Tech
lingliu@cc.gatech.edu

CP4

Heterogeneous Data Fusion Via Space Alignment Using Nonmetric Multidimensional Scaling

This paper aims to align heterogeneous data spaces into one common space, which makes it possible to analyze relationships between them. We propose a novel graph embedding framework and one such method based on nonmetric multidimensional scaling (NMDS). The NMDS criteria using distance rank orders effectively handles both the deformation of original spaces and the alignment between them. Experimental results show its advantages over existing methods using multi-lingual data and document-speech data.

Jaegul Choo
College of Computing
Georgia Institute of Technology
joyfull@cc.gatech.edu

Shawn Bohn, Grant Nakamura, Amanda White
Pacific Northwest National Laboratory
shawn.bohn@pnl.gov, grant.nakamura@pnl.gov,
amanda.white@pnl.gov

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

CP4

A Bayesian Nonparametric Joint Factor Model for Learning Shared and Individual Subspaces from Multiple Data Sources

Joint analysis of multiple data sources is becoming increasingly popular in transfer learning, multi-task learning and cross-domain data mining. One promising approach to model the data jointly is through learning the shared and individual factor subspaces. However, performance of this approach depends on the subspace dimensionalities and the level of sharing needs to be specified a priori. To this end, we propose a nonparametric joint factor analysis framework for modeling multiple related data sources. Our model utilizes the hierarchical beta process as a nonparametric prior to automatically infer the number of shared and individual factors. For posterior inference, we provide a Gibbs sampling scheme using auxiliary variables. The effectiveness of the proposed framework is validated through its application on two real world problems – transfer learning in text and image retrieval.

Sunil K. Gupta
Deakin University, Geelong Waurn Ponds Campus,
Australia
sunil.gupta@deakin.edu.au

Dinh Phung, Svetha Venkatesh
Deakin University, Geelong Waurn Ponds Campus
Victoria, Australia
dinh.phung@deakin.edu.au,
svetha.venkatesh@deakin.edu.au

CP4

Adaptive Multi-Task Sparse Learning with An Application to Fmri Study

In this paper, we propose two adaptive multi-task learning methods, adaptive multi-task lasso and adaptive multi-task elastic-net. Both of them can simultaneously conduct model estimation and variable selection across different tasks. Under weak assumptions, we establish the asymptotic oracle property. As a case study, we apply the adaptive multi-task elastic-net to a cognitive science problem, where one wants to discover a compact semantic basis for predicting fMRI images. We show that the adaptive multi-task sparse learning method achieves superior performance and provides some insights into how the brain represents the meanings of words.

Xi Chen
Carnegie Mellon University
School of Computer Science
xichen@cs.cmu.edu

Jingrui He, Rick Lawrence
IBM T.J. Watson Research Center
jingruhe@us.ibm.com, ricklawr@us.ibm.com

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

CP4

Heterogeneous
Datasets Representation and Learning using Diffusion Maps and Laplacian Pyramids

The diffusion maps and geometric harmonics provide a method for describing and extending the geometry of high dimensional datasets. These methods suffers from two limitations: First, the assumption that the attributes of the processed dataset are comparable. Second, application of the geometric harmonics requires setting for the correct scale. We propose a method for learning heterogeneous datasets by using diffusion maps for unifying heterogeneous dataset and by replacing the geometric harmonics with Laplacian pyramid extensions.

Neta Rabin
Yale University
neta.rabin@yale.edu

CP4

Learning from Heterogeneous Sources Via Gradient Boosting Consensus

Multiple data sources containing different types of features may be available for a given task. For instance, users' profiles can be used to build recommendation systems. In addition, a model can also use users' historical behaviors and social networks to infer users' interests on related products. We argue that it is desirable to collectively use any available multiple heterogeneous data sources in order to build effective learning models. We call this framework *heterogeneous learning*. In our proposed setting, data sources can include (i) non-overlapping features, (ii) non-overlapping instances, and (iii) multiple networks (i.e. graphs) that connect instances. In this paper, we propose a general optimization framework for heterogeneous learning, and devise a corresponding learning model from gradient boosting. The idea is to minimize the empirical loss with two constraints: (1) There should be consensus among the predictions of overlapping instances (if any) from different data sources; (2) Connected instances in graph datasets may have similar predictions. The objective function is solved by stochastic gradient boosting trees. Furthermore, a weighting strategy is designed to emphasize informative data sources, and deemphasize the noisy ones. We formally prove that the proposed strategy leads to a tighter error bound. This approach consistently outperforms a standard concatenation of data sources on movie rating prediction, number recognition and terrorist attack detection tasks. We observe that the proposed model can improve out-of-sample error rate by as much as 80%.

Xiaoxiao Shi
Computer Department, University of Illinois at Chicago
xiao.x.shi@gmail.com

Jean-Francois Paiement, David Grangier
AT&T Labs
jpaiement@research.att.com, grangier@research.att.com

Philip Yu
University of Illinois at Chicago
psyu@cs.uic.edu

CP5

Marbles: Mining Association Rules Buried in Long Event Sequences

Episodes are sequential patterns that describe events that often occur in the vicinity of each other. In this paper we propose an algorithm that mines association rules between two episodes. We introduce two novel confidence measures for the rules, and aim to limit the output by eliminating redundant rules. We define the class of closed rules, a class that contains all non-redundant output. To make the algorithm efficient, we use pruning steps along the way.

Boris Cule, Nikolaj Tatti, Bart Goethals
University of Antwerp
boris.cule@ua.ac.be, nikolaj.tatti@ua.ac.be, bart.goethals@ua.ac.be

CP5

Class Relevant Pattern Mining in Output-Polynomial Time

The set of so-called relevant patterns is a subset of all itemsets particularly suited for pattern-based classification tasks. So far, no efficient algorithm has been developed for computing the set of relevant patterns: all existing solutions have a worst-case complexity which is exponential in the size of the input and output. In this paper, we investigate new properties of the relevant patterns and develop, thereupon, the first algorithm whose runtime is polynomial in the size of the input and output. As we show in the experimental section, this result is not only of theoretical interest but also of practical importance, often reducing the search space by orders of magnitude.

Henrik Grosskreutz
Fraunhofer IAIS
henrik.grosskreutz@iais.fraunhofer.de

CP5

Mining Patterns in Networks Using Homomorphism

In recent years many algorithms have been developed for finding patterns in graphs and networks. A disadvantage of these algorithms is that they use subgraph isomorphism to determine the support of a graph pattern; subgraph isomorphism is a well-known NP complete problem. In this paper, we propose an alternative approach which mines tree patterns in networks by using subgraph *homomorphism*. The advantage of homomorphism is that it can be computed in polynomial time, which allows us to develop an algorithm that mines tree patterns in arbitrary graphs in incremental polynomial time. Homomorphism however entails two problems not found when using isomorphism: (1) two patterns of different size can be equivalent; (2) patterns of unbounded size can be frequent. In this paper we formalize these problems and study solutions that easily fit within our algorithm.

Anton Dries
Universitat Pompeu Fabra
anton.dries@upf.edu

Siegfried Nijssen
Katholieke Universiteit Leuven
siegfried.nijssen@cs.kuleuven.be

CP5

Scalable Induction of Probabilistic Real-Time Automata Using Maximum Frequent Pattern Based Clustering

The paper presents a scalable method for learning probabilistic real-time automata (PRTAs), a new type of model that captures the dynamics of multi-dimensional event logs. In multi-dimensional event logs, events are described by several features instead of only one symbol. Moreover, it is not clear up front which events occur in an event log. The learning method to find a PRTA that models such an event log is based on the state merging of a prefix tree acceptor, which is guided by a clustering to determine the states of the automaton. To make the overall approach scalable, an online clustering method based on maximum frequent patterns (MFPs) is used. The approach is evaluated on a synthetic, a biological and a medical data set. The results show that the induction of automata using

MFP-based clustering gives easy to understand and stable automata, but most importantly, makes it scalable to large data sets.

Jana Schmidt
Institut fuer Informatik, TU Muenchen
jana.schmidt@in.tum.de

Sonja Ansorge
TU Muenchen
sonja.ansorge@gmx.de

Stefan Kramer
Johannes Gutenberg-Universitaet Mainz
kramer@informatik.uni-mainz.de

CP5
Slim: Directly Mining Descriptive Patterns

Mining small, useful, and high-quality sets of patterns has recently become an important topic in data mining. The standard approach is to first mine many candidates, and then to select a good subset. However, the pattern explosion generates such enormous amounts of candidates that by post-processing it is virtually impossible to analyse dense or large databases in any detail. We introduce Slim, an any-time algorithm for mining high-quality sets of itemsets directly from data. We use MDL to identify the best set of itemsets as that set that describes the data best. To approximate this optimum, we iteratively use the current solution to determine what itemset would provide most gain—estimating quality using an accurate heuristic. Without requiring a pre-mined candidate collection, Slim is parameter-free in both theory and practice. Experiments show we mine high-quality pattern sets; while evaluating orders-of-magnitude fewer candidates than our closest competitor, Krimp, we obtain much better compression ratios—closely approximating the locally-optimal strategy. Classification experiments independently verify we characterise data very well.

Koen Smets, Jilles Vreeken
Universiteit Antwerpen
koen.smets@ua.ac.be, jilles.vreeken@ua.ac.be

CP6
Beam Methods for the Profile Hidden Markov Model

The Profile Hidden Markov Model (PHMM) is commonly used to represent biological sequences. We present a method for transforming the Profile HMM into an equivalent standard HMM where each transition is associated with a single emission. Using this transformation, we develop a beam method, which includes a novel variational adaptation of the infinite-HMM beam sampling technique, to create a fast inference algorithm. We evaluate our algorithm on both synthetic data and protein sequence datasets, showing that our beam method can lead to considerable improvements in runtime while maintaining the model's ability to concisely represent sequences.

Samuel J. Blasiak, Huzefa Rangwala, Kathryn Laskey
George Mason University
sblasiak@gmu.edu, rangwala@cs.gmu.edu, klaskey@gmu.edu

CP6
Optimal Distance Estimation Between Compressed Data Series

Most real-world data contain repeated or periodic patterns. This suggests that they can be effectively represented and compressed using only a few coefficients of an appropriate complete orthogonal basis (e.g., Fourier, Wavelets, Karhunen Loeve expansion or Principal Components). In the face of ever increasing data repositories and given that most mining operations are distance-based, it is vital to perform accurate distance estimation directly on the compressed data. However, distance estimation when the data are represented using different sets of coefficients is still a largely unexplored area. This work studies the optimization problems related to obtaining the tightest lower/upper bound on the distance based on the available information. In particular, we consider the problem where a distinct set of coefficients is maintained for each sequence, and the L_2-norm of the compression error is recorded. We establish the properties of optimal solutions, and leverage the theoretical analysis to develop a fast algorithm to obtain an exact solution to the problem. The suggested solution provides the tightest provable estimation of the L_2-norm or the correlation, and executes at least two order of magnitudes faster than a numerical solution based on convex optimization. The contributions of this work extend beyond the purview of periodic data, as our methods are applicable to any sequential or high-dimensional data as well as to any orthogonal data transformation used for the underlying data compression scheme.

Nikolaos Freris, Michail Vlachos
IBM Research
nif@zurich.ibm.com, michalis0@gmail.com

Serdar Kozat
Koc University
Istanbul, Turkey
skozat@ku.edu.tr

CP6
Transformation Based Ensembles for Time Series Classification

Until recently, the vast majority of data mining time series classification (TSC) research has focused on alternative distance measures for 1-Nearest Neighbour (1-NN) classifiers based on either the raw data, or on compressions or smoothing of the raw data. Despite the extensive evidence in favour of 1-NN classifiers with Euclidean or Dynamic Time Warping distance, there has also been a flurry of recent research publications proposing classification algorithms for TSC. Generally, these classifiers describe different ways of incorporating summary measures in the time domain into more complex classifiers. Our hypothesis is that the easiest way to gain improvement on TSC problems is to simply transform into an alternative data space where the discriminatory features are more easily detected. To test our hypothesis, we perform a range of benchmarking experiments in the time domain, before evaluating nearest neighbour classifiers on data transformed into the power spectrum, the autocorrelation function, and the principal component space. We demonstrate that on some problems there is dramatic improvement in the accuracy of classifiers built on the transformed data over classifiers built in the time domain, but that there is also a wide variance in accuracy for a particular classifier built on different data transforms. To overcome this variability, we propose a simple transformation based ensemble, then demonstrate that

it improves performance and reduces the variability of classifiers built in the time domain only. Our advice to a practitioner with a real world TSC problem is to try transforms before developing a complex classifier; it is the easiest way to get a potentially large increase in accuracy, and may provide further insights into the underlying relationships that characterise the problem.

Anthony Bagnall, Luke Davis, Jon Hills, Jason Lines
University of East Anglia
Anthony.Bagnall@uea.ac.uk, luke.davis@uea.ac.uk, j.hills@uea.ac.uk, j.lines@uea.ac.uk

CP6
Mining Compressing Sequential Patterns

Compression based pattern mining has been successfully applied to many data mining tasks. We propose an approach based on the minimum description length principle to extract sequential patterns that compress a database of sequences well. We show that mining compressing patterns is NP-Hard and belongs to the class of inapproximable problems. We propose two heuristic algorithms to mining compressing patterns. The first uses a two-phase approach similar to Krimp for itemset data. To overcome performance with the required candidate generation we propose GoKrimp, an effective greedy algorithm that directly mines compressing patterns. We conduct an empirical study on six real-life datasets to compare the proposed algorithms by run time, compressibility, and classification accuracy using the patterns found as features for SVM classifiers.

Hoang Thanh Lam
TU Eindhoven
t.l.hoang@tue.nl

Fabian Moerchen, Dmitriy Fradkin
Siemens Corporation, Corporate Research
fabian.moerchen@siemens.com,
dmitriy.fradkin@siemens.com

Toon Calders
TU Eindhoven
t.calders@tue.nl

CP6
Simplex Distributions for Embedding Data Matrices over Time

Early stress recognition is of great relevance in precision plant protection. Pre-symptomatic water stress detection is of particular interest, ultimately helping to meet the challenge of “How to feed a hungry world?’. Due to the climate change, this is of considerable political and public interest. Due to its large-scale and temporal nature, e.g., when monitoring plants using hyperspectral imaging, and the demand of physical meaning of the results, it presents unique computational problems in scale and interpretability. However, big data matrices over time also arise in several other real-life applications such as stock market monitoring where a business sector is characterized by the ups and downs of each of its companies per year or topic monitoring of document collections. Therefore, we consider the general problem of embedding data matrices into Euclidean space over time without making any assumption on the generating distribution of each matrix. To do so, we represent all data samples by means of convex combinations of only few extreme ones computable in linear time. On the simplex spanned by the extremes, there are then natural candidates for distributions inducing distances between and in turn embeddings of the data matrices. We evaluate our method across several domains, including synthetic, text, and financial data as well as a large-scale dataset on water stress detection in plants with more than 3 billion matrix entries. The results demonstrate that the embeddings are meaningful and fast to compute. The stress detection results were validated by a domain expert and conform to existing plant physiological knowledge.

Kristian Kersting
Fraunhofer IAIS
kristian.kersting@iais.fraunhofer.de

Mirwaes Wahabzada
Fraunhofer IAIS
Sankt Augustin, Germany
mirwaes.wahabzada@iais.fraunhofer.de

Christoph Roemer
Institute of Geodesy and Geoinformation
University of Bonn, Germany
roemer@igg.uni-bonn.de

Christian Thurau
Fraunhofer IAIS, Sankt Augustin, Germany
christian.thurau@iais.fraunhofer.de

Agim Ballvora
Institute of Crop Science and Resource Conservation
Plant Breeding, University of Bonn, Germany
ballvora@uni-bonn.de

Uwe Rascher
FZ Jülich
u.rascher@fz-juelich.de

Jens Leon
Institute of Crop Science and Resource Conservation
Plant Breeding, University of Bonn, Germany
j.leon@uni-bonn.de

Christian Bauckhage
Fraunhofer IAIS
christian.bauckhage@iais.fraunhofer.de

Lutz Pluemer
Institute of Geodesy and Geoinformation
University of Bonn, Germany
pluemer@igg.uni-bonn.de

CP7
Bayesian Supervised Multilabel Learning with Coupled Embedding and Classification

We introduce a novel *Bayesian supervised multilabel learning* method that combines linear dimensionality reduction with linear binary classification. We present a deterministic variational approximation approach to learn the proposed probabilistic model for multilabel classification. Experiments show that the proposed approach achieves good performance values in terms of hamming loss, macro F_1, and micro F_1 on held-out test data. The low-dimensional embeddings obtained by our method are also very useful for exploratory data analysis.

Mehmet Gönen
Department of Information and Computer Science

Aalto University School of Science
mehmet.gonen@aalto.fi

CP7
Multi-Objective Multi-Label Classification

Multi-label classification refers to the task of predicting potentially multiple labels for a given instance. Conventional multi-label classification approaches focus on the single objective setting, where the learning algorithm optimizes over a single performance criterion (e.g. *Ranking Loss*) or a heuristic function. The basic assumption is that the optimization over one single objective can improve the overall performance of multi-label classification and meet the requirements of various applications. However, in many real applications, an optimal multi-label classifier may need to consider the tradeoffs among multiple conflicting objectives, such as minimizing *Hamming Loss* and maximizing *Micro F1*. In this paper, we study the problem of *multi-objective multi-label classification* and propose a novel solution (called Moml) to optimize over multiple objectives simultaneously. Note that optimization objectives may be conflicting, thus one cannot identify a single solution that is optimal on all objectives. Our Moml algorithm finds a set of *non-dominated solutions* which are optimal according to the different tradeoffs of the multiple objectives. So users can flexibly construct various combined predictive models from the solution set, which helps to provide more meaningful classification results in different application scenarios. Empirical studies on real-world tasks demonstrate that the Moml can effectively boost the overall performance of multi-label classification, not limiting to the optimization objectives.

Chuan Shi
Beijing University of Posts and Telecommunications
shichuan@bupt.edu.cn

Xiangnan Kong, Philip Yu
University of Illinois at Chicago
kongxn@gmail.com, psyu@cs.uic.edu

Bai Wang
Beijing University of Posts and Telecommunications
wangbai@bupt.edu.cn

CP7
A Distributed Kernel Summation Framework for General-Dimension Machine Learning

Kernel summations are a ubiquitous key computational bottleneck in many data analysis methods. We provide the first distributed implementation of kernel summation framework that can utilize: 1) various types of deterministic and probabilistic approximations; 2) any multi-dimensional binary utilizing both distributed and shared memory parallelism; 3) a dynamic load balancing scheme. We show scalability results for kernel density estimation on a subset of the Sloan Digital Sky Survey Data up to 6,144 cores.

Dongryeol Lee, Richard Vuduc, Alexander Gray
Georgia Institute of Technology
dongryel@cc.gatech.edu, richie@cc.gatech.edu, agray@cc.gatech.edu

CP7
Subtree Replacement in Decision Tree Simplification

The current availability of efficient algorithms for decision tree induction makes intricate post-processing techniques worth to be investigated both for efficiency and effectiveness. We study the simplification operator of subtree replacement, also known as *grafting*, originally implemented in the C4.5 system. We present a parametric bottom-up algorithm integrating grafting with the standard pruning operator, and analyze its complexity in terms of the number of nodes visited. Immediate instances of the parametric algorithm include extensions of error based, reduced error, minimum error, and pessimistic error pruning. Experimental results show that the computational cost of grafting is paid off by statistically significant smaller trees without accuracy loss.

Salvatore Ruggieri
Dipartimento di Informatica, Universita' di Pisa
ruggieri@di.unipi.it

CP7
Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information

We propose a new matrix completion algorithm—Kernelized Probabilistic Matrix Factorization (KPMF), which effectively incorporates external side information into the matrix factorization process. Unlike Probabilistic Matrix Factorization (PMF), which assumes an independent latent vector for each row (and each column) with Gaussian priors, KMPF works with latent vectors spanning all rows (and columns) with Gaussian Process (GP) priors. Hence, KPMF explicitly captures the underlying (nonlinear) covariance structures across rows and columns. This crucial difference greatly boosts the performance of KPMF when appropriate side information, e.g., users' social network in recommender systems, is incorporated. We demonstrate the efficacy of KPMF through two different applications: 1) recommender systems and 2) image restoration.

Tinghui Zhou
Carnegie Mellon University
tinghuiz@cmu.edu

Hanhuai Shan
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
shan@cs.umn.edu

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

Guillermo Sapiro
University of Minnesota
Dept Electrical & Computer Engineering
guille@umn.edu

CP8
On Dynamic Link Inference in Heterogeneous Networks

Network and linked data have become quite prevalent in recent years because of the ubiquity of the web and social media applications, which are inherently network oriented. Such networks are massive, dynamic, contain a lot of content, and may evolve over time in terms of the underlying structure. In this paper, we will study the problem of dynamic link inference in *temporal* and *heterogeneous* information networks. The problem of dynamic link inference is extremely challenging in massive and heterogeneous information network because of the challenges associated with the dynamic nature of the network, and the different types of nodes and attributes in it. Both the topology and type information need to be used effectively for the link inference process. We propose an effective two-level scheme which makes efficient macro- and micro-decisions for combining structure and content in a *dynamic and time-sensitive* way. The time-sensitive nature of the links is leveraged in order to perform effective link prediction. We illustrate the effectiveness of our technique over a number of real data sets.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Yan Xie, Philip Yu
University of Illinois at Chicago
yxie8@uic.edu, psyu@cs.uic.edu

CP8

Parameter-Free Identification of Cohesive Subgroups in Large Attributed Graphs

Given a graph with node attributes, how can we find meaningful patterns such as clusters, bridges, and outliers? Attributed graphs appear in real world in the form of social networks with user interests, gene interaction networks with gene expression information, phone call networks with customer demographics, and many others. In effect, we want to group the nodes into clusters with similar connectivity and homogeneous attributes. Most existing graph clustering algorithms either consider only the connectivity structure of the graph and ignore the node attributes, or require several user-defined parameters such as the number of clusters. We propose PICS, a novel, parameter-free method for mining *attributed graphs*. Two key advantages of our method are that (1) it requires *no* user-specified parameters such as the number of clusters and similarity functions, and (2) its running time scales *linearly* with total graph and attribute size. Our experiments show that PICS reveals meaningful and insightful patterns and outliers in both synthetic and real data sets, including call networks, political books, political blogs, and collections from Twitter and YouTube which have more than 70K nodes and 30K attributes.

Leman Akoglu
Carnegie Mellon University
lakoglu@cs.cmu.edu

Hanghang Tong
IBM T.J. Watson
htong@us.ibm.com

Brendan Meeder, Christos Faloutsos
Carnegie Mellon University
bmeeder@cs.cmu.edu, christos@cs.cmu.edu

CP8

Structural Analysis in Multi-Relational Social Networks

Modern social networks often consist of multiple relations among individuals. Understanding the structure of such multi-relational network is essential. In sociology, one way of structural analysis is to identify different positions and roles using blockmodels. In this paper, we generalize stochastic blockmodels to *Generalized Stochastic Blockmodels* (GSBM) for performing positional and role analysis on multi-relational networks. Our GSBM generalizes many different kinds of *Multivariate Probability Distribution Function* (MVPDF) to model different kinds of multi-relational networks. In particular, we propose to use *multivariate Poisson distribution* for multi-relational social networks.

Bing Tian Dai, Freddy Chua, Ee-Peng Lim
Singapore Management University
btdai@smu.edu.sg, freddy.chua.2009@phdis.smu.edu.sg, eplim@smu.edu.sg

CP8

Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model

In many real-world situations, different and often opposite opinions, innovations, or products are competing with one another for their social influence in a networked society. In this paper, we study competitive influence propagation in social networks under the competitive linear threshold (CLT) model, an extension to the classic linear threshold model. Under the CLT model, we focus on the problem that one entity tries to block the influence propagation of its competing entity as much as possible by strategically selecting a number of seed nodes that could initiate its own influence propagation. We call this problem the influence blocking maximization (IBM) problem. We prove that the objective function of IBM in the CLT model is submodular, and thus a greedy algorithm could achieve $1 - 1/e$ approximation ratio. However, the greedy algorithm requires Monte-Carlo simulations of competitive influence propagation, which makes the algorithm not efficient. We design an efficient algorithm CLDAG, which utilizes the properties of the CLT model, to address this issue. We conduct extensive simulations of CLDAG, the greedy algorithm, and other baseline algorithms on real-world and synthetic datasets. Our results show that CLDAG is able to provide best accuracy in par with the greedy algorithm and often better than other algorithms, while it is two orders of magnitude faster than the greedy algorithm.

Xinran He, Guojie Song
Peking University
xinranhe@pku.edu.cn, gjsong@pku.edu.cn

Wei Chen
Microsoft Research Asia
weic@microsoft.com

Qingye Jiang
Columbia University
qj2116@columbia.edu

CP8

A Framework for the Evaluation and Management of Network Centrality

Network-analysis literature is rich in node-centrality measures that quantify the centrality of a node as a function of the (shortest) paths of the network that go through it. Existing work focuses on defining instances of such measures and designing algorithms for the specific combinatorial problems that arise for each instance.In this work, we propose a unifying definition of centrality that subsumes all path-counting based centrality definitions: e.g., stress, betweenness or paths centrality. We also define a generic algorithm for computing this generalized centrality measure for every node and every group of nodes in the network. Next, we define two optimization problems: k-Group Centrality Maximization and k-Edge Centrality Boosting. In the former, the task is to identify the subset of k nodes that have the largest group centrality. In the latter, the goal is to identify up to k edges to add to the network so that the centrality of a node is maximized. We show that both of these problems can be solved efficiently for arbitrary centrality definitions using our general framework. In a thorough experimental evaluation we show the practical utility of our framework and the efficacy of our algorithms.

Vatche Ishakian, Dora Erdos, Evimaria Terzi, Azer Bestavros
Boston University
visahak@bu.edu, edori@cs.bu.edu, evimaria@cs.bu.edu, best@cs.bu.edu

CP9

Feature Selection "Tomography' — Illustrating That Optimal Feature Filtering Is Hopelessly Ungeneralizable

Feature filtering methods are used in high-dimensional domains to quickly score each feature independently. We provide a new empirical method to reveal the *feature preference surface* for a given situation. This visualization reveals new insights for feature filtering: (a) Existing functions do not match the surfaces we revealed. (b) The shape of the surfaces varies and depends on more factors than have been studied at once in the existing literature in feature filtering.

George Forman
HP Labs
ghforman@hpl.hp.com

CP9

A Bayesian Markov-Switching Model for Sparse Dynamic Network Estimation

Inferring Dynamic Bayesian Networks (DBNs) from multivariate time series data is a key step towards the understanding of complex systems as it reveals important dependency relationship underlying such systems. Most of the traditional approaches assume a "static' DBN. Yet in many relevant applications, such as those arising in biology and social sciences, the dependency structures may vary over time. In this paper, we introduce a sparse Markov-switching vector autoregressive model to capture the structural changes in the dependency relationships over time. Our approach accounts for such structural changes via a set of latent state variables, which are modeled by a discrete-time discrete-state Markov process. Assuming that the underlying structures are sparse, we estimate the networks at each state through the hierarchical Bayesian group Lasso, so as to efficiently capture dependencies with lags greater than one time unit. For computation, we develop an efficient algorithm based on the Expectation-Maximization method. We demonstrate the strength of our approach through simulation studies and a real data set concerning climate change.

Huijing Jiang
IBM T.J. Watson Research Center
huijiang@us.ibm.com

Aurelie Lozano
IBM Research
T. J. Watson Research Center
aclozano@us.ibm.com

Fei Liu
IBM T.J. Watson Research Center
feiliu@us.ibm.com

CP9

Feature Selection over Distributed Data Streams through Optimization

Monitoring data streams in a distributed system has attracted considerable interest in recent years. The task of feature selection (e.g., by monitoring the information gain of various features) requires a very high communication overhead when addressed using straightforward centralized algorithms. While most of the existing algorithms deal with monitoring simple aggregated values such as frequency of occurrence of stream items, motivated by recent contributions based on geometric ideas we present an alternative approach. The proposed approach enables monitoring values of an arbitrary threshold function over distributed data streams through constraints applied separately on each stream. We report numerical experiments on a real–world data that detect instances where communication between nodes is required, and compare the approach and the results to those recently reported in the literature.

Jacob Kogan
umbc
kogan@umbc.edu

CP9

Sampling Strategies to Evaluate the Performance of Unknown Predictors

The focus of this paper is on how to select a small sample of examples for labeling that can help us to evaluate many different classification models unknown at the time of sampling. We are particularly interested in studying the sampling strategies for problems in which the prevalence of the two classes is highly biased toward one of the classes. The evaluation measures of interest we want to estimate as accurately as possible are those obtained from the contingency table. We provide a careful theoretical analysis on sensitivity, specificity, and precision and show how sampling strategies should be adapted to the rate of skewness in data in order to effectively compute the three aforementioned evaluation measures.

Hamed Valizadegan, Saeed Amizadeh, Milos Hauskrecht
University of Pittsburgh
hamed@cs.pitt.edu, saeed@cs.pitt.edu, milos@cs.pitt.edu

CP9

Learning Hierarchical Relationships among Partially Ordered Objects with Heterogeneous At-

tributes and Links

Objects linking with many other objects in an information network may imply various semantic relationships. In this work we study a generic form of relationship along which objects can form a tree-like structure, a pervasive structure in various domains. We formalize the problem of uncovering hierarchical relationships in a supervised setting. We propose a discriminative undirected graphical model which integrates a wide range of features and rules by defining potential functions with simple forms.

Chi Wang
University of Illinois at Urbana-Champaign
chiwang1@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

Qi Li, Xiang Li, Wen-Pin Lin, Heng Ji
City University of New York
liqiearth@gmail.com, jackieiuu729@gmail.com, danniellin@gmail.com, hengjicuny@gmail.com

CP10
Transfer Learning of Distance Metrics by Cross-Domain Metric Sampling Across Heterogeneous Spaces

In this paper, we examine a new angle to the transfer learning problem, where we examine the problem of distance function learning. Specifically, we focus on the problem of how our knowledge of distance functions in one domain can be transferred to a new domain. A good semantic understanding of the feature space is critical in providing the domain specific understanding for setting up good distance functions. Unfortunately, not all domains have feature representations which are equally interpretable. For example, in some domains such as text, the semantics of the feature representation are clear, as a result of which it is easy for a domain expert to set up distance functions for specific kinds of semantics. In the case of image data, the features are semantically harder to interpret, and it is harder to set up distance functions, especially for particular semantic criteria. In this paper, we focus on the problem of transfer learning as a way to close the semantic gap between different domains, and show how to use correspondence information between two domains in order to set up distance functions for the semantically more challenging domain.

Guo-Jun Qi
Beckman Institute
University of Illinois at Urbana-Champaign
qi4@illinois.edu

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Thomas Huang
University of Illinois at Urbana-Champaign
huang@ifp.uiuc.edu

CP10
Transfer Topic Modeling with Ease and Scalability

The increasing volume of *short* texts generated on social media sites, such as Twitter or Facebook, creates a great demand for effective and efficient topic modeling approaches. While latent Dirichlet allocation (LDA) can be applied, it is not optimal due to its weakness in handling short texts with fast-changing topics and scalability concerns. In this paper, we propose a transfer learning approach that utilizes abundant labeled documents from other domains (such as Yahoo! News or Wikipedia) to improve topic modeling, with better model fitting and result interpretation. Specifically, we develop *Transfer Hierarchical* LDA (thLDA) model, which incorporates the label information from other domains via informative priors. In addition, we develop a parallel implementation of our model for large-scale applications. We demonstrate the effectiveness of our thLDA model on both a microblogging dataset and standard text collections including AP and RCV1 datasets.

Jeon-Hyung Kang, Jun Ma, Yan Liu
University of Southern California
jeonhyuk@usc.edu, junma@usc.edu, yanliu@usc.edu

CP10
Dual Transfer Learning

In this paper, we propose a novel approach, Dual Transfer Learning (DTL), which simultaneously learns the marginal and conditional distributions, and exploits the duality between them in a principled way. The key idea behind DTL is that learning one distribution can help to learn the other. This duality property leads to mutual reinforcement when adapting both distributions across domains to transfer knowledge. Experiments demonstrate the effectiveness of our proposed approach.

Mingsheng Long
Department of Computer Science and Technology
Tsinghua University
longmingsheng@gmail.com

Jianmin Wang, Guiguang Ding
School of Software
Tsinghua University
jimwang@tsinghua.edu.cn, dinggg@tsinghua.edu.cn

Wei Cheng
Department of Computer Science
University of North Carolina at Chapel Hill
weicheng@cs.unc.edu

Xiang Zhang
Department of Electrical Engineering and Computer Science
Case Western Reserve University
xiang.zhang@case.edu

Wei Wang
Department of Computer Science
University of North Carolina at Chapel Hill
weiwang@cs.unc.edu

CP10
Transfer Significant Subgraphs Across Graph Databases

A key step of graph classification is to identify informative subgraphs that encode label information. For instance, in drug efficacy prediction, the drugs (chemical compounds) effective against the same disease usually contain similar chemical-subgraphs effective to control the disease. Then, one can use such chemical subgraphs to identify effective drugs. We call these subgraphs *significant subgraphs*. In this paper, the aim is to utilize the significant subgraphs from related graph datasets to help label graphs of the target dataset. For example, we utilize the breast cancer drug data, and transfer the anti-cancer subgraphs to help label another set of drug data against lung cancer. To do so, we propose a Bayesian-based transfer learning model. The key idea is to first evaluate the similarity between the target and source datasets by estimating the degree they share on their significant subgraphs. This dataset similarity is then used to judiciously select significant subgraphs from similar (related) datasets to the target dataset. An optimization problem is devised to maximize the likelihood that the selected subgraphs are significant in the target dataset. The objective function is further proven to have the antimonotone property which can help prune the search space significantly. Sixteen sets of experiments show that the proposed algorithm can effectively reduce the error rates by as much as 40%. More importantly, it is 10 times faster than the comparison models, which include unsupervised and supervised significant subgraph mining algorithms.

Xiaoxiao Shi
Computer Department, University of Illinois at Chicago
xiao.x.shi@gmail.com

Xiangnan Kong, Philip Yu
University of Illinois at Chicago
kongxn@gmail.com, psyu@cs.uic.edu

CP11

Sor: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and Its Healthcare Applications

As more clinical information with increasing diversity become available for analysis, a large number of features can be constructed and leveraged for predictive modeling. Feature selection is a classic analytic component that faces new challenges due to the new applications: How to handle a diverse set of high dimensional features? How to select features with high predictive power, but low redundant information? How to design methods that can select globally optimal features with theoretical guarantee? How to incorporate and extend existing knowledge driven approach? In this paper, we present Scalable Orthogonal Regression (SOR), an optimization-based feature selection method with the following novelties: 1) Scalability: SOR achieves nearly linear scale-up with respect to the number of input features and the number of samples; 2) Optimality: SOR is formulated as an alternative convex optimization problem with theoretical convergence and global optimality guarantee; 3) Low-redundancy: thanks to the orthogonality objective, SOR is designed specifically to select less redundant features without sacrificing quality; 4) Extendability: SOR can enhance an existing set of preselected features by adding additional features that complement the existing feature set but still with strong predictive power. We present evaluation results showing that SOR consistently outperforms state of the art feature selection methods in a range of quality metrics on several real world data sets. We demonstrate a case study of a large-scale clinical application for predicting early onset of Heart Failure (HF) using real Electronic Health Records (EHRs) data of over 10K patients for over 7 years. Leveraging SOR, we are able to construct accurate and robust predictive models and derive potential clinical insights.

Dijun Luo
The University of Texas at Arlington
dijun.luo@gmail.com

Fei Wang, Jimeng Sun, Marianthi Marka
IBM T.J. Watson Research Center
fwang@us.ibm.com, jimeng@us.ibm.com,
mmarkat@us.ibm.com

Jianying Hu, Shahram Ebadollahi
IBM
jyhu@us.ibm.com, ebad@us.ibm.com

CP11

IntruMine: Mining Intruders in Untrustworthy Data of Cyber-Physical Systems

A Cyber-Physical System (CPS) integrates physical (i.e., sensor) devices with cyber (i.e., informational) components to form a situation-aware system that responds intelligently to dynamic changes in real-world. It has wide application to scenarios of traffic control, environment monitoring and battlefield surveillance. This study investigates the specific problem of intruder mining in CPS: With a large number of sensors deployed in a designated area, the task is real time detection of intruders who enter the area, based on untrustworthy data. We propose a method called IntruMine to detect and verify the intruders. IntruMine constructs monitoring graphs to model the relationships between sensors and possible intruders, and computes the position and energy of each intruder with the link information from these monitoring graphs. Finally, a confidence rating is calculated for each potential detection, reducing false positives in the results. IntruMine is a generalized approach. Two classical methods of intruder detection can be seen as special cases of IntruMine under certain conditions. We conduct extensive experiments to evaluate the performance of IntruMine on both synthetic and real datasets and the experimental results show that IntruMine has better effectiveness and efficiency than existing methods.

Lu-An Tang, Quanquan Gu, Xiao Yu, Jiawei Han
UIUC
tang18@uiuc.edu, qgu3@illinois.edu, xiaoyu1@illinois.edu, hanj@illinois.edu

Thomas La Porta
PSU
tlp@cse.psu.edu

Alice Leung
BBN Technology
aleung@bbn.com

Tarek Abdelzaher
UIUC
zaher@illinois.edu

Lance Kaplan
U.S. Army Lab
lance.m.kaplan.civ@mail.mil

CP11

Robust Reputation-Based Ranking on Bipartite

Rating Networks

With the growth of the Internet and E-commerce, bipartite rating networks are ubiquitous. In such bipartite rating networks, there exist two types of entities: the users and the objects, where users give ratings to objects. A fundamental problem in such networks is how to rank the objects by user's ratings. Although it has been extensively studied in the past decade, the existing algorithms either cannot guarantee convergence, or are not robust to the spammers. In this paper, we propose six new reputation-based algorithms, where the users' reputation is determined by the aggregated difference between the users' ratings and the corresponding objects' rankings. We prove that all of our algorithms converge into a unique fixed point. The time and space complexity of our algorithms are linear w.r.t. the size of the graph, thus they can be scalable to large datasets. Moreover, our algorithms are robust to the spamming users. We evaluate our algorithms using three real datasets. The experimental results confirm the effectiveness, efficiency, and robustness of our algorithms.

Rong-Hua Li, Jeffery Xu Yu, Xin Huang, Hong Cheng
The Chinese University of Hong Kong
rhli@se.cuhk.edu.hk, yu@se.cuhk.edu.hk,
xhuang@se.cuhk.edu.hk, hcheng@se.cuhk.edu.hk

CP11
Mining Massive Archives of Mice Sounds with Symbolized Representations

The house mouse has long been an important model organism in biology and medicine to address human diseases. Advances in sensor technology have created a situation where our ability to collect data far outstrips our ability to analyze it manually. In this work we show a novel technique for mining mice vocalizations directly in the visual (spectrogram) space and the use of similarity search, classification, motif discovery and contrast set mining in this domain.

Jesin Zakaria
3337 Utah Street
Riverside, CA-92507
jzaka001@ucr.edu

Sarah Rotschafer
Department of Psychology
University of California Riverside
srots001@ucr.edu

Abdullah Mueen
University of California, Riverside
mueen@cs.ucr.edu

Khaleel Razak
Department of Psychology
University of California Riverside
khaleel@ucr.edu

Eamonn Keogh
University of California, Riverside
eamonn@cs.ucr.edu

MS1
Efficient Monte Carlo Computation of Fisher Information Matrix using Prior Information

Abstract not available at time of publication.

Sonjoy Das
University at Buffalo
sonjoy@buffalo.edu

James Spall
Johns Hopkins University
james.spall@ jhuapl.edu

Roger Ghanem
University of Southern California
Aerospace and Mechanical Engineering and Civil Engineering
ghanem@usc.edu

MS1
Probabilistic Models of Past Climate Change

Abstract not available at time of publication.

Julien Emile-Geay, Dominique Guillot
University of Southern California
Los Angeles, CA 90089 0740, USA
julieneg@usc.edu, dguillot@usc.edu

Tapio Schneider
California Institute of Technology
tapio@caltech.edu

Bala Rajaratnam
Stanford University
brajarat@stanford.edu

MS1
Diffusion on Random Manifolds

Abstract not available at time of publication.

Hadi Meidani
University of Southern California
meidani@usc.edu

Roger Ghanem
University of Southern California
Aerospace and Mechanical Engineering and Civil Engineering
ghanem@usc.edu

MS1
A Priori Testing of Adaptive Sampling and Sparse PC Representations for Ocean General Circulation Models

Abstract not available at time of publication.

Justin Winokur
Johns Hopkins University
jwinokur@jhu.edu

Patrick R. Conrad
MIT
prconrad@mit.edu

Ihab Sraj, Alen Alexanderian
Johns Hopkins University

israj@jhu.edu, aalexa20@jhu.edu

Mohamed Iskandarani
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
MIskandarani@rsmas.miami.edu

Ashwanth Srinivasan
University of Miami
asrinivasan@rsmas.miami.edu

Youssef M. Marzouk
Massachusetts Institute of Technology
ymarz@mit.edu

Omar Knio
Duke University
amk@duke.edu

PP1

On Influential Node Discovery in Dynamic Social Networks

The problem of maximizing influence spread has been widely studied in social networks, because of its tremendous number of applications in determining critical points in a social network for information dissemination. All the techniques proposed in the literature are inherently static in nature, which are designed for social networks with a fixed set of links. However, many forms of *social interactions* are *transient* in nature, with relatively short periods of interaction. Any influence spread may happen only during the period of interaction, and the probability of spread is a function of the corresponding interaction time. Furthermore, such interactions are quite fluid and evolving, as a result of which the topology of the underlying network may change rapidly, as new interactions form and others terminate. In such cases, it may be desirable to determine the influential nodes based on the dynamic interaction patterns. Alternatively, one may wish to discover the most likely starting points for a *given infection pattern*. We will propose methods which can be used both for optimization of information spread, as well as the backward tracing of the source of influence spread. We will present experimental results illustrating the effectiveness of our approach on a number of real data sets.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Shuyang Lin, Philip Yu
University of Illinois at Chicago
slin38@cs.uic.edu, psyu@cs.uic.edu

PP1

Event Detection in Social Streams

Social networks generate a large amount of text content over time because of continuous interaction between participants. The mining of such *social streams* is more challenging than traditional text streams, because of the presence of both text content and implicit network structure within the stream. The problem of event detection is also closely related to clustering, because the events can only be inferred from *aggregate* trend changes in the stream. In this paper, we will study the two related problems of clustering and event detection in social streams. We will study both the supervised and unsupervised case for the event detection problem. We present experimental results illustrating the effectiveness of incorporating network structure in event discovery over purely content-based methods.

Charu C. Aggarwal
IBM T. J. Watson Research Center
charu@us.ibm.com

Karthik Subbian
University of Minnesota
karthik@umn.edu

PP1

Query-based Biclustering using Formal Concept Analysis

Abstract not available at time of publication.

Faris Alqadah
Johns Hopkins University
faris.alqadah@gmail.com

Joel S. Bader
Johns Hopkins University
Department of Biomedical Engineering
joel.bader@jhu.edu

Rajul Anand, Chandan Reddy
Wayne State University
rajulanand@wayne.edu, reddy@cs.wayne.edu

PP1

Granger Causality Analysis in Irregular Time Series

In this paper, we propose a nonparametric generalization of the Granger graphical models called Generalized Lasso Granger (GLG) to uncover the temporal dependencies from *Irregular Time Series*, whose observations are not sampled at equally-spaced time stamps. Via theoretical analysis and extensive experiments, we verify the effectiveness of our model. Furthermore, we apply GLG to the application dataset of $\delta^{18}O$ isotope of Oxygen records in Asia to discover the moisture transportation patterns in a 800-year period.

Mohammad Taha Bahadori, Yan Liu
University of Southern California
mohammab@usc.edu, yanliu.cs@usc.edu

PP1

Clustering Based on Yukawa Potential

Inspired by the clustering phenomenon of nucleus, we propose a novel dynamic clustering algorithm based on Yukawa potential (Yupc). Each data object is regarded as a particle following the basic rules of movements in the Yukawa potential field. After several time intervals, similar objects gradually aggregate together and form clear clusters. Natural clusters of different shapes, densities, sizes, numbers and distributions can be detected by Yupc, reflecting the intrinsic structure of the original data set.

Xue Bai, Zezhen Lin, Yun Xiong, Yangyong Zhu
Fudan University
xuebai@fudan.edu.cn, justinlin722@gmail.com,
yunx@fudan.edu.cn, yyzhu@fudan.edu.cn

PP1

Balancing Prediction and Recommendation Accuracy: Hierarchical Latent Factors for Preference Data

Recent works in Recommender Systems (RS) have investigated the relationships between the prediction accuracy, i.e. the ability of a RS to minimize cost functions in estimating users' preferences, and the accuracy of the recommendation list provided to users. Algorithms, which focus on the minimization of cost functions, have shown to achieve a weak recommendation accuracy, and vice versa. We present a Bayesian probabilistic hierarchical approach for RS designed to meet both prediction and recommendation accuracy.

Ettore Ritacco, Nicola Barbieri, Giuseppe Manco,
Riccardo Ortale
ICAR-CNR
ritacco@icar.cnr.it, barbieri@icar.cnr.it,
manco@icar.cnr.it, ortale@icar.cnr.it

PP1

Deterministic Cur for Improved Large-Scale Data Analysis: An Empirical Study

Low-rank approximations which are computed from selected rows and columns of a given data matrix have attracted considerable attention lately. They have been proposed as an alternative to the SVD because they nat- urally lead to interpretable decompositions which was shown to be successful in application such as fraud de- tection, fMRI segmentation, and collaborative filtering. The CUR decomposition of large matrices, for exam- ple, samples rows and columns according to a proba- bility distribution that depends on the Euclidean norm of rows or columns or on other measures of statistical leverage. At the same time, there are various deter- ministic approaches that do not resort to sampling and were found to often yield factorization of superior qual- ity with respect to reconstruction accuracy. However, these are hardly applicable to large matrices as they typically suffer from high computational costs. Con- sequently, many practitioners in the field of data min- ing have abandon deterministic approaches in favor of randomized ones when dealing with todays large-scale data sets. In this paper, we empirically disprove this prejudice. We do so by introducing a novel, linear-time, deterministic CUR approach that adopts the recently introduced Simplex Volume Maximization approach for column selection. The latter has already been proven to be successful for NMF-like decompositions of matrices of billions of entries. Our exhaustive empirical study on more than 30 synthetic and real-world data sets demon- strates that it is also beneficial for CUR-like decomposi- tions. Compared to other deterministic CUR-like meth- ods, it provides comparable reconstruction quality but operates much faster so that it easily scales to matrices of billions of elements. Compared to sampling-based methods, it provides competitive reconstruction quality while staying in the same run-time complexity class.

Christian Thurau, Kristian Kersting,
Christian Bauckhage
Fraunhofer IAIS
cthurau@gmail.com, kristian.kersting@iais.fraunhofer.de,
christian.bauckhage@iais.fraunhofer.de

PP1

Combining Active Learning and Dynamic Dimensionality Reduction

To date, many active learning techniques have been developed for acquiring labels when training data is limited. However, an important aspect of the problem has often been neglected or just mentioned in passing: the curse of dimensionality. Yet, the curse of dimensionality poses even greater challenges in the case of limited data, which is precisely the setup for active learning. Reducing the dimensions is not a trivial task, however, as the correct number of dimensions depends on a number of factors including the training data size, the number of classes, the discriminative power of the features, and the underlying classification model. Moreover, active learning is typically applied in an iterative manner where the number of labels is smaller in the earlier iterations compared to the later ones. We propose an adaptive dimensionality reduction technique that determines the appropriate number of dimensions for each active learning iteration, utilizing the labeled and unlabeled data effectively to learn more accurate models. Extensive experiments comparing various approaches and parameter settings show that the proposed method improves performance drastically on three real-world text classification tasks.

Mustafa Bilgic
Illinois Institute of Technology
mbilgic@iit.edu

PP1

Context-Aware Search for Personal Information Management Systems

We present a novel context-aware desktop search framework by leveraging Hidden Markov Model to capture the relationships between user's access actions and activity states. The model is learned from user's past access history and is used to predict user's current activity upon the submission of some keyword query. We further propose a ranking scheme with this predicted context information incorporated. Experimental evaluation demonstrates its enhancement to user's search experience.

Jidong Chen
EMC Research China
EMC Corporation
jidong.chen@emc.com

Wentao Wu
Fudan University
wentaowu@fudan.edu.cn

Hang Guo
EMC Research China
hang.guo@emc.com

Wei Wang
Fudan University
weiwang1@fudan.edu.cn

PP1

Mining Social Dependencies in Dynamic Interaction Networks

User-to-user interactions have become ubiquitous in Web 2.0. Users exchange emails, post on newsgroups, tag web pages, co-author papers, etc. Through these interactions, users co-produce or co-adopt content items (e.g., words in emails, tags in social bookmarking sites). We model such

dynamic interactions as a user interaction network, which relates users, interactions, and content items over time. After some interactions, a user may produce content that is more similar to those produced by other users previously. We term this effect *social dependency*, and we seek to mine from such networks the degree to which a user may be socially dependent on another user over time. We propose a *Decay Topic Model* to model the evolution of a user's preferences for content items at the topic level, as well as a *Social Dependency Metric* that quantifies the extent of social dependency based on interactions and content changes. Our experiments on two user interaction networks induced from real-life datasets show the effectiveness of our approach.

Freddy Chua, Hady Lauw, Ee-Peng Lim
Singapore Management University
freddycct@gmail.com, hadywlauw@smu.edu.sg, eplim@smu.edu.sg

PP1

Detecting Irregularly Shaped Significant Spatial and Spatio-Temporal Clusters

Detecting significant overdensity or underdensity clusters in spatio-temporal data is critical for many real-world applications. Most existing approaches are designed to deal with regularly shaped clusters such as circular, elliptic and rectangular ones, but cannot work well on irregularly shaped clusters. In this paper, we propose GridScan, a grid-based approach for detecting irregularly shaped spatial clusters. In GridScan, a cluster is asymptotically described by a set of connected grid cells and is computed by a fast greedy region-growing algorithm with elaborating cluster merging in the process. The time complexity of GridScan is linear to the number of grids, making it scalable to very large datasets. A prospective spatio-temporal cluster detection approach, GridScan-Pro, is also proposed by extending GridScan. Experiments and a case study in the epidemic scenario demonstrate that our approaches greatly outperform existing ones in terms of accuracy, efficiency, and scalability.

Weishan Dong, Xin Zhang, Li Li, Changhua Sun, Lei Shi, Wei Sun
IBM Research - China
dongweis@cn.ibm.com, zxin@cn.ibm.com, lilichina@cn.ibm.com, schangh@cn.ibm.com, shllsh@cn.ibm.com, weisun@cn.ibm.com

PP1

Contextual Collaborative Filtering Via Hierarchical Matrix Factorization

Matrix factorization (MF) has been demonstrated to be one of the most competitive techniques for collaborative filtering. However, state-of-the-art MFs do not consider contextual information, where ratings can be generated under different environments. For example, users select items under various situations, such as happy mood vs. sad, mobile vs. stationary, movies vs. book, etc. Under different contexts, the preference of users are inherently different. The problem is that MF methods uniformly decompose the rating matrix, and thus they are unable to factorize for different contexts. To amend this problem and improve recommendation accuracy, we introduce a "hierarchical' factorization model by considering the local context when performing matrix factorization. The intuition is that: as ratings are being generated from heterogeneous environments, certain user and item pairs tend to be more similar to each other than others, and hence they ought to receive more collaborative information from each other. To take the contextual information into consideration, the proposed "contextual collaborative filtering' approach splits the rating matrix hierarchically by grouping similar users and items together, and factorizes each sub-matrix locally under different contexts. By building an ensemble model, the approach further avoids over-fitting with less parameter tuning. We analyze and demonstrate that the proposed method is a model-averaging gradient boosting model, and its error rate can be bounded. Experimental results show that it outperforms three state-of-the-art algorithms on a number of real-world datasets (MovieLens, Netflix, etc). The source code and datasets are available for download `http://www.cse.ust.hk/~ezhong/code/sdm12hmf.zip`.

Erheng Zhong
Sun Yat-Sen University
ezhong@cse.ust.hk

Wei Fan
IBM T.J.Watson Research,
weifan@us.ibm.com

Qiang Yang
Department of Computer Science,
Hong Kong University of Science
qyang@cse.ust.hk

PP1

Active Learning with Monotonicity Constraints

In many applications of data mining it is known beforehand that the response variable should be increasing (or decreasing) in the attributes. We propose two algorithms to exploit such monotonicity constraints for active learning in ordinal classification in two different settings. The basis of our approach is the observation that if the class label of an object is given, then the monotonicity constraints may allow the labels of other objects to be inferred. For instance, from knowing that loan applicant a is rejected, it can be concluded that all applicants that score worse than a on all criteria should be rejected as well. We propose two heuristics to determine good query points. These heuristics make a selection based on a point's potential to infer the labels of other points. The algorithms, each implemented with the proposed heuristics, are evaluated on artificial and real data sets to study their performance. We conclude that exploitation of monotonicity constraints can be very beneficial in active learning.

Ad Feelders, Nicola Barile
Universiteit Utrecht
A.J.Feelders@uu.nl, n.barile@uu.nl

PP1

Pseudo Cold Start Link Prediction with Multiple Sources in Social Networks

Link prediction is an important task in social networks and data mining. Most existing researches therefore approach this problem by exploring the topological structure of the social network using only one source of information. In this work, we introduce the pseudo cold start link prediction with multiple source. We propose a two-phase supervised method. We assess our method empirically over a large

data collection obtained from Youtube.

Liang Ge
The Sate University of New York at Buffalo
liangge@buffalo.edu

PP1
Discovering Context-Aware Influential Objects

IIt is very helpful for a user to get a *moderate* amount of information highly related to his/her immediate *context* (e.g., location, time, discussion topics) during the exploration of digital object collections (e.g., articles, web pages, blogs). For instance, in investigating a research topic, a researcher may be very interested in finding articles that are most related to the articles he/she already read on this topic, which we consider as “context’ in this paper. To facilitate users’ exploration, we introduce the problem of discovering Context-aware Influential Objects (CIO) from a collection of digital objects with influence relationships. Although there is a large amount of work in detecting direct influence degree between objects to denote how strong an object influences others, very few works utilize such direct influence to find influential objects for a context. To discover CIOs for a context consisting of several objects of a user’s interest, the first challenge is to meaningfully measure the *collective* influence of an object over a context considering both the direct influence and the *indirectly* derived influence, which is not taken into consideration by most “query by example’ approaches. We propose an aggregation framework to formulate the collective influence among objects by leveraging both direct and indirect influence. The second challenge is to discover CIOs *efficiently.* We present three approaches to calculate collective influence of an object over a context from an influence graph. In particular, the first approach utilizes the breadth-first-search paradigm; the other approaches make use of the topological sorting of graph nodes and perform context-aware search using *push* and *pull* mechanisms. We show experimental results on real datasets to demonstrate the effectiveness and efficiency of the proposed methodologies.

Huiping Cao
Computer Science
New Mexico State University
hcao@cs.nmsu.edu

Yangpai Liu, Yifan Hao
New Mexico State University
lypmoon@nmsu.edu, yifan@nmsu.edu

Peng Han, Xinda Zeng
Chongqing Academy of Science and Technology, China
han.peng@ciat-cq.org, shinda1020@gmail.com

PP1
Monitoring and Mining Insect Sounds in Visual Space

Monitoring animals by the sounds they produce is an important and challenging task, whether the application is outdoors in a natural habitat, or in the controlled environment of a laboratory setting. In the former case the density and diversity of animal sounds can act as a measure of biodiversity. In the latter case, researchers often create control and treatment groups of animals, expose them to different interventions, and test for different outcomes. One possible manifestation of different outcomes may be changes in the bioacoustics of the animals. With such a plethora of important applications, there have been significant efforts to build bioacoustic classification tools. However, we argue that most current tools are severely limited. They often require the careful tuning of many parameters (and thus huge amounts of training data), they are too computationally expensive for deployment in resource-limited sensors, they are specialized for a very small group of species, or they are simply not accurate enough to be useful. In this work we introduce a novel bioacoustic recognition/classification framework that mitigates or solves all of the above problems. We propose to classify animal sounds in the visual space, by treating the texture of their spectrograms as an acoustic fingerprint using a recently introduced parameter-free texture measure as a distance measure. We further show that by searching for the most representative acoustic fingerprint we can significantly outperform other techniques in terms of speed and accuracy.

Yuan Hao, Bilson J. Campana, Eamonn Keogh
University of California, Riverside
yhao@cs.ucr.edu, bcampana@cs.ucr.edu, eamonn@cs.ucr.edu

PP1
Image Mining of Historical Manuscripts to Establish Provenance

The recent digitization of more than twenty million books has been led by initiatives from countries wishing to preserve their cultural heritage and by commercial endeavors, such as the Google Print Library Project. Within a few years a significant fraction of the worlds books will be online. For millions of intact books and tens of millions of loose pages, the provenance of the manuscripts may be in doubt or completely unknown, thus denying historians an understanding of the context of the content. In some cases it may be possible for human experts to regain the provenance by examining linguistic, cultural and/or stylistic clues. However, such experts are rare and this investigation is clearly a time-consuming process. One technique used by experts to establish provenance is the examination of the ornate initial letters appearing in the questioned manuscript. By comparing the initial letters in the manuscript to annotated initial letters whose origin is known, the provenance can be determined. In this work we show for the first time that we can reproduce this ability with a computer algorithm. We leverage off a recently introduced technique to measure texture similarity and show that it can recognize initial letters with an accuracy that rivals or exceeds human performance. A brute force implementation of this measure would require several years to process a single large book; however, we introduce a novel lower bound that allows us to process the books in minutes.

Bing Hu
University of California, Riverside
University of California, Riverside
bhu002@ucr.edu

PP1
RP-growth: Top-k Mining of Relevant Patterns with Minimum Support Raising

This paper proposes RP-growth, an efficient top-k mining algorithm for the patterns highly relevant to the class of interest. RP-growth conducts branch-and-bound search using anti-monotonic upper bounds of the relevance score, and its pruning strategy is successfully translated to minimum support raising, a standard pruning strategy in top-k mining. Furthermore, RP-growth introduces an aggressive pruning strategy based on the notion called weakness. Experimental results on text classification exhibit the efficiency and the usefulness of RP-growth.

Yoshitaka Kameya, Taisuke Sato
Tokyo Institute of Technology
kameya@mi.cs.titech.ac.jp, sato@mi.cs.titech.ac.jp

PP1

Fast Random Walk Graph Kernel

Random walk graph kernel has been used as an important tool for various data mining tasks including classification and similarity computation. Despite its usefulness, however, it suffers from the expensive computational cost which is at least $O(n^3)$ or $O(m^2)$ for graphs with n nodes and m edges. In this paper, we propose Ark, a set of fast algorithms for random walk graph kernel computation. Ark is based on the observation that real graphs have much lower intrinsic ranks, compared with the orders of the graphs. Ark exploits the low rank structure to quickly compute random walk graph kernels in $O(n^2)$ or $O(m)$ time. Experimental results show that our method is up to $97,865$ times faster than the existing algorithms, while providing more than 91.3% of the accuracies.

U Kang
Carnegie Mellon University
Computer Science Department
ukang@cs.dot.cmu.dot.edu

Hanghang Tong
IBM T.J. Watson
htong@us.ibm.com

Jimeng Sun
IBM T.J. Watson Research Center
jimeng@us.ibm.com

PP1

Tracking Spatio-Temporal Diffusion in Climate Data

A forest canopy forms a critical platform for complex interactions between the vegetation and the atmosphere boundary layer and is considered as a crucial piece for environmental scientists in their understanding of the ecosystem and its response to the climate change. Microfronts represent a class of these interactions characterized by a moving mass of air that introduce fluctuations in ambient temperature and humidity on small spatial and temporal scales. In this paper, we present a joint spatio-temporal hidden markov model that simultaneously incorporates neighborhood dependencies in space and time. We show that our approach can trace the diffusion of microfronts more effectively than several baseline methods over a sensor data from Brazilian rainforest and a synthetically generated dataset.

Jaya Kawale, Aditya Pal
University of Minnesota
kawale@cs.umn.edu, apal@cs.umn.edu

Rob Fatland
Microsoft Research
rob.fatland@microsoft.com

PP1

Group Sparsity in Nonnegative Matrix Factorization

A recent challenge in data analysis for science and engineering is that data are often represented in a structured way. In particular, many data mining tasks have to deal with group-structured prior information, where features or data items are organized into groups. In this paper, we develop group sparsity regularization methods for nonnegative matrix factorization (NMF). NMF is an effective data mining tool that has been widely adopted in text mining, bioinformatics, and clustering, but a principled approach to incorporating group information into NMF has been lacking in the literature. Motivated by an observation that features or data items within a group are expected to share the same sparsity pattern in their latent factor representation, we propose mixed-norm regularization to promote group sparsity in the factor matrices of NMF. Group sparsity improves the interpretation of latent factors. Efficient convex optimization methods for dealing with the mixed-norm term are presented along with computational comparisons between them. Application examples of the proposed method in factor recovery, semi-supervised clustering, and multilingual text analysis are demonstrated.

Jingu Kim
Georgia Institute of Technology
jingu@cc.gatech.edu

Renato C. Monteiro
Georgia Institute of Technology
School of ISyE
monteiro@isye.gatech.edu

Haesun Park
Georgia Institute of Technology
hpark@cc.gatech.edu

PP1

Global Linear Neighborhoods for Efficient Label Propagation

Graph-based semi-supervised learning improves classification by combining labeled and unlabeled data through label propagation. In this paper, we propose to learn a nonnegative low-rank graph to capture global linear neighborhoods, under the assumption that each data point can be linearly reconstructed from weighted combinations of its direct neighbors and reachable indirect neighbors. Large scale experiments on UCI datasets and gene expression datasets showed label propagation based on global linear neighborhoods achieved more accurate classification results.

Ze Tian, Rui Kuang
Dept Computer Science
University of Minnesota
tianze@cs.umn.edu, kuang@cs.umn.edu

PP1

Generalized Similarity Kernels for Efficient Sequence Classification

String kernel-based machine learning methods have yielded great success in practical tasks of structured/sequential data analysis such as document topic elucidation, music genre classification, protein superfamily and fold prediction. However, typical string kernel methods rely on *sym-*

bolic Hamming-distance based matching which may not necessarily reflect the underlying (e.g., physical) similarity between sequence fragments. In this work we propose a novel computational framework that uses more “precise’, *general similarity metrics* $\mathcal{S}(\cdot,\cdot)$ and distance-preserving embeddings with string kernels and improves upon state-of-the-art on a number of sequence analysis tasks such as music, and biological sequence classification.

Pavel P. Kuksa
Rutgers University
pkuksa@nec-labs.com

Imdadullah Khan
Gulf University for Science and Technology
imdadk@gmail.com

Vladimir Pavlovic
Rutgers University
vladimir@cs.rutgers.edu

PP1
Detecting Extreme Rank Anomalous Collections

Anomaly or outlier detection has a wide range of applications, including fraud and spam detection. Most existing studies focus on detecting point anomalies, i.e., individual, isolated entities. However, there is an increasing number of applications in which anomalies do not occur individually, but in small collections. Unlike the majority, entities in an anomalous collection tend to share certain extreme behavioral traits. The knowledge essential in understanding why and how the set of entities becomes outliers would only be revealed by examining at the collection level. A good example is web spammers adopting common spamming techniques. To discover this kind of anomalous collections, we introduce a novel definition of anomaly, called *Extreme Rank Anomalous Collection*. We propose a statistical model to quantify the anomalousness of such a collection, and present an exact as well as a heuristic algorithms for finding top-K extreme rank anomalous collections. We apply the algorithms on real Web spam data to detect spamming sites, and on IMDB data to detect unusual actor groups. Our algorithms achieve higher precisions compared to existing spam and anomaly detection methods. More importantly, our approach succeeds in finding meaningful anomalous collections in both datasets.

Hanbo Dai, Feida Zhu, Ee-Peng Lim, Hwee Hwa Pang,
Hady Lauw
Singapore Management University
hanbo.dai.2008@smu.edu.sg, fdzhu@smu.edu.sg,
eplim@smu.edu.sg, hhpang@smu.edu.sg,
hadywlauw@smu.edu.sg

PP1
Visualizing Variable-Length Time Series Motifs

The problem of time series motif discovery has received a lot of attention from researchers in the past decade. Most existing work on finding time series motifs require that the length of the motifs be known in advance. However, such information is not always available. In addition, motifs of different lengths may co-exist in a time series dataset. In this work, we develop a motif visualization system based on grammar induction. We demonstrate that grammar induction in time series can effectively identify repeated patterns without prior knowledge of their lengths. The motifs discovered by the visualization system are of variable lengths in two ways. Not only can the *inter-motif* subsequences be of different lengths, the *intra-motif* subsequences also are not restricted to have identical length—a unique property that is desirable, but has not been seen in the literature.

Yuan Li, Jessica Lin
George Mason University
Department of Computer Science
ylif@gmu.edu, jessica@cs.gmu.edu

Tim Oates
Department of Computer Science and Electrical
Engineering
University of Maryland Baltimore County, Baltimore,
USA
oates@cs.umbc.edu

PP1
Which Distance Metric Is Right: An Evolutionary K-Means View

We study the impact of monotone metrics on K-means clustering. By revealing the order-preserving property and proving that the cluster centroid is a good approximator of their respective optimal centers, we show K-means cannot differentiate the cosine-monotone metrics. Then an evolutionary framework is proposed to enable inspection of these metrics. Most importantly, this paper furthers our understanding of the impact of the metrics on the optimization process of K-means.

Chuanren Liu
Rutgers, the State University of New Jersey
chuanren.liu@rutgers.edu

Tianming Hu
Dongguan University of Technology
tmhu@ieee.org

Yong Ge, Hui Xiong
Rutgers, the State University of New Jersey
yongge@rutgers.edu, hxiong@rutgers.edu

PP1
Constructing Training Sets for Outlier Detection

Outlier detection often works in an unsupervised manner due to the difficulty of obtaining enough training data. Since outliers are rare, one has to label a very large dataset to include enough outliers in the training set, with which classifiers could sufficiently learn the concept of outliers. Labeling a large training set is costly for most applications. However, we could just label suspected instances identified by unsupervised methods. In this way, the number of instances to be labeled could be greatly reduced. Based on this idea, we propose CISO, an algorithm Constructing training set by Identifying Suspected Outliers. In this algorithm, instances in a pool are first ranked by an unsupervised outlier detection algorithm. Then, suspected instances are selected and hand-labeled, and all remaining instances receive label of inlier. As such, all instances in the pool are labeled and used in the training set. We also propose Budgeted CISO (BCISO), with which user could set a fixed budget for labeling. Experiments show that both algorithms achieve good performance compared to other methods when the same amount of labeling effort are used.

Liping Liu
EECS Oregon State University
liping.liulp@gmail.com

Xiaoli Z. Fern
Oregon State University
xfern@eecs.oregonstate.edu

PP1

A Flexible Open-Source Toolbox for Scalable Complex Graph Analysis

The Knowledge Discovery Toolbox (KDT) enables domain experts to perform complex analyses of huge datasets on supercomputers using a high-level language without grappling with the difficulties of writing parallel code, calling parallel libraries, or becoming a graph expert. KDT provides a flexible Python interface to a small set of high-level reusable graph operations; composing these operations produces graph analysis algorithms scalable to graphs on the order of 10 billion edges or greater.

Adam Lugowski
UC Santa Barbara
alugowski@cs.ucsb.edu

Aydin Buluc
Lawrence Berkeley National Laboratory
abuluc@lbl.gov

David Alber
Microsoft
david.alber@microsoft.com

John R. Gilbert
Dept of Computer Science
University of California, Santa Barbara
gilbert@cs.ucsb.edu

Steve Reinhardt
Cray, Inc.
spr@cray.com

Yun Teng, Andrew Waranis
UC Santa Barbara
yunteng@umail.ucsb.edu, andrewwaranis@umail.ucsb.edu

PP1

Fast Robustness Estimation in Large Social Graphs: Communities and Anomaly Detection

Given a large social graph, what can we say about its robustness? In this work, we are trying to answer the above question studying the *expansion properties* of large social graphs. We present a measure which characterizes the robustness of a graph and serves as global measure of the community structure (or lack thereof), and we show how to compute it efficiently. We present extensive experimental results on both static and time-evolving real networks.

Fragkiskos D. Malliaros
Department of Computer Engineering and Informatics
University of Patras
malliaro@ceid.upatras.gr

Vasileios Megalooikonomou
Department of Computer Engineering and Informatics
University of Patras and Temple University
vasilis@ceid.upatras.gr

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

PP1

On Finding Joint Subspace Boolean Matrix Factorizations

Abstract not available at time of publication.

Pauli Miettinen
Max Planck Institute for Informatics
pauli.miettinen@mpi-inf.mpg.de

PP1

Generalized Optimization Framework for Graph-Based Semi-Supervised Learning

We develop a generalized optimization framework for graph-based semi-supervised learning. The framework gives as particular cases the Standard Laplacian, Normalized Laplacian and PageRank based methods. We have also provided new probabilistic interpretation based on random walks and characterized the limiting behaviour of the methods. The random walk based interpretation allows us to explain differences between the performances of methods with different smoothing kernels. It appears that the PageRank based method is robust with respect to the choice of the regularization parameter and the labelled data. We illustrate our theoretical results with two realistic datasets, characterizing different challenges: Les Miserables characters social network and Wikipedia hyper-link graph. The graph-based semi-supervised learning classifies the Wikipedia articles with very good precision and perfect recall employing only the information about the hyper-text links.

Marina M. Sokol, Konstantin Avrachenkov
INRIA Sophia Antipolis
marina.sokol@inria.fr, k,avrachenkov@sophia.inria.fr

Paulo Goncalves
INRIA Rhone
paulo.goncalves@ens-lyon.fr

Alexey Mishenin
St. Petersburg State University
alexey.mishenin@gmail.com

PP1

A Tree-Based Kernel for Graphs

This paper proposes a new tree-based kernel for graphs. Graphs are decomposed into multisets of ordered Directed Acyclic Graphs (DAGs) and a family of kernels is defined by application of tree kernels extended to the DAG domain. We focus our attention on the efficient development of one member of this family. A technique for speeding up the computation is given, as well as theoretical bounds and practical evidence of its feasibility.

Nicolo' Navarin, Giovanni Da San Martino, Alessandro Sperduti

University of Padova
Department of Mathematics
nnavarin@math.unipd.it, dasan@math.unipd.it, sperduti@math.unipd.it

PP1

Density-Based Projected Clustering over High Dimensional Data Streams

Clustering of high dimensional data streams is an important problem in many application domains, a prominent example being network monitoring. Several approaches have been lately proposed for solving independently the different aspects of the problem. There exist methods for clustering over full dimensional streams and methods for finding clusters in subspaces of high dimensional static data. Yet only a few approaches have been proposed so far which tackle both the stream and the high dimensionality aspects of the problem simultaneously. In this work, we propose a new density-based projected clustering algorithm, HDDStream, for high dimensional data streams. Our algorithm summarizes both the data points and the dimensions where these points are grouped together and maintains these summaries online, as new points arrive over time and old points expire due to ageing. Our experimental results illustrate the e effectiveness and the efficiency of HDDStream and also demonstrate that it could serve as a trigger for detecting drastic changes in the underlying stream population, like bursts of network attacks.

Eirini C. Ntoutsi
Ludwig-Maximilians-Universität München (LMU)
ntoutsi@dbs.ifi.lmu.de

Arthur Zimek
LMU Munich
zimek@dbs.ifi.lmu.de

Themis Palpanas
Information Engineering and Computer Science Department
(DISI), University of Trento, Italy
themis@disi.unitn.eu

Peer Kröger
Ludwig-Maximilians-Universität München
kroeger@dbs.ifi.lmu.de

Hans-Peter Kriegel
Ludwig-Maximilians University Munich
kriegel@dbs.ifi.lmu.de

PP1

A Novel Approximation to Dynamic Time Warping Allows Anytime Clustering of Massive Time Series Datasets

Given the ubiquity of time series data, the data mining community has spent significant time investigating the best time series similarity measure to use for various tasks and domains. After more than a decade of extensive efforts, there is increasing evidence that Dynamic Time Warping (DTW) is very difficult to beat. Given that, recent efforts have focused on making the intrinsically slow DTW algorithm faster. For the similarity-search task, an important subroutine in many data mining algorithms, significant progress has been made by replacing the vast majority of expensive DTW calculations with cheap-to-compute lower bound calculations. However, these lower bound based optimizations do not directly apply to clustering, and thus for some realistic problems, clustering with DTW can take days or weeks. In this work, we show that we can mitigate this untenable lethargy by casting DTW clustering as an anytime algorithm. At the heart of our algorithm is a novel data-adaptive approximation to DTW which can be quickly computed, and which produces approximations to DTW that are much better than the best currently known linear-time approximations. We demonstrate our ideas on real world problems showing that we can get virtually all the accuracy of a batch DTW clustering algorithm in a fraction of the time.

Qiang Zhu, Gustavo E. Batista,
Thanawin Rakthanmanon, Emaonn Keogh
University of California, Riverside
qzhu@cs.ucr.edu, gbatista@cs.ucr.edu, rakthant@cs.ucr.edu, eamonn@cs.ucr.edu

PP1

Nearest-Neighbor Search on a Time Budget via Max-Margin Trees

Many high-profile applications pose high-dimensional nearest-neighbor search problems. Yet, it still remains difficult to achieve fast query times for state-of-the-art approaches which use multidimensional trees for either exact or approximate search, possibly in combination with hashing approaches. Moreover, a number of these applications only have a limited amount of time to answer nearest-neighbor queries. However, we observe empirically that the correct neighbor is often found early within the tree-search process, while the bulk of the time is spent on verifying its correctness. Motivated by this, we propose an algorithm for finding the best neighbor given any particular time limit, and develop a new data structure, the *max-margin tree*, to achieve accurate results even with small time budgets. Max-margin trees perform better in the limited-time setting than current commonly-used data structures such as the *kd*-tree and the more recently developed RP-tree data structure.

Parikshit Ram
School of Computational Science and Engineering
Georgia Institute of Technology
p.ram@gatech.edu

Dongryeol Lee, Alexander Gray
Georgia Institute of Technology
dongryel@cc.gatech.edu, agray@cc.gatech.edu

PP1

Efficient Clustering of Metagenomic Sequences Using Locality Sensitive Hashing

The new generation of genomic technologies have allowed researchers to determine the collective DNA of organisms (e.g. microbes) co-existing as communities across the ecosystem (e.g. within the human host). There is a need for the computational approaches to analyze and annotate the large volumes of available sequence data from such microbial communities (metagenomes). In this paper, we developed an efficient and accurate metagenome clustering approach that uses the locality sensitive hashing (LSH) technique to approximate the computational complexity associated with comparing sequences. We introduce the use of fixed-length, gapless subsequences for improving the sensitivity of the LSH-based similarity function. We evaluate the performance of our algorithm on two metagenome datasets associated with microbes existing across diffferent human skin locations. Our empirical results show the strength of the developed approach in comparison to three state-of-the-art sequence clustering algorithms with regards to computational efficiency and clustering quality. We also demonstrate practical significance for the developed clustering algorithm, to compare bacterial diversity and structure across diffferent skin locations.

Zeehasham Rasheed, Huzefa Rangwala, Daniel Barbara
George Mason University
zrasheed@gmu.edu, rangwala@cs.gmu.edu, dbarbara@gmu.edu

PP1
On Evaluation of Outlier Rankings and Outlier Scores

Outlier detection research is currently focusing on the development of new methods and on improving the computation time for these methods. Evaluation however is rather heuristic, often considering just precision in the top k results or using the area under the ROC curve. These evaluation procedures do not allow for assessment of similarity between methods. Judging the similarity of or correlation between two rankings of outlier scores is an important question in itself but it is also an essential step towards meaningfully building outlier detection ensembles, where this aspect has been completely ignored so far. In this study, our generalized view of evaluation methods allows both to evaluate the performance of existing methods as well as to compare different methods w.r.t. their detection performance. Our new evaluation framework takes into consideration the class imbalance problem and offers new insights on similarity and redundancy of existing outlier detection methods. As a result, the design of effective ensemble methods for outlier detection is considerably enhanced.

Arthur Zimek, Erich Schubert, Remigius Wojdanowski
LMU Munich
zimek@dbs.ifi.lmu.de, schube@dbs.ifi.lmu.de, wojdanowski@dbs.ifi.lmu.de

Hans-Peter Kriegel
Ludwig-Maximilians University Munich
kriegel@dbs.ifi.lmu.de

PP1
Regularized Structured Output Learning with Partial Labels

We consider the problem of learning structured output probabilistic models with training examples having partial labels. Partial label scenarios arise commonly in web applications such as hierarchical and multi-label classification. We solve the learning problem with partial labels by incorporating entropy and label distribution or correlation regularizations along with marginal likelihood maximization. We develop probabilistic taxonomy and multi-label classifier models, and provide the ideas needed for expanding their usage to the partial labels scenario.

Sundararajan Sellamanickam, Charu Tiwari
Yahoo! Labs, Bangalore, India
ssrajan@yahoo-inc.com, charu@yahoo-inc.com

Sathiya Keerthi Selvaraj
Yahoo! Labs, Santa Clara, CA
selvarak@yahoo-inc.com

PP1
The Similarity Between Stochastic Kronecker and Chung-Lu Graph Models

The *Stochastic Kronecker Graph* (SKG) model has been chosen as a benchmark by the Graph500 steering committee, but there is little understanding of the properties of this model. We show that the parallel variant of the edge-configuration model given by Chung and Lu (CL) is similar to the SKG model. Our experiments suggest that the graph distribution represented by SKG is almost the same as that given by a CL model. Also, CL appears to fit real data as well as SKG.

C. Seshadhri, Ali Pinar
Sandia National Labs
scomand@sandia.gov, apinar@sandia.gov

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

PP1
Wigm: Discovery of Subgraph Patterns in a Large Weighted Graph

Many research areas have begun representing massive data sets as very large graphs. Thus, graph mining has been an active research area in recent years. Most of the graph mining research focuses on mining unweighted graphs. However, weighted graphs are actually more common. The weight on an edge may represent the likelihood or logarithmic transformation of likelihood of the existence of the edge or the strength of an edge, which is common in many biological networks. In this paper, a weighted subgraph pattern model is proposed to capture the importance of a subgraph pattern and our aim is to find these patterns in a large weighted graph. Two related problems are studied in this paper: (1) discovering all patterns with respect to a given minimum weight threshold and (2) finding k patterns with the highest weights. The weighted subgraph patterns do not possess the anti-monotonic property and in turn, most of existing subgraph mining methods could not be directly applied. Fortunately, the **1-extension** property is identified so that a bounded search can be achieved. A novel weighted graph mining algorithm, namely WIGM, is devised based on the 1-extension property. Last but not least, real and synthetic data sets are used to show the effectiveness and efficiency of our proposed models and algorithms.

Wei Su, Jiong Yang
Case Western Reserve University
wei.su@case.edu, jiong.yang@case.edu

Shirong Li
Aliyun Inc
shirong.li@alibaba-inc.com

Mehmet Dalkilicb
Indiana University
dalkilic@indiana.edu

PP1

Legislative Prediction Via Random Walks over a Heterogeneous Graph

In this article, we propose a random walk-based model to predict legislators votes on a set of bills. In particular, we first convert roll call data, i.e. the recorded votes and the corresponding deliberative bodies, to a heterogeneous graph, where both the legislators and bills are treated as vertices. Three types of weighted edges are then computed accordingly, representing legislators social and political relations, bills semantic similarity, and legislator-bill vote relations. Through performing two-stage random walks over this heterogeneous graph, we can estimate legislative votes on past and future bills. We apply this proposed method on real legislative roll call data of the United States Congress and compare to state-of-the-art approaches. The experimental results demonstrate the superior performance and unique prediction power of the proposed model.

Jun Wang
IBM Thomas J. Watson Research Center
Business Analytics and Mathematical Sciences Department
wangjun@us.ibm.com

Kush Varshne, Aleksandra Mojsilovic
IBM Thomas J. Watson Research Center
krvarshn@us.ibm.com, aleksand@us.ibm.com

PP1

An Iterative and Re-Weighting Framework for Rejection and Uncertainty Resolution in Crowdsourcing

We propose an Iterative Re-weighted Consensus Maximization framework to address the missing and uncertain label problem. The intuitive idea is to use an iterated framework to estimate each labeler's hidden competence and formulate it as a spectral clustering problem in the functional space, in order to minimize the overall loss given missing and uncertain information. One main advantage of the proposed method from state-of-the-art Bayesian model averaging based approaches is that it uncovers the intrinsic consistency among different set of answers and mines the best possible ground truth.

Sihong Xie
University of Illinois at Chicago
sxie6@uic.edu

Wei Fan
IBM T.J.Watson Research
wei.fan@gmail.com

Philip Yu
University of Illinois at Chicago
psyu@cs.uic.edu

PP1

Citation Prediction in Heterogeneous Bibliographic Networks

To reveal information hiding in link space of bibliographical networks, link analysis has been studied from different perspectives in recent years. In this paper, we address a novel problem namely citation prediction, that is: given information about authors, topics, target publication venues as well as time of certain research paper, finding and predicting the citation relationship between a query paper and a set of previous papers. Considering the gigantic size of relevant papers, the loosely connected citation network structure as well as the highly skewed citation relation distribution, citation prediction is more challenging than other link prediction problems which have been studied before. By building a meta-path based prediction model on a topic discriminative search space, we here propose a two-phase citation probability learning approach, in order to predict citation relationship effectively and efficiently. Experiments are performed on real-world dataset with comprehensive measurements, which demonstrate that our framework has substantial advantages over commonly used link prediction approaches in predicting citation relations in bibliographical networks.

Xiao Yu, Quanquan Gu
UIUC
xiaoyu1@illinois.edu, qgu3@illinois.edu

Mianwei Zhou
University of Illinois at Urbana Champaign
zhou18@illinois.edu

Jiawei Han
UIUC
hanj@illinois.edu

PP1

Mining Multi-Label Data Streams Using Ensemble-Based Active Learning

Data stream classification has drawn increasing attention from the data mining community in recent years, where a large number of stream classification models were proposed. However, most existing models were merely focused on mining from single-label data streams. Mining from multi-label data streams has not been fully addressed yet. On the other hand, although some recent work touched the multi-label stream mining problem, they never consider the expensive labeling cost issue, preventing them from real-world applications. To this end, we study, in this paper, a challenging problem that mining from multi-label data streams with limited labeling resource. Specifically, we propose an ensemble-based active learning framework to handle the large volume of stream data, expensive labeling cost and concept drifting problems on multi-label data streams. Experiments on both synthetic and real world data sets demonstrate the performance of the proposed method.

Peng Wang, Peng Zhang, Li Guo
Institute of Information Engineering
Chinese Academy of Sciences
peng860215@gmail.com, zhangpeng04@gmail.com, guoli@ict.ac.cn

PP1

Feature Selection for High-Dimensional Integrated Data

Motivated by the problem of identifying correlations between genes or features of two related biological systems, we propose a model of *feature selection* in which only a subset of the predictors X_t are dependent on the multidimensional variate Y, and the remainder of the predictors constitute a "noise set' X_u independent of Y. Using Monte Carlo simulations, we investigated the relative performance of two methods: thresholding and singular-value decomposition, in combination with stochastic optimization to determine “empirical bounds’ on the small-sample accuracy of an asymptotic approximation. We demonstrate utility of the thresholding and SVD feature selection methods with respect to a recent infant intestinal gene expression and metagenomics dataset.

Charles Y. Zheng
Texas A and M
charles.y.zheng@gmail.com

Scott Schwartz
Texas Agrilife
sschwartz@ag.tamu.edu

Robert Chapkin
Texas A and M
Integrative Nutrition and Complex Diseases
r-chapkin@tamu.edu

Raymond Carroll
Texas A and M
Statistics
carroll@stat.tamu.edu

Ivan Ivanov
Texas A and M
Veterinary Physiology and Pharmacology
iivanov@cvm.tamu.edu

Organizer and Speaker Index

2012 SIAM
International Conference
on **DATA MINING**

April 26-28, 2012

Disney's Paradise Pier Hotel
Anaheim, California, USA

Italicized names indicate session organizers.

Italicized names indicate session organizers.

Notes

SDM12 Budget

Conference Budget
April 26-28, 2012
Anaheim, CA

Expected Paid Attendance 225

Revenue	
Registration	$85,595
Total	$85,595
Direct Expenses	
Printing	$1,500
Organizing Committee	$1,700
Invited Speaker	$10,400
Food and Beverage	$15,700
Telecomm	$2,100
AV and Equipment (rental)	$7,300
Room (rental)	$0
Advertising	$900
Conference Staff Labor	$17,700
Proceedings	$5,300
Other (supplies, staff travel, freight, exhibits, misc.)	$3,900
Total Direct Expenses:	$66,500
Support Services: *	
Services covered by Revenue	$19,095
Services covered by SIAM	$30,121
Total Support Services:	$49,216
Total Expenses:	$115,716

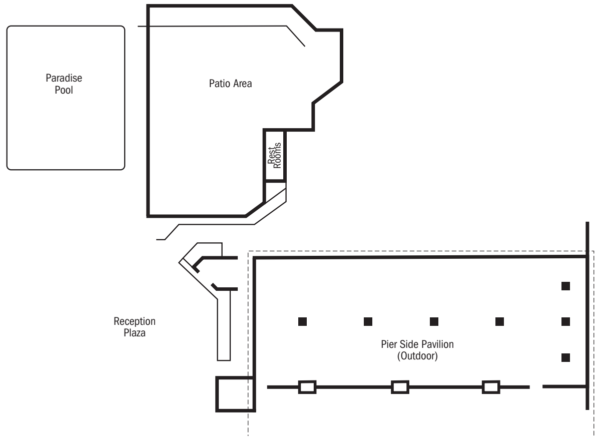
* Support services includes customer service (who handle registration), accounting, computer support, shipping, marketing and other SIAM support staff. It also includes a share of the computer systems and general items (building expenses in the SIAM HQ).

Disney's Paradise Pier Hotel Map

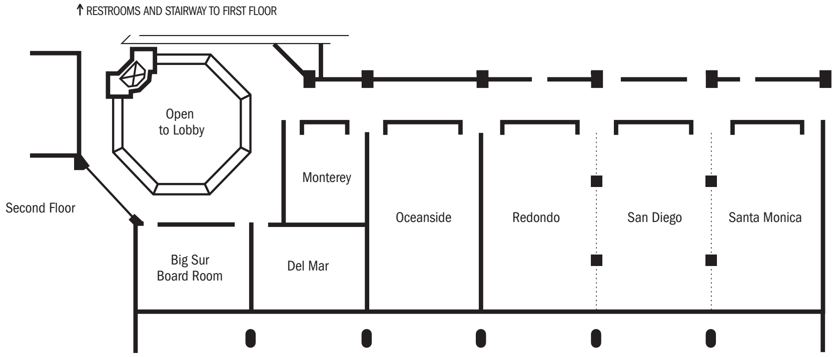
Disney's PARADISE PIER HOTEL

CONVENTION CENTER

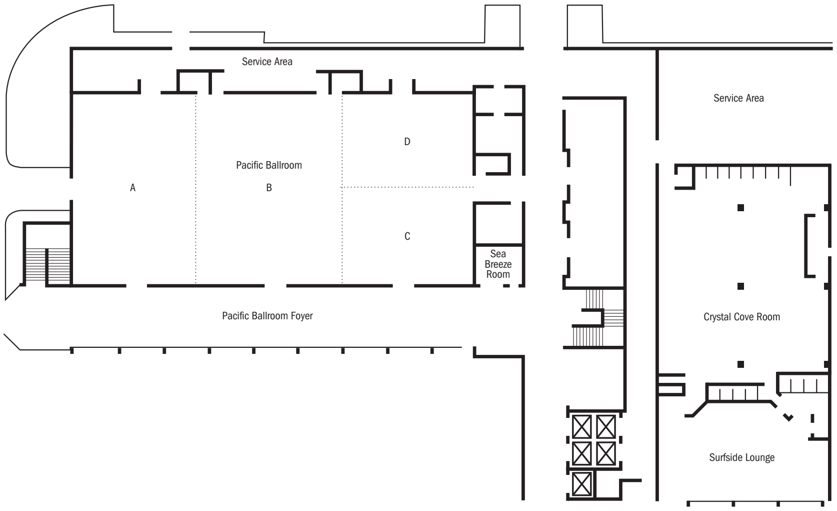
PIER SIDE PAVILION



SECOND FLOOR MEETING ROOMS



PACIFIC BALLROOM



FSC logo text box indicating size & layout of logo. Conlins to insert logo.