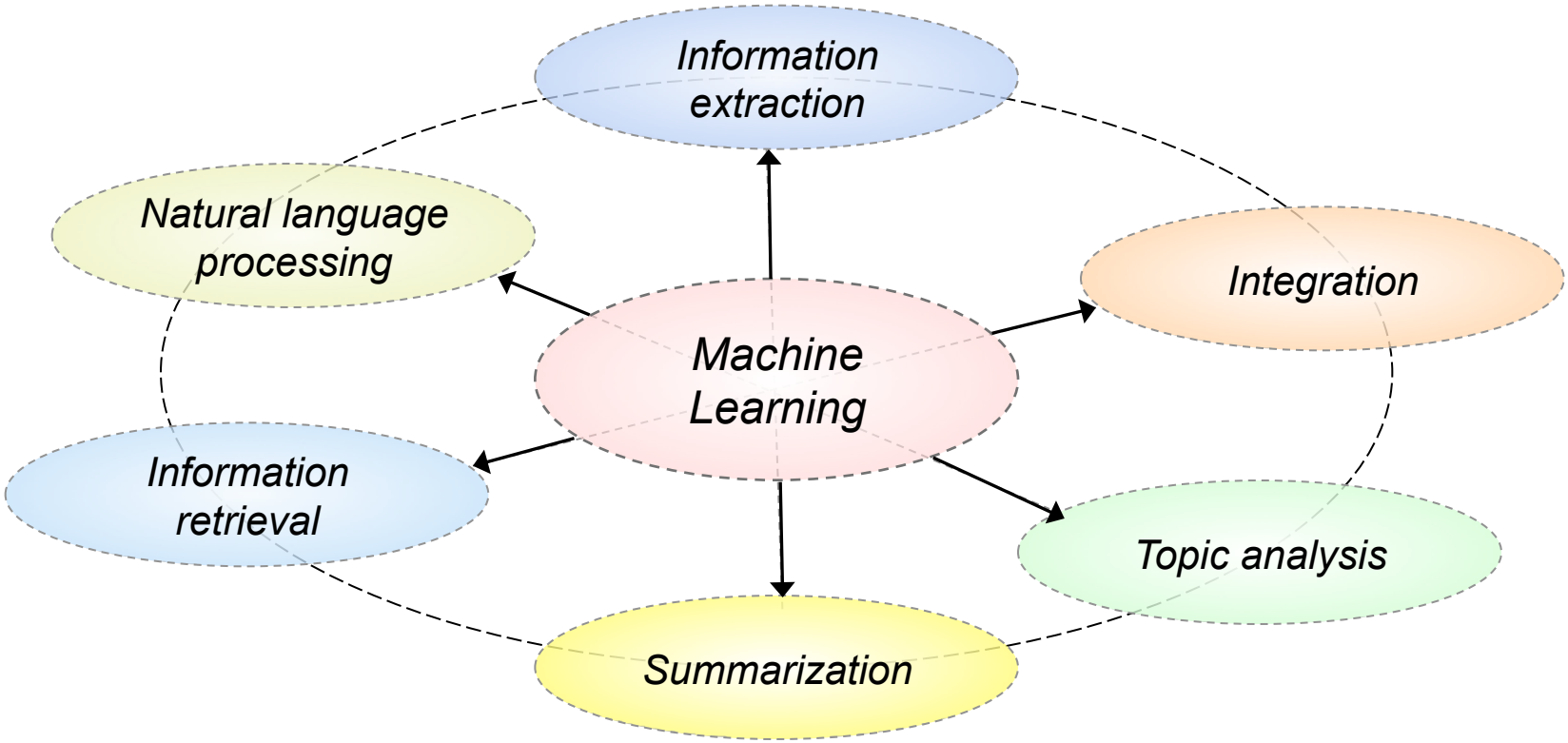
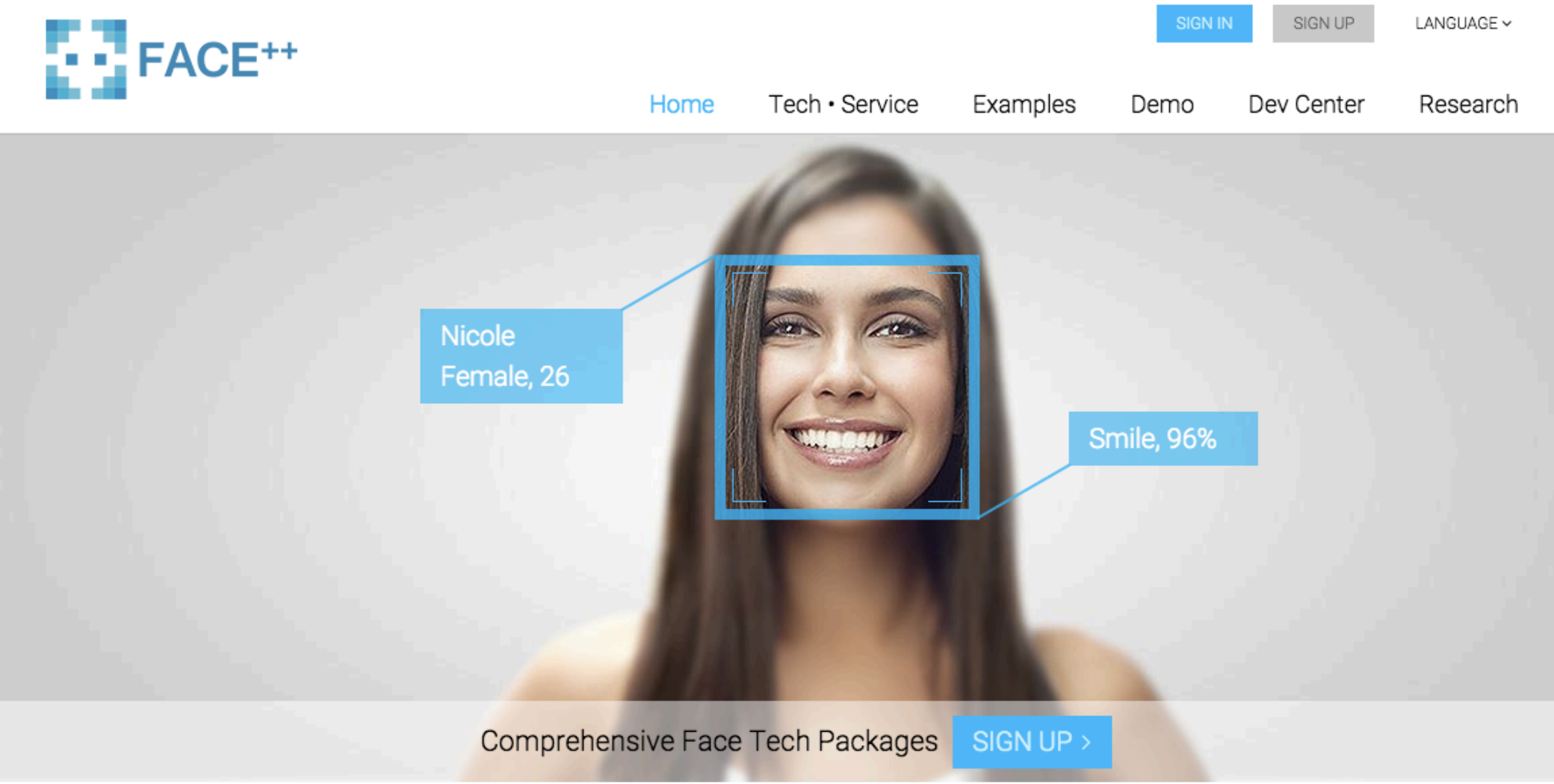
Why Machine Learning?

Jie Tang
Department of Computer Science & Technology
Tsinghua University
2015

Machine Learning

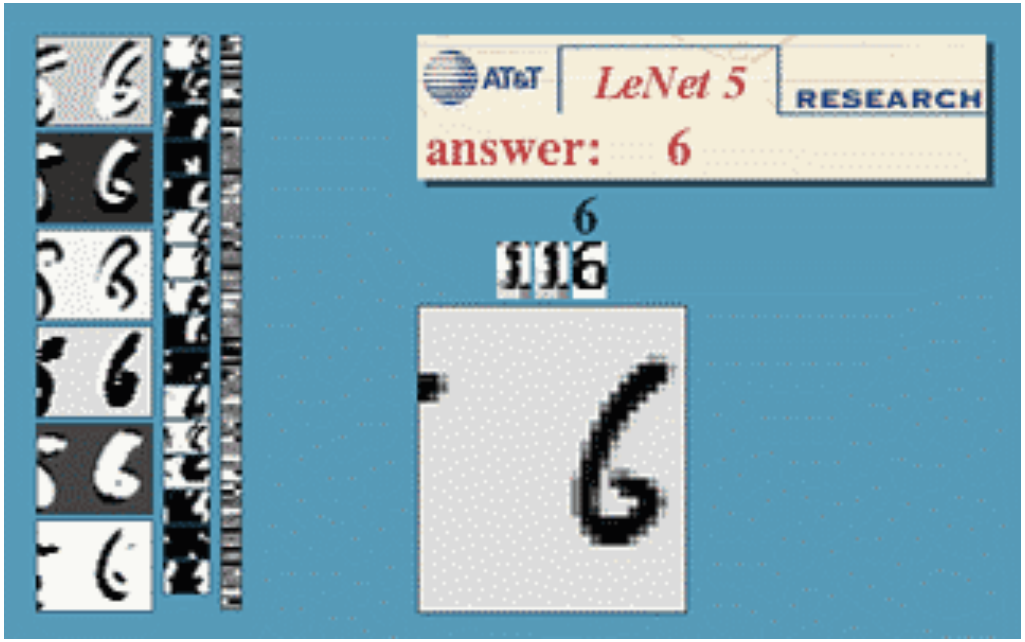


Face Recognition



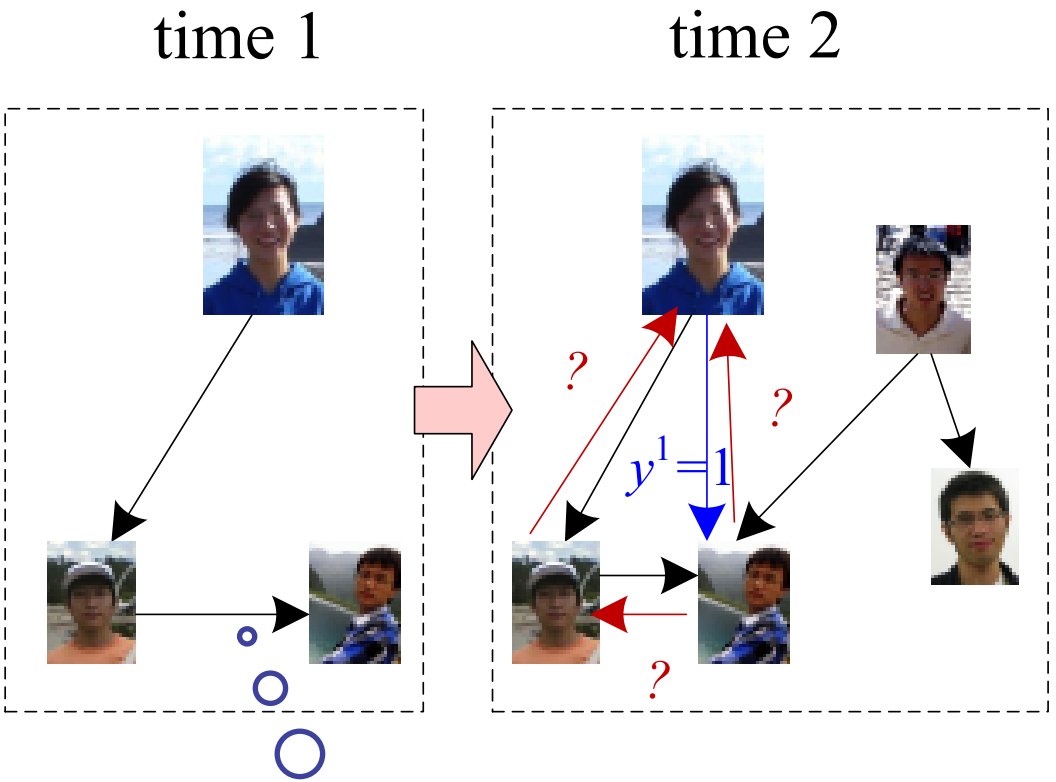
Handwritten Recognition

- Convolutional neural networks for handwritten recognition



- http://yann.lecun.com/exdb/lenet/index.html

Follow back Prediction



When you follow a friend in Twitter, how likely he will follow back?

IE from Web Page (IJCAI'07)

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open -source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

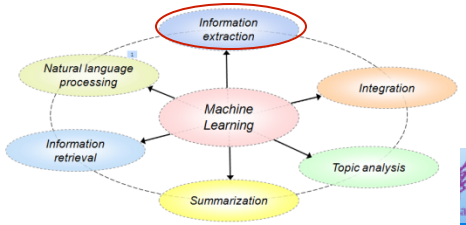
* Microsoft
Gates

* Microsoft
Bill Veghte

* Microsoft
VP

Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..



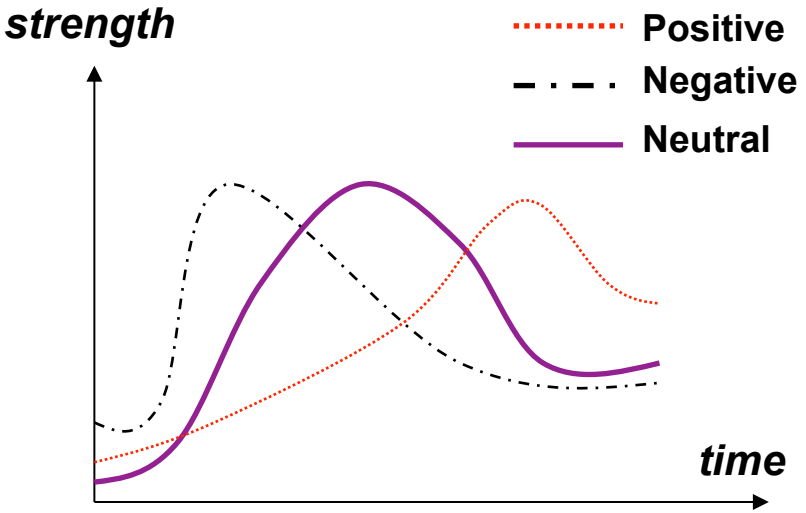
Sentiment Summary (WWW'08)

Topic-sentiment summary

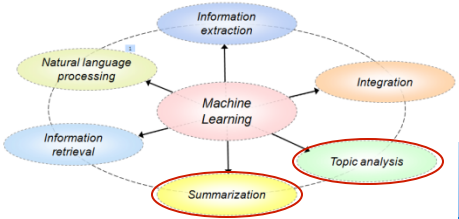
Query: *Dell Laptop*

	positive	negative	neutral
Topic 1 (*Price*)	• it is the best site and they show Dell coupon code as early as possible	• Even though Dell's price is cheaper, we still don't want it. • ……	• mac pro vs. dell precision: a price comparis.. • DELL is trading at $24.66
Topic 2 (*Battery*)	• One thing I really like about this Dell battery is the Express Charge feature.	• my Dell battery sucks • Stupid Dell laptop battery • ……	• i still want a free battery from dell.. • ……

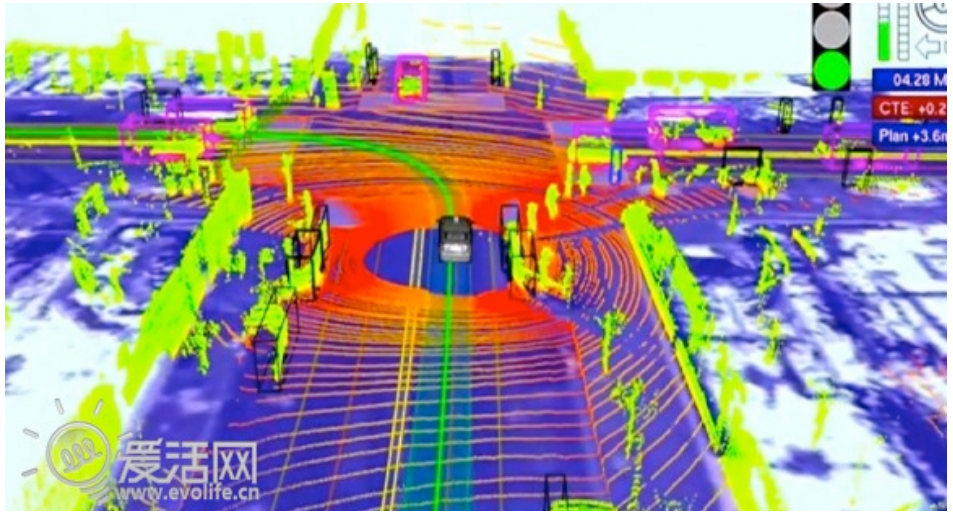
Topic-sentiment dynamics (Topic = *Price*)

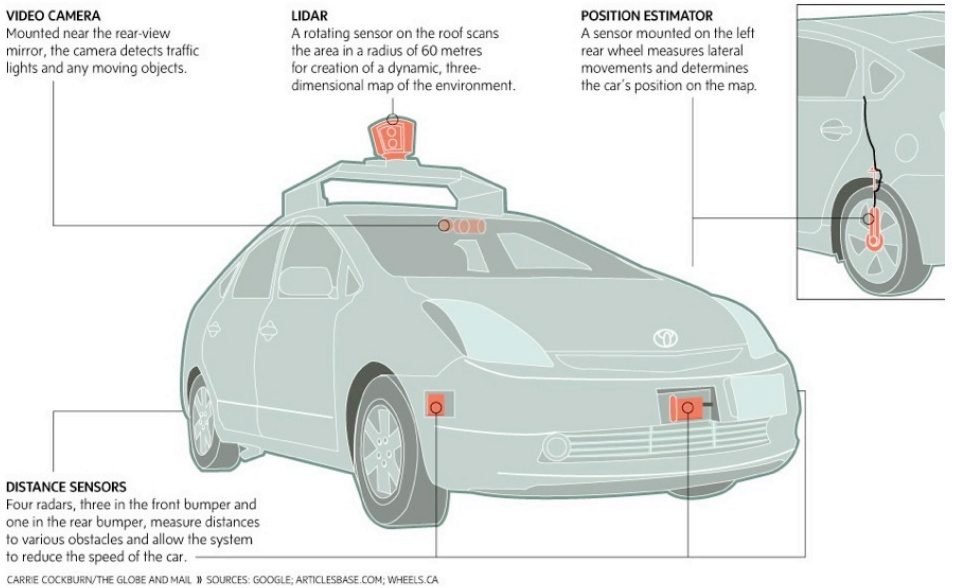


Unsupervised/Semi-supervised discovery of topics and different sentiments of the topics



Google driverless car





Google's self-driving cars last month hit a 480,000-miles-driven milestone.

--2012/09

The system combines information gathered from Google Street View with artificial intelligence software that combines input from video cameras inside the car, a LIDAR sensor on top of the vehicle, radar sensors on the front of the vehicle and a position sensor attached to one of the rear wheels that helps locate the car's position on the map

IPhone Siri



Siri understands what you say, knows what you mean, and even talks back.

Siri (pronounced /ˈsɪri/) is an intelligent personal assistant and knowledge navigator which works as an application for Apple's iOS. The application uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. Apple claims that the software adapts to the user's individual preferences over time and personalizes results, and performing tasks such as finding recommendations for nearby restaurants, or getting directions

IBM Watson

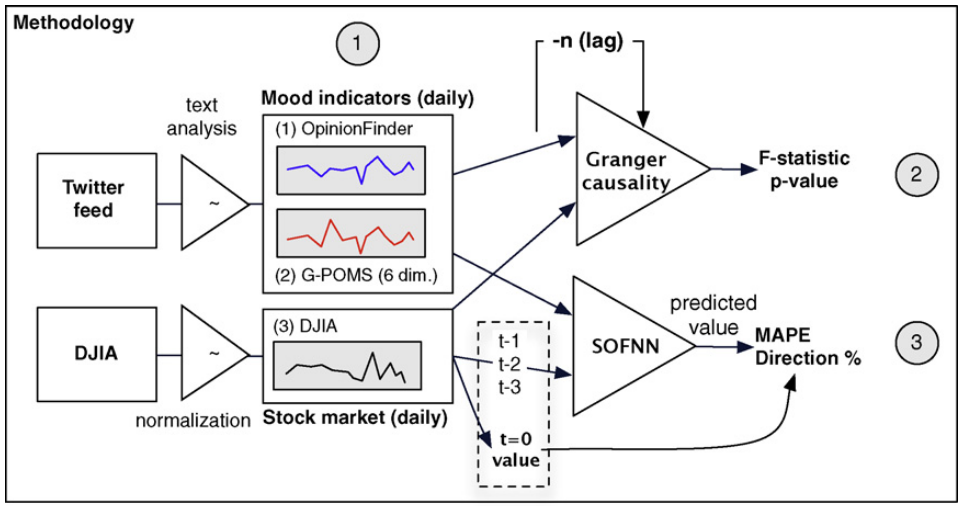


In 2011, as a test of its abilities, Watson competed on the quiz show Jeopardy!, in the show's only human-versus-machine match-up to date

Watson is a Question answering (QA) computing system built by IBM. IBM describes it as "an application of advanced Natural Language Processing, Information Retrieval, Knowledge Representation and Reasoning, and Machine Learning technologies to the field of open domain question answering" which is "built on IBM's *DeepQA* technology for hypothesis generation, massive evidence gathering, analysis, and scoring

Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia, but was not connected to the Internet during the game. For each clue, Watson's three most probable responses were displayed on the television screen. Watson consistently outperformed its human opponents on the game's signaling device.

Stock Prediction using Twitter Mood



The paper investigates whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can indeed be used to predict the stock market.

Two tools are used to measure variations in the public mood from Twitter's tweets from February 28, 2008 to December 19, 2008. The resulting public mood time series are correlated to the Dow Jones Industrial Average (DJIA) to assess their ability to predict changes in the DJIA over time.

The results indicate that the prediction ccuracy of standard stock market prediction models is significantly improved when certain mood dimensions are included.

Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, Volume 2, Issue 1, March 2011, Pages 1–8

What is Machine Learning?

Data Structures

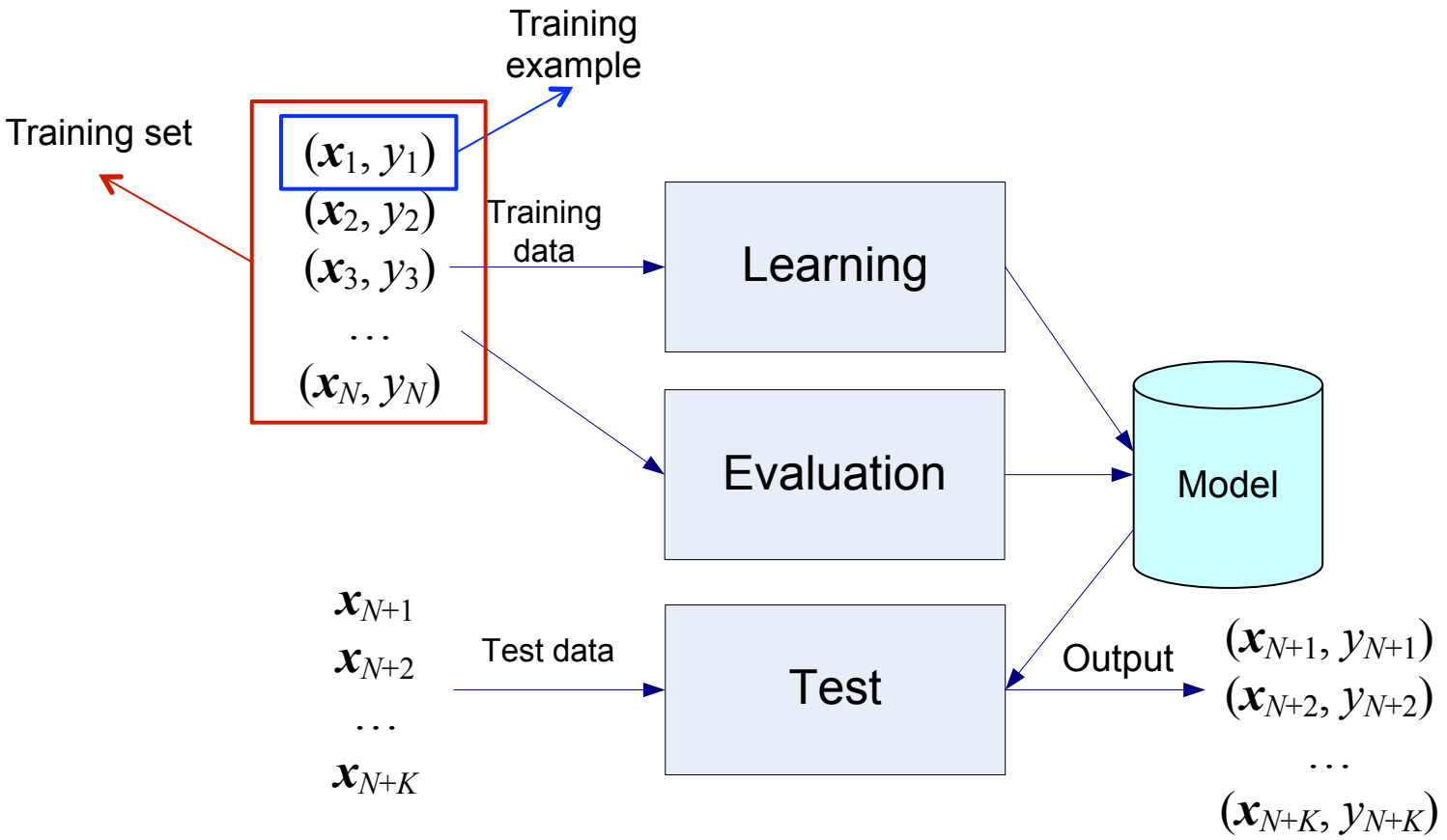
- An example with *p* features

$$\boldsymbol{x}_i=[x_{i1}, x_{i2},\ldots, x_{ip}]$$

- Data (set) matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nf} & \cdots & x_{Np} \end{bmatrix}$$

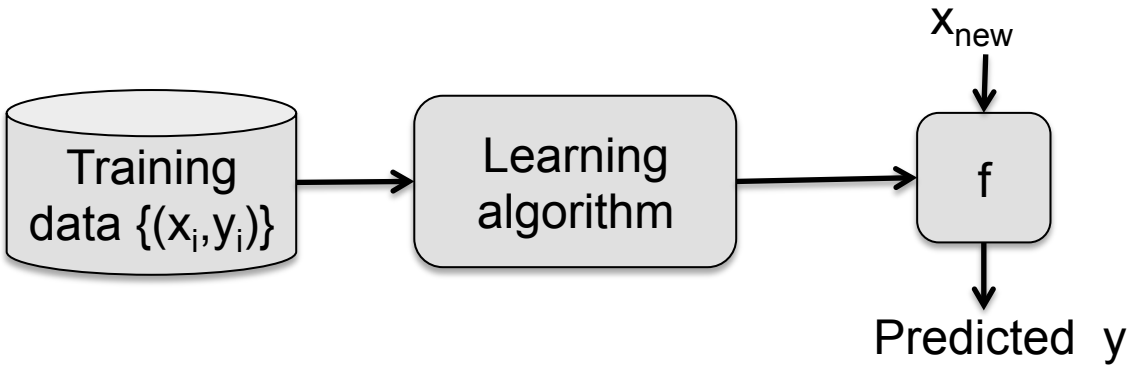
Supervised Learning



Supervised Learning

- Given a training set $S=\{(x_1, y_1),(x_2, y_2),\ldots,(x_N, y_N)\}$, and $x_i \in X=R^m$, $i=1,2,\ldots,N$
- To learn a function $f(x)$, which can best fit the data

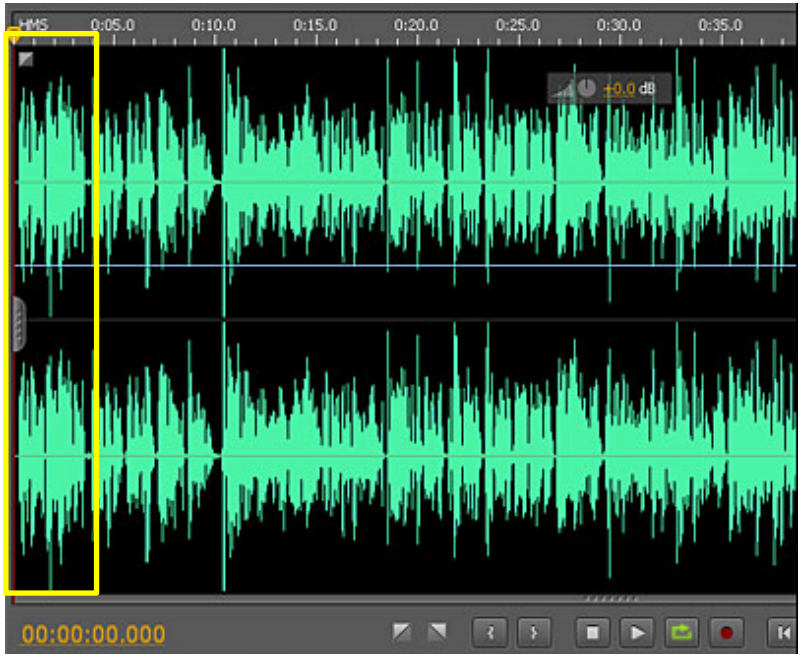
$$y = f(x)$$



Regression: When y is continuous (i.e., $y_i \in Y=R$), we call the learning problem as regression.

Classification: When y can take on only a small number of discrete values (e.g., the binary case $y_i \in Y=\{1,-1\}$), we call the learning problem as classification.

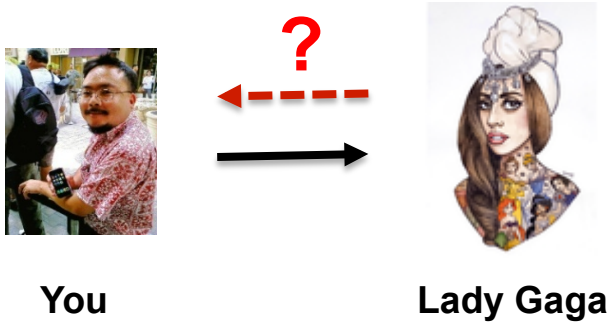
Siri's first task: Speech-to-Text



Who is the presidetn of US?

$\boldsymbol{x}_1$={Hz, Position, Volume} y_1 = "Who"

Example: Follow-back Prediction



When a follow b, we have a new instance x_i;

If b immediately follow back a, we say y_i=1;

otherwise y_i=0;

x_i={#fans of a, #fans of b, #common friends between a and b}

Example: IE using CRF

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open -source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

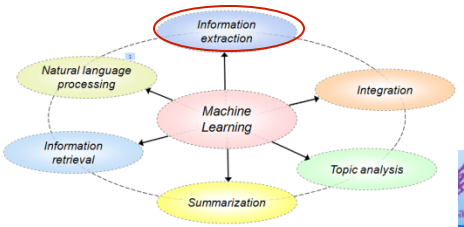
* Microsoft
Gates

* Microsoft

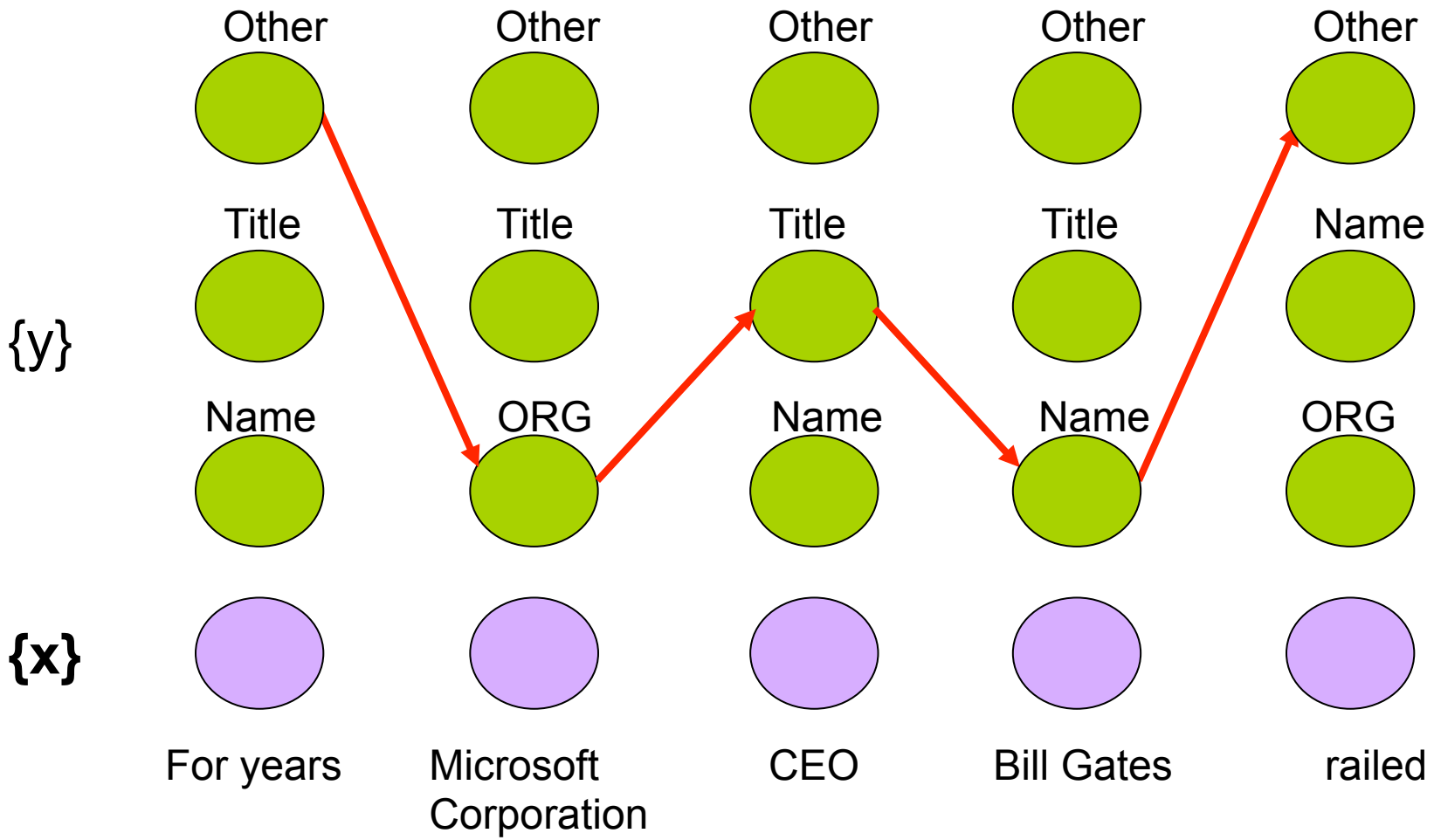
* Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..



Example: IE using CRF (cont.)



x_1={For, years, PREP, NOUN} y_1 = Other

Our Stories on ML

Email Data Cleaning (KDD'05)

From: SY <sandeep....@gmail.com> - Find messages by this author
Date: Mon, 4 Apr 2005 11:29:28 +0530
Subject: Re: ..How to do addition??

Hi Ranger,
Your design of Matrix
class is not___ good.
what are you doing with two
matrices in a single class?make class Matrix as follows

import java.io.*;

Extra line break

Missing space
Extra space
Missing period

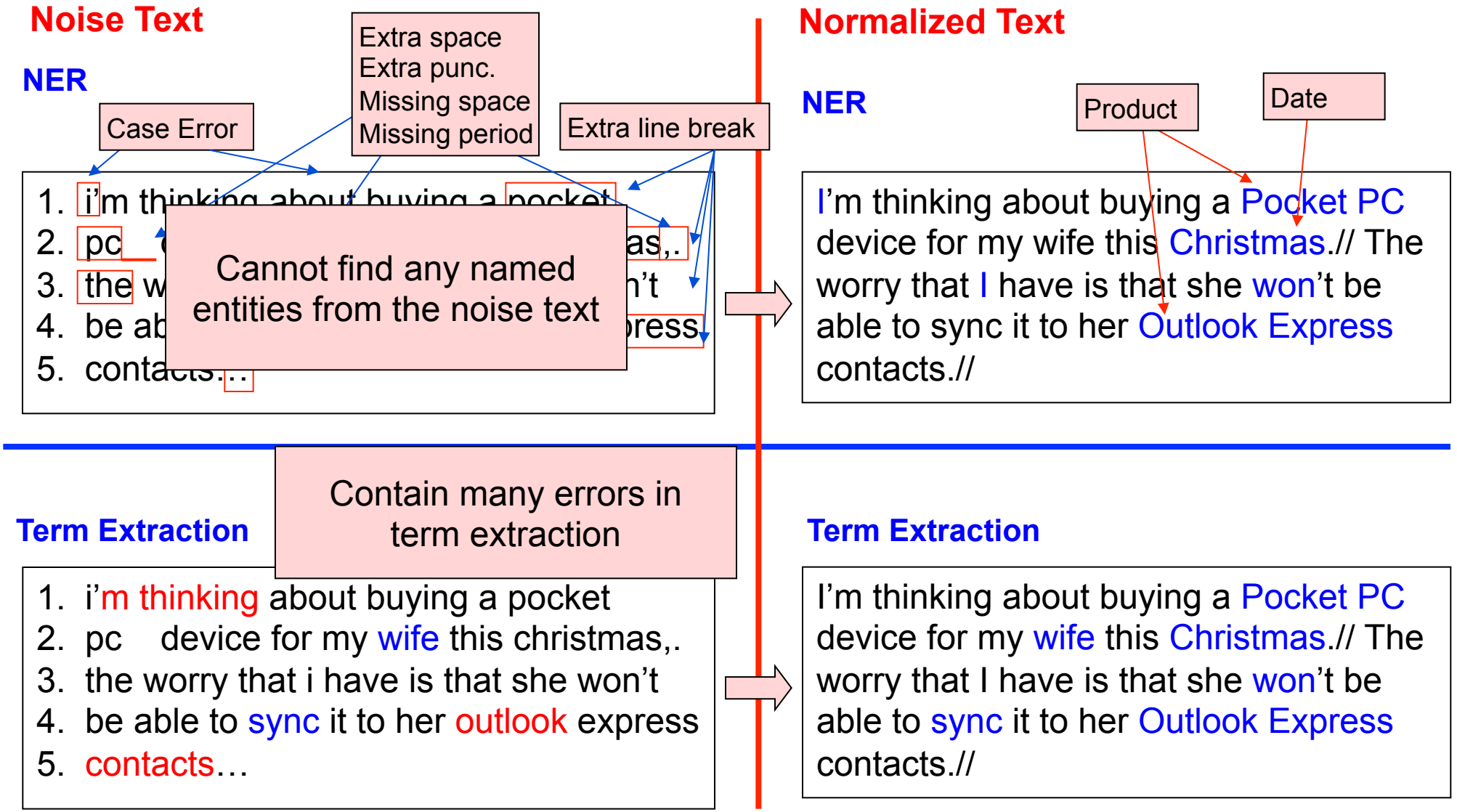
Case errors

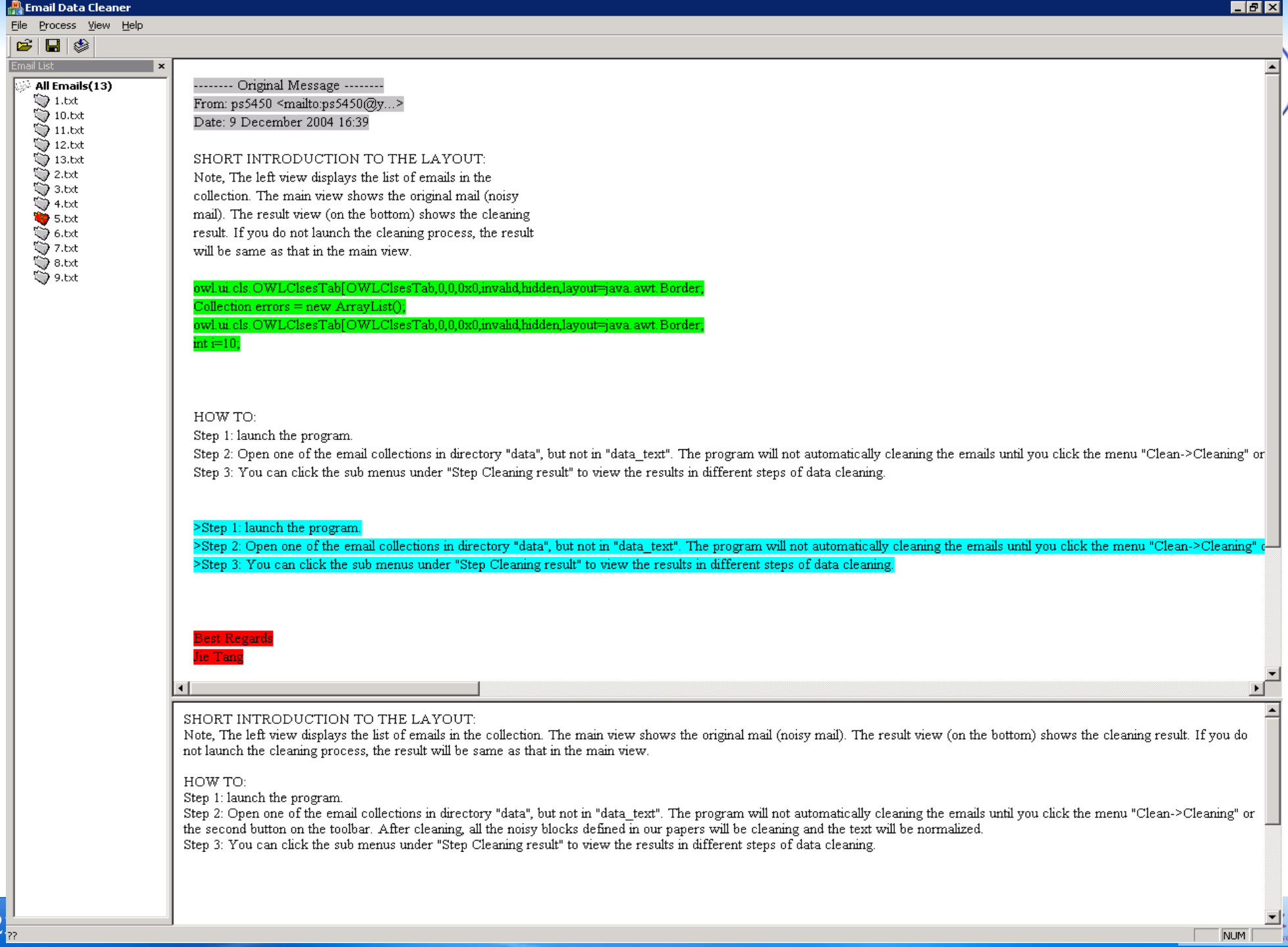
Hi Ranger,
Your design of Matrix class is not good. What are you doing with two matrices in a single class?
Make class Matrix as follows:

Tel: 011-243600808
Homepage: http://www.it.com/~Sandeep/

On Apr 3, 2005 5:33 PM, ranger <asiri....@gmail.com> wrote:
> Hi... I want to perform the addtion in my Matrix class. I got the program to
> enter 2 Matricx and diaplay them. Hear is the code of the Matrix class and
> TestMatrix class. I'm glad If anyone can let me know how to do the addition.....Tnx

Text Normalization (ACL'07)




Email Data Cleaner
File Process View Help
Email List
All Emails(13)
1.txt
10.txt
11.txt
12.txt
13.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
-------- Original Message --------
From: ps5450 <mailto:ps5450@y...>
Date: 9 December 2004 16:39
SHORT INTRODUCTION TO THE LAYOUT:
Note, The left view displays the list of emails in the collection. The main view shows the original mail (noisy mail). The result view (on the bottom) shows the cleaning result. If you do not launch the cleaning process, the result will be same as that in the main view.
owl.ui.cls.OWLClsesTab[OWLClsesTab,0,0,0x0,invalid,hidden,layout=java.awt.Border;
Collection errors = new ArrayList();
owl.ui.cls.OWLClsesTab[OWLClsesTab,0,0,0x0,invalid,hidden,layout=java.awt.Border;
int i=10;
HOW TO:
Step 1: launch the program.
Step 2: Open one of the email collections in directory "data", but not in "data_text". The program will not automatically cleaning the emails until you click the menu "Clean->Cleaning" or
Step 3: You can click the sub menus under "Step Cleaning result" to view the results in different steps of data cleaning.
>Step 1: launch the program.
>Step 2: Open one of the email collections in directory "data", but not in "data_text". The program will not automatically cleaning the emails until you click the menu "Clean->Cleaning"
>Step 3: You can click the sub menus under "Step Cleaning result" to view the results in different steps of data cleaning.
Best Regards
Jie Tang
SHORT INTRODUCTION TO THE LAYOUT:
Note, The left view displays the list of emails in the collection. The main view shows the original mail (noisy mail). The result view (on the bottom) shows the cleaning result. If you do not launch the cleaning process, the result will be same as that in the main view.
HOW TO:
Step 1: launch the program.
Step 2: Open one of the email collections in directory "data", but not in "data_text". The program will not automatically cleaning the emails until you click the menu "Clean->Cleaning" or the second button on the toolbar. After cleaning, all the noisy blocks defined in our papers will be cleaning and the text will be normalized.
Step 3: You can click the sub menus under "Step Cleaning result" to view the results in different steps of data cleaning.
??
NUM

Information is (tree-)structured…



Qiong (Joan) Luo

Assistant Professor
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon,
Hong Kong (Zipcode free!)

Office: Room *3554* (via lifts 25,26), Academic Building
Phone: (852) 2358-6995
Fax: (852) 2358-1477
Email: luo@cse.ust.hk
URL: http://www.cse.ust.hk/~luo

Bio | Entry@DBLP | Publications | Research | Services | Teaching & Advising

Qiong Luo's Short Bio

Last updated on August 23, 2008

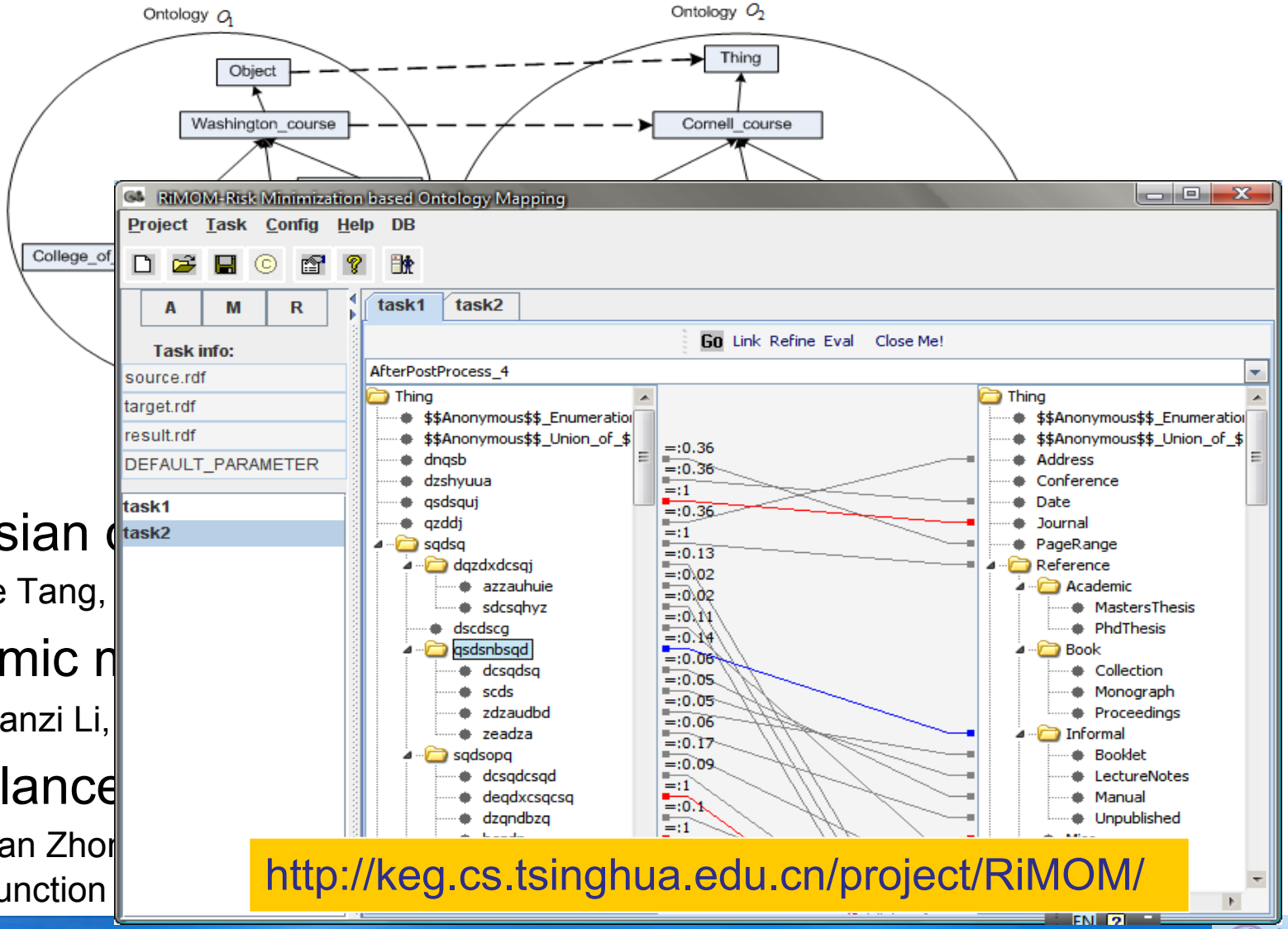
Qiong Luo is an Assistant Professor at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology (HKUST). Her research interests are in database systems, with a focus on architecture-conscious databases, query processing in pervasive computing and sensor networks, and web data management. Qiong received her Ph.D. in Computer Sciences from the University of Wisconsin-Madison in 2002, her M.S. and B.S. in Computer Sciences from Beijing (Peking) University, China in 1997 and 1992 respectively.

- Bingsheng He, Wenbin Fang, Qiong Luo, Naga K. Govindaraju, and Tuyong Wang. Mars: A MapReduce Framework on Graphics Processors. PACT 2008.
- Xiangye Xiao, Qiong Luo, Dan Hong, Hongbo Fu, Xing Xie, and Wei-Ying Ma. Browsing on Small Displays by Transforming Web Pages into Hierarchically Structured Sub-Pages. Accepted for publication at ACM Transactions on the Web (TWeb), 2008.
- Bingsheng He and Qiong Luo. Cache-Oblivious Databases: Limitations and Opportunities. ACM Transactions on Database Systems (TODS), June 2008.
- Qiong Luo, Jeffrey F. Naughton, Wenwei Xue. Form-based proxy caching for database-backed web sites: keywords and functions. VLDB J. 17(3): 489-513 (2008)

Academic Services

- Conference/Workshop organization:
  - **DaMoN**(International Workshop on Data Management for New Hardware), in conjuction with SIGMOD. Program Chair, 2008; General Chair, 2007.
  - **DMSN**(International Workshop on Data Management for Sensor Networks), in conjuction with VLDB. Member of Steering Committee, 2008; General Chair, 2007; Program Co-chair, 2006.
  - **ICDE**(IEEE International Conference on Data Engineering), Tutorial Co-chair, 2007.
  - **SAC**(ACM Symposium on Applied Computing), The Internet Data Management track. Track Co-chair, 2004.
- Program committee:
  - **APWeb** (Asia Pacific Web Conference) 2004, 2006-07.
  - **COMAD** (International Conference on Management of Data) 2005-06.
  - **DaMoN**(International Workshop on Data Management for New Hardware), in conjuction with SIGMOD. 2006
  - **DASFAA** (International Conference on Database Systems for Advanced Applications), 2005, 2006 (Industrial Track), 2008.

Ontology mapping



- Bayesian
  - Jie Tang,
- Dynamic
  - Juanzi Li,
- Unbalance
  - Qian Zhor
    Function

http://keg.cs.tsinghua.edu.cn/project/RiMOM/

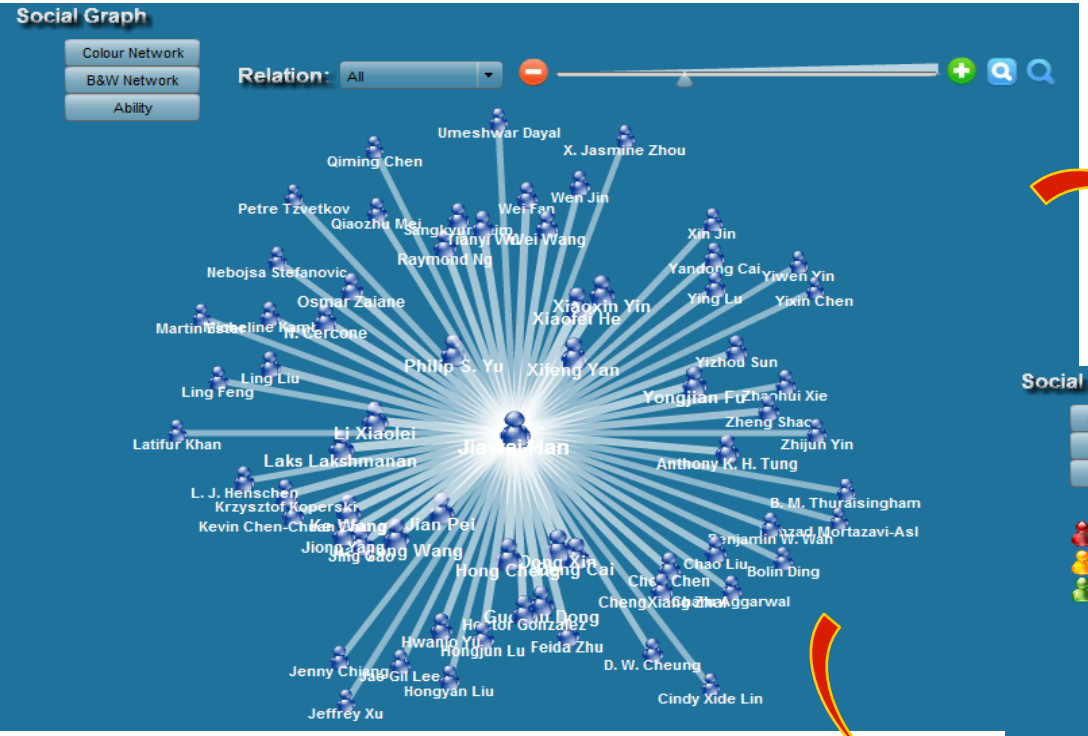
Learning to infer social ties

- Nobody exists merely in one social network.
  - Public network vs. private network
  - Business network vs. family network
- However, existing networks (e.g., Facebook and Twitter) are trying to lump everyone into one big network
  - FB tries to solve this problem via lists/groups
  - however…
- Google circles



Example 1. From BW to Color

(KDD'10)





Example 2. From BW to Color

(PKDD'11 best student paper runnerup)

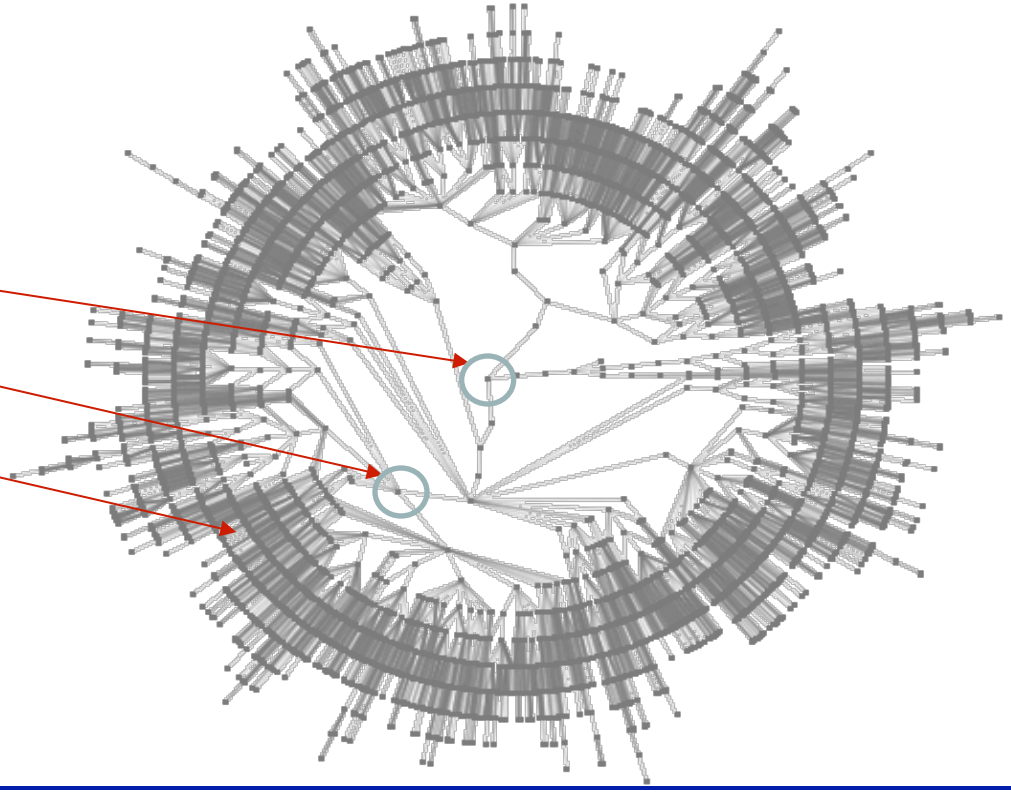
Enterprise email network

How to infer

CEO

Manager

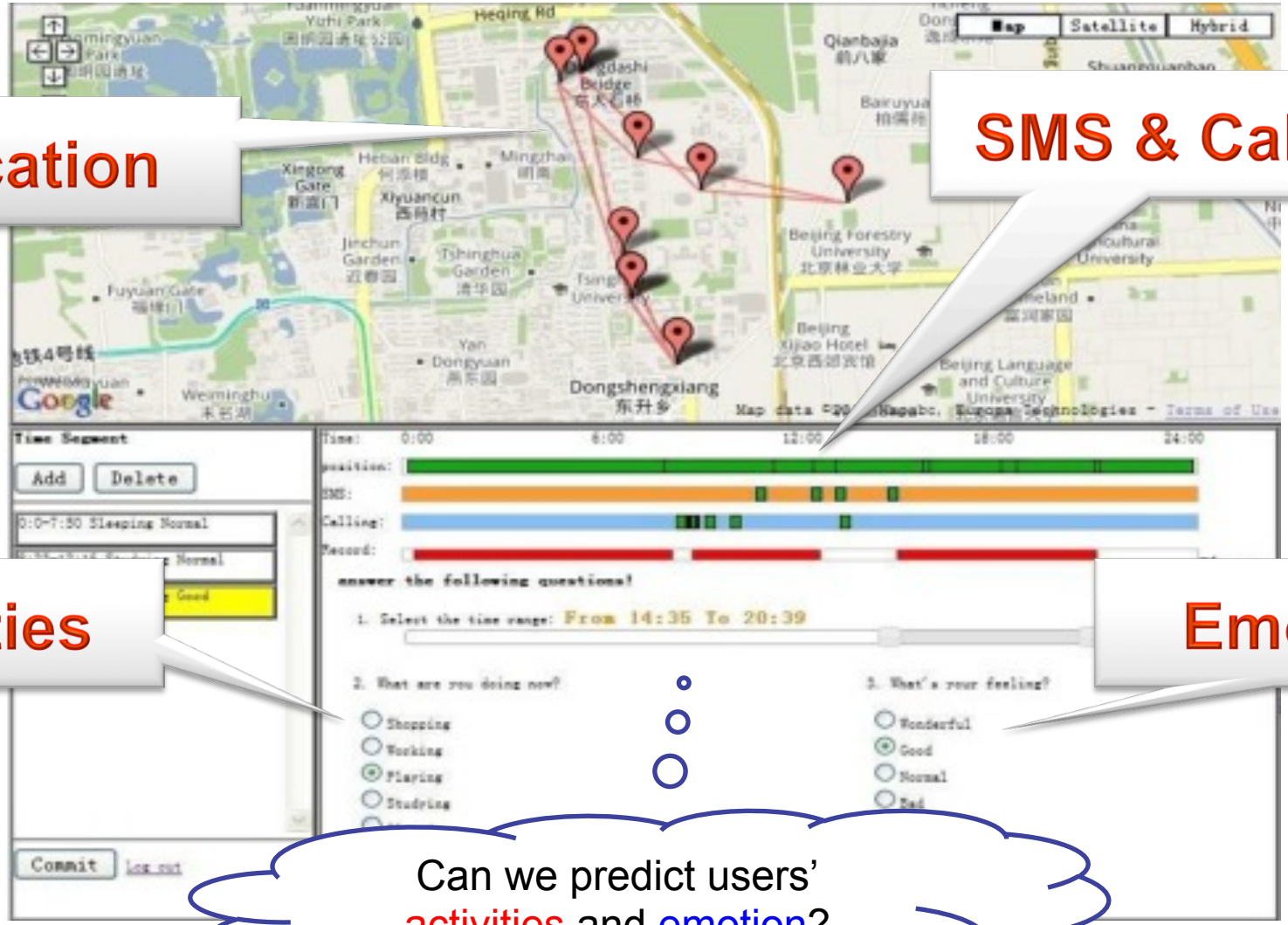
Employee



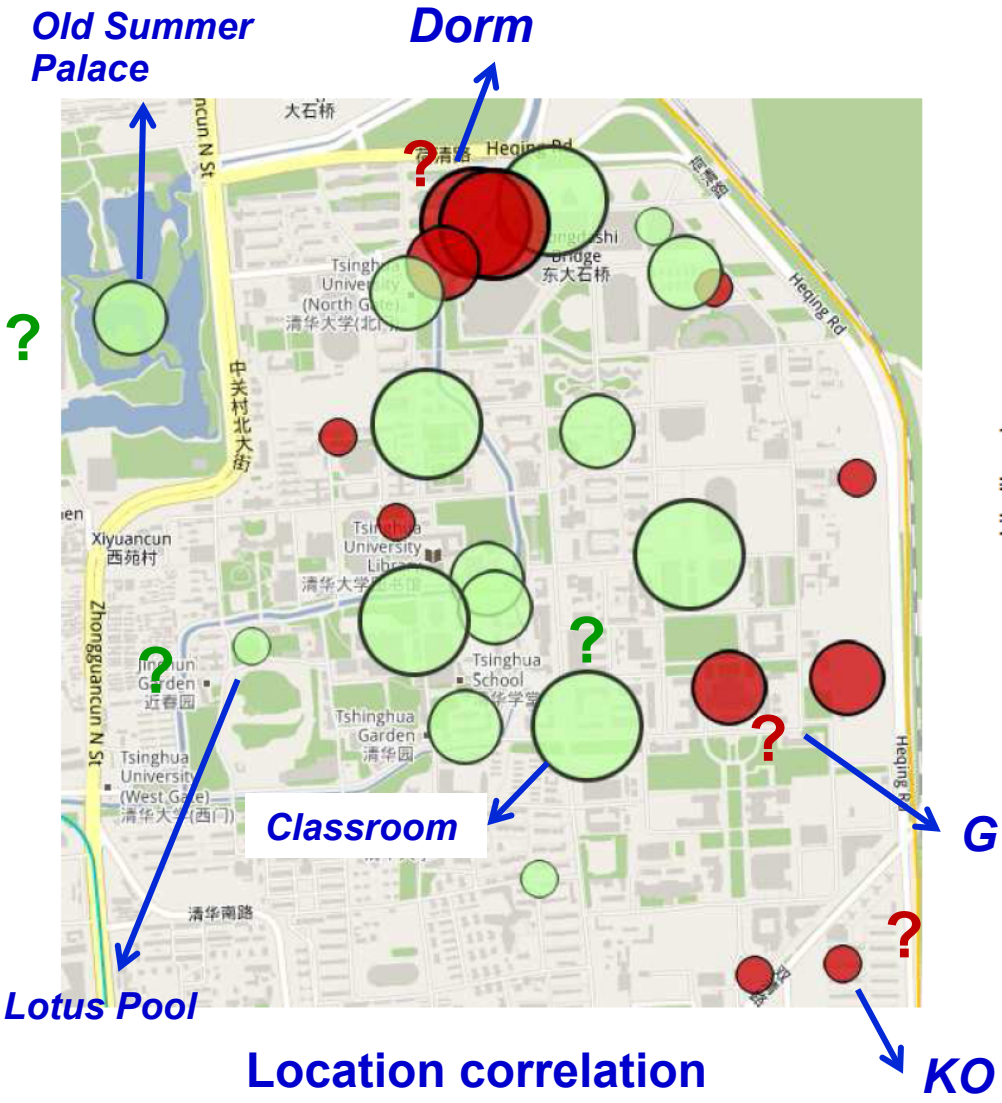
User interactions may form *implicit groups*

Learning to predict actions and emotions

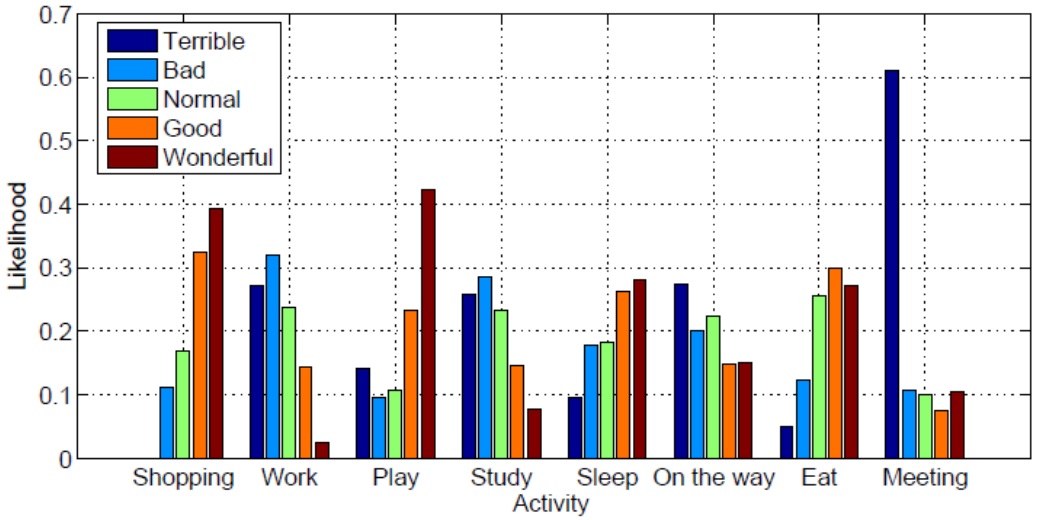
(KDD'10, ICDM'10, IEEE TAC)



Observations

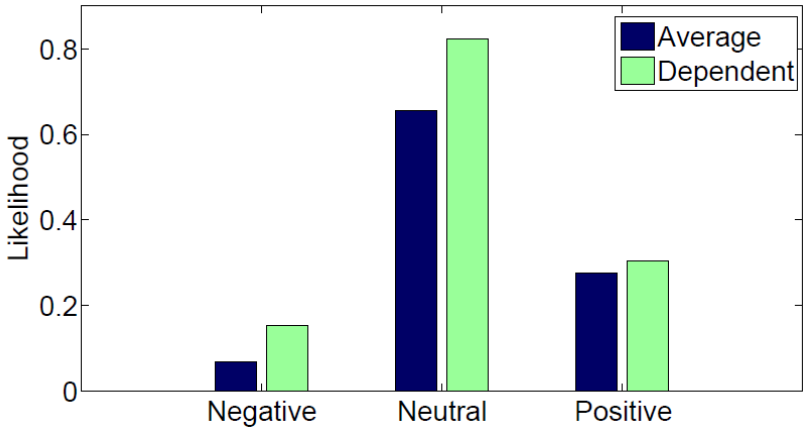


Location correlation
(Red-happy)



Activity correlation

Observations (cont.)

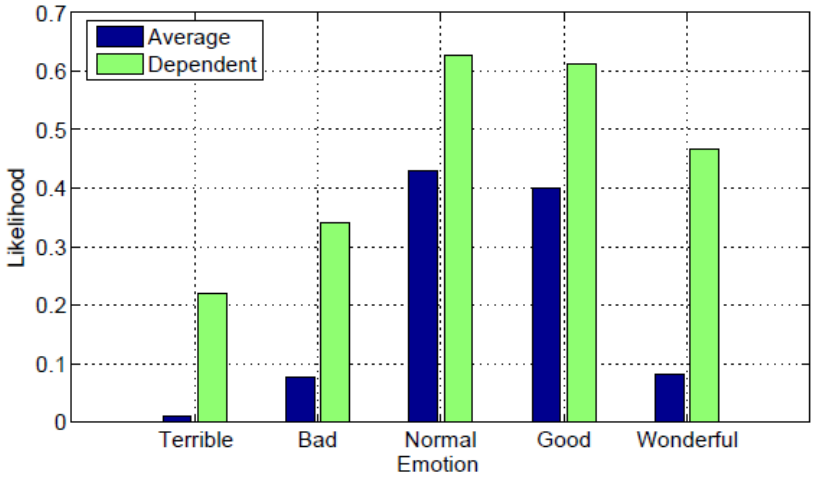


(a) Social correlation

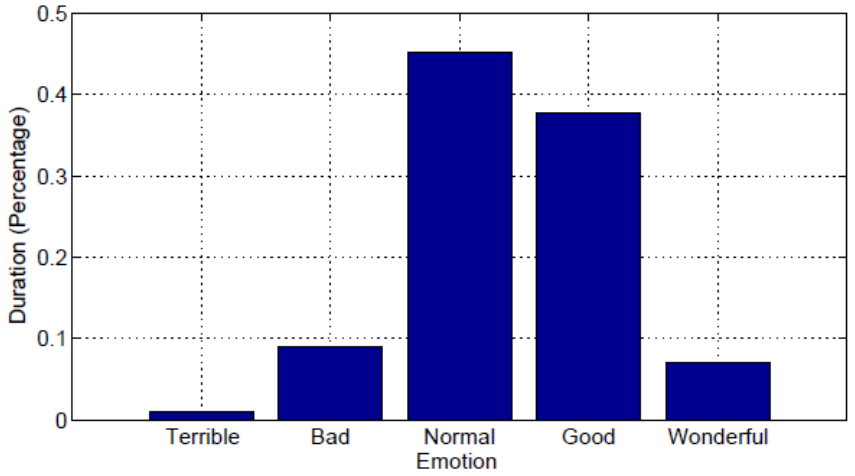


(a) Implicit groups by emotions

Social correlation



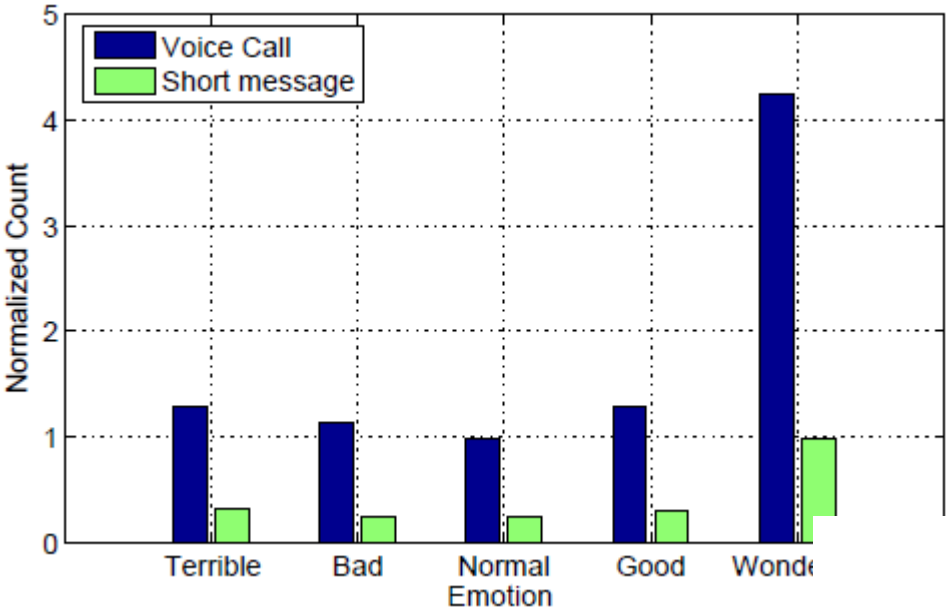
(a) Temporal correlation



(b) Time duration

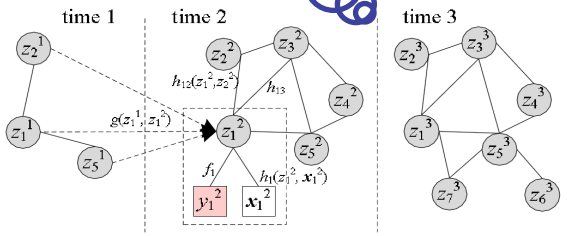
Temporal correlation

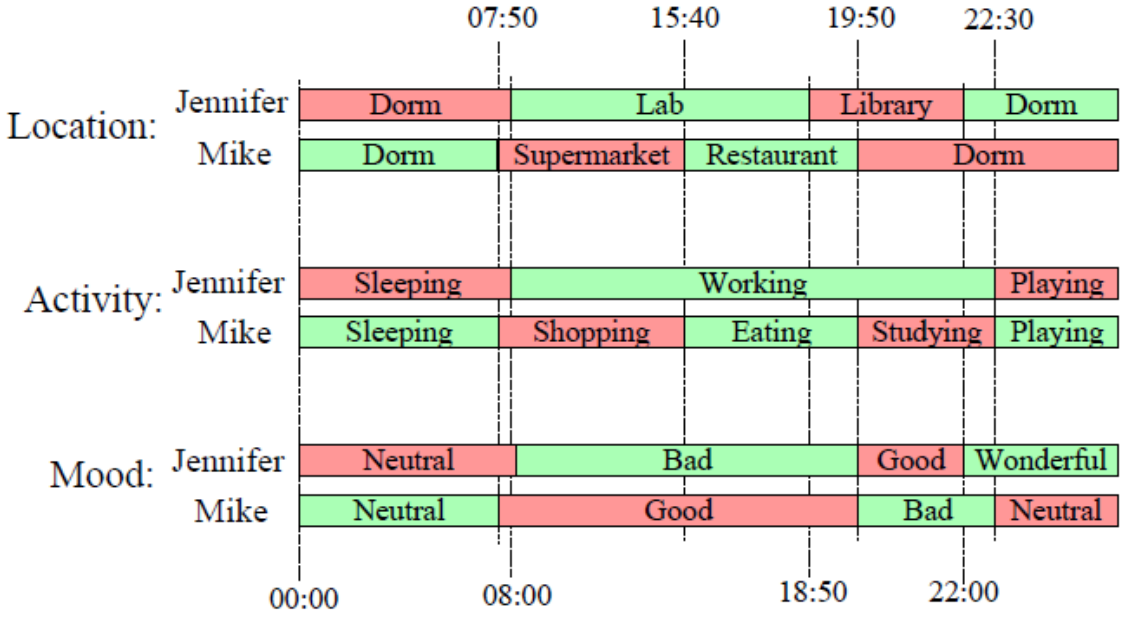
Observations (cont.)



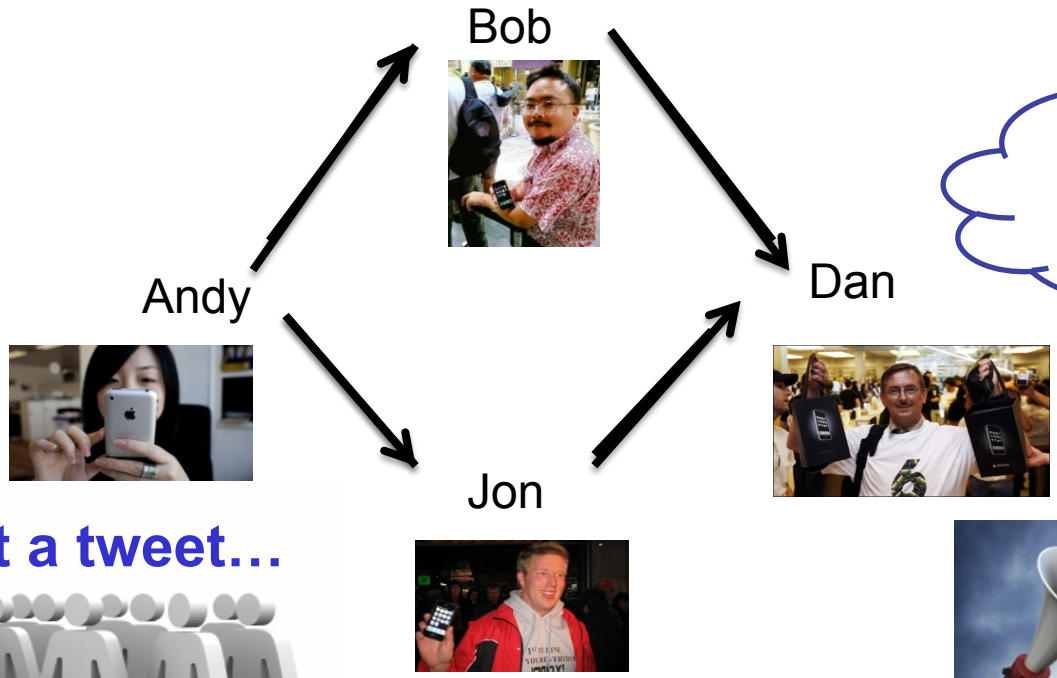
Calling (SMS) correlation

We should not split the data into different time windows …





Retweet Predicting (CIKM'10)



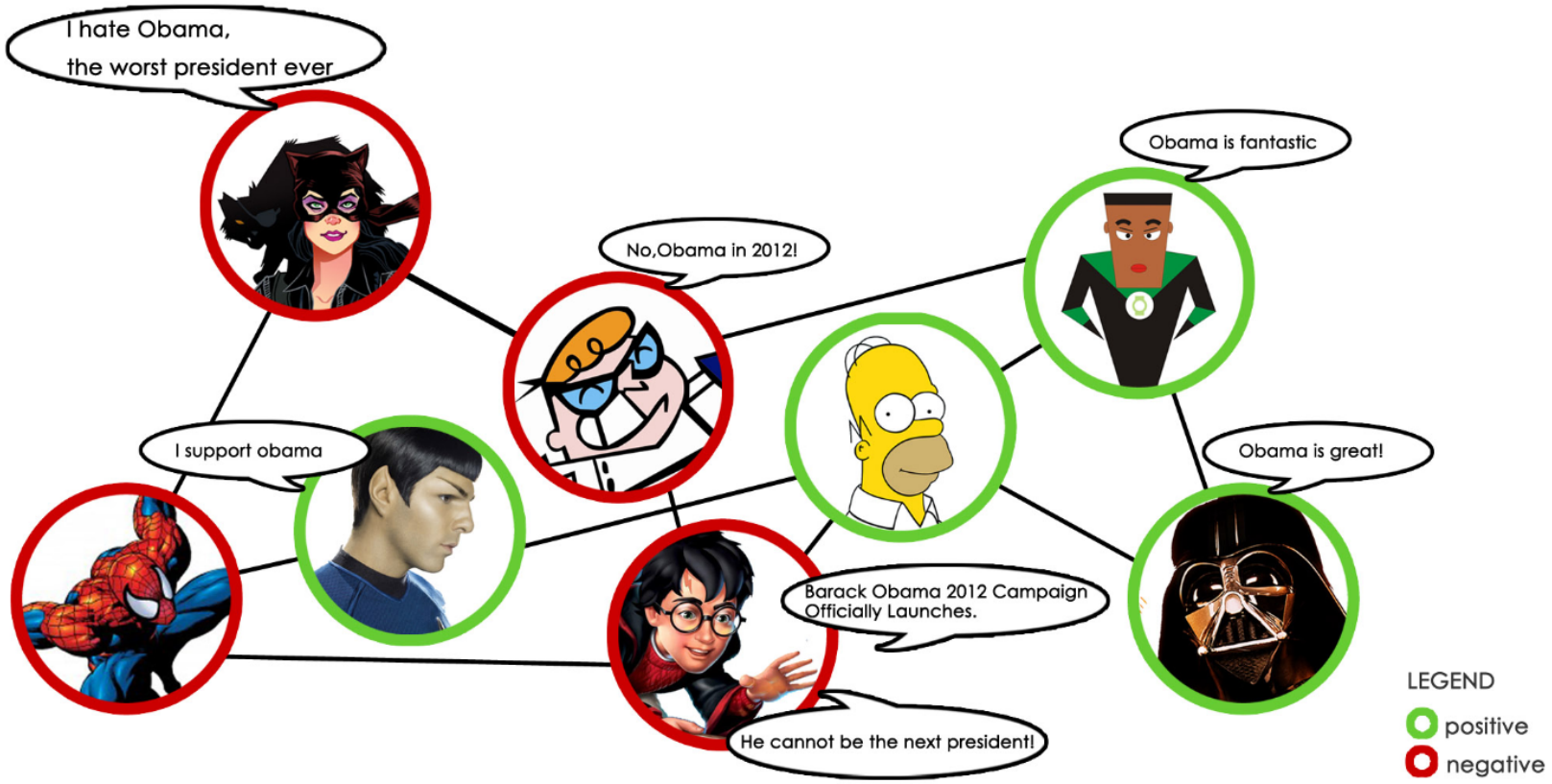
Who will retweet it?

When you post a tweet…





Sentiment Analysis in Social Networks
(KDD'11)



Modelling Paying Behavior in Game Social Networks

Zhanpeng Fang[+], Xinyu Zhou[+], Jie Tang[+], Wei Shao[#], A.C.M. Fong[*], Longjun Sun[#], Ying Ding[-], Ling Zhou[+], and Jarder Luo[+]

[+]Tsinghua University

[#]Tencent Corporation

[*]Auckland University of Technology

[-]Indiana University

Billion Dollar Industry

- Facebook[1]
  - 250 million monthly players
  - 200 games with >1 million active players
  - 12% revenue

- Tencent[2] (Market Cap: ~150B $)
  - >400 million players
  - 50% revenue



[1] Facebook 2013 First Quarter Report
[2] Tencent 2013 Anual Report

Billion Dollar Industry

- Facebook[1]
  - 250 million monthly players
  - 200 games with >1 million active players

Not only keep players playing, but also make them pay.

  - >400 million players
  - 50% revenue



[1] Facebook 2013 First Quarter Report
[2] Tencent 2013 Anual Report

What we do

- Given users' data in online games, predict:

Free users -> Paying users

What we do

- Given users' data in online games, predict:

Free users -> Paying users

- Our goal:
  - Fundamental factors
  - Social effect
  - Predictive model

Two games: DNF

- Dungeon & Fighter Online (DNF)
  - Fight enemies by individuals or groups
  - 400+ million users
  - 2nd largest online game in China.



Two games: QQ Speed

- QQ Speed
  - Car racing against other users
  - 200+ million users



Datasets

- Statistics of the datasets

Category	Type	QQSpeed	DNF
User	all users	7.60M	347K
	free users	$\sim 10^6$	$\sim 10^5$
	paying users	$\sim 10^6$	$\sim 10^5$
	new payers	$\sim 10^5$	$\sim 10^4$
Relationship	co-playing	134M	7.30M
Guild	guilds	600K	49.6K
	co-guild	66.7M	51.7M
Activity	activity types	58	64
	activity logs	44.7B	5.71B
Date span	from	2013.6.20	2013.4.1
	to	2013.8.20	2013.6.30

Datasets

- Statistics of the datasets

Category	Type	QQSpeed	DNF
User	all users	7.60M	347K
	free users	$\sim 10^6$	$\sim 10^5$
	paying users	$\sim 10^6$	$\sim 10^5$
	new payers	$\sim 10^5$	$\sim 10^4$
Relationship	co-playing	134M	7.30M
Guild	guilds	600K	49.6K
	co-guild	66.7M	51.7M
Activity	activity types	58	64
	activity logs	44.7B	5.71B
Date span	from	2013.6.20	2013.4.1
	to	2013.8.20	2013.6.30

Datasets

- Statistics of the datasets

Category	Type	QQSpeed	DNF
User	all users	7.60M	347K
	free users	$\sim 10^6$	$\sim 10^5$
	paying users	$\sim 10^6$	$\sim 10^5$
	new payers	$\sim 10^5$	$\sim 10^4$

Date span of data



Free users: **No $**

Paying users: **$**

New Payers: **No $** **$**

Datasets

- Statistics of the datasets

Category	Type	QQSpeed	DNF
User	all users	7.60M	347K
	free users	$\sim 10^6$	$\sim 10^5$
	paying users	$\sim 10^6$	$\sim 10^5$
	new payers	$\sim 10^5$	$\sim 10^4$
Relationship	co-playing	134M	7.30M
Guild	guilds	600K	49.6K
	co-guild	66.7M	51.7M
Activity	activity types	58	64
	activity logs	44.7B	5.71B
Date span	from	2013.6.20	2013.4.1
	to	2013.8.20	2013.6.30

Online Test

- Test setting
  - Two groups: *test group* and *control group*.
  - Send messages to invite the user to attend a promotion activity.



Online Test Results

- Online test 1
  - Test the effectiveness of our approach in online scenario.
  - Test group: LCFM
  - Control group: Prior strategy

	Online Test 1 2013.12.27 - 2014.1.3		Online Test 2 2014.1.24 - 2014.1.27		
Group name	test group	control group	test group	control group	prior group
Group size	600K	200K	400K	400K	200K
#Message read	345K	106K	229K	215K	106K
Message read rate	57.50%	53.00%	57.25%	53.75%	53.00%
#Message clicked	47584	7466	23325	20922	6299
Message clicked rate	7.93%	3.73%	5.83%	5.23%	3.15%
Lift_Ratio	196.87%	0%	126.81%	73.40%	0%

Online Test Results

- Online test 2
  - Test the contribution of social factors in online scenario.
  - Test group: LCFM
  - Control group: LCFM – Social effect features

	Online Test 1 2013.12.27 - 2014.1.3		Online Test 2 2014.1.24 - 2014.1.27		
Group name	test group	control group	test group	control group	prior group
Group size	600K	200K	400K	400K	200K
#Message read	345K	106K	229K	215K	106K
Message read rate	57.50%	53.00%	57.25%	53.75%	53.00%
#Message clicked	47584	7466	23325	20922	6299
Message clicked rate	7.93%	3.73%	5.83%	5.23%	3.15%
Lift_Ratio	196.87%	0%	126.81%	73.40%	0%

AMiner.org

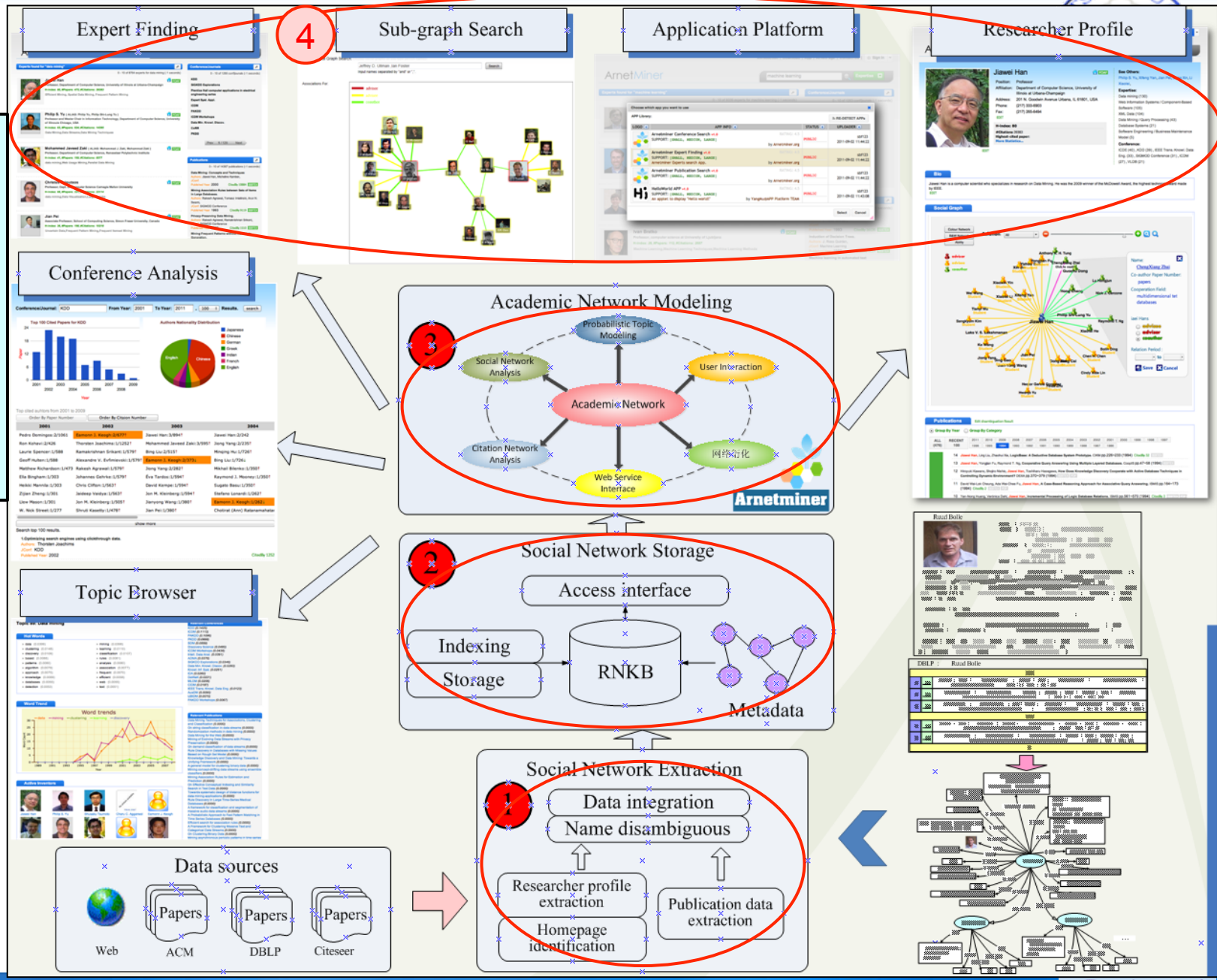
- Academic research social network analysis and mining system

提供全面的研究者网络分析与挖掘功能

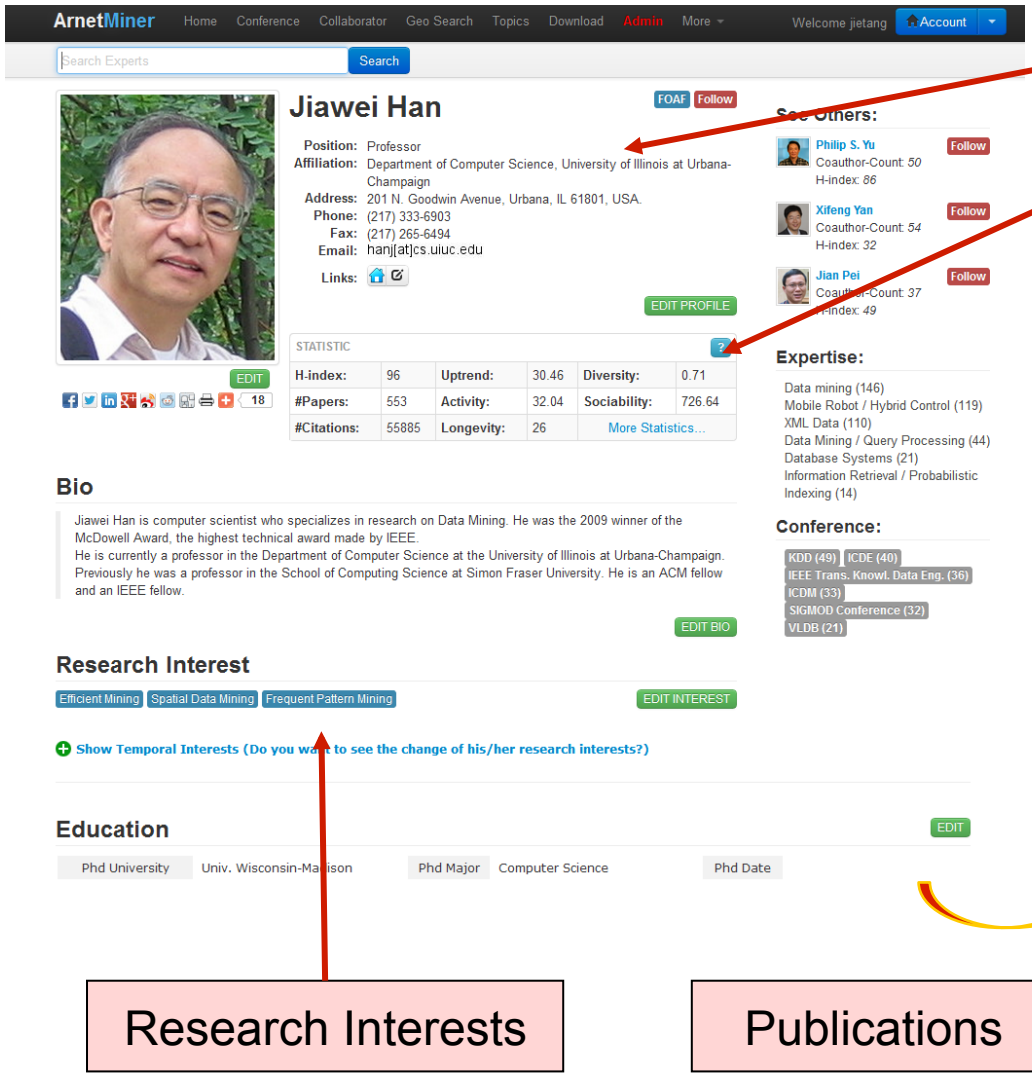
Papers published: ACM TKDD, IEEE TKDE, J. Informetrics, KDD'08-12, WWW'12, SIGMOD'09, IJCAI'09

http://aminer.org/

- Researcher profile extraction
- Expert finding
- Social network search
- Topic browser
- Conference analysis
- ArnetApp platform



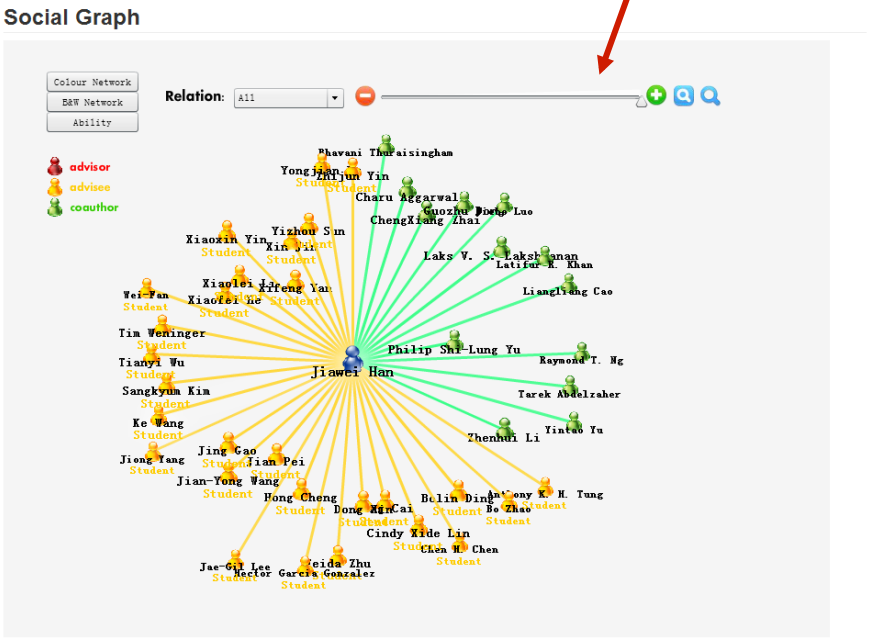
Person Search



Basic Info.

Citation statistics

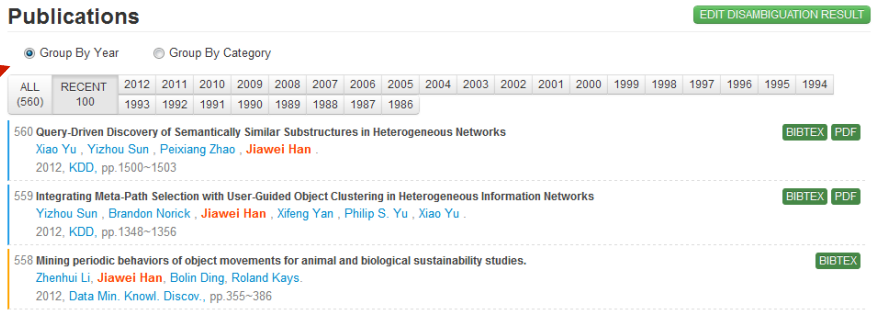
Social Network



Social Graph

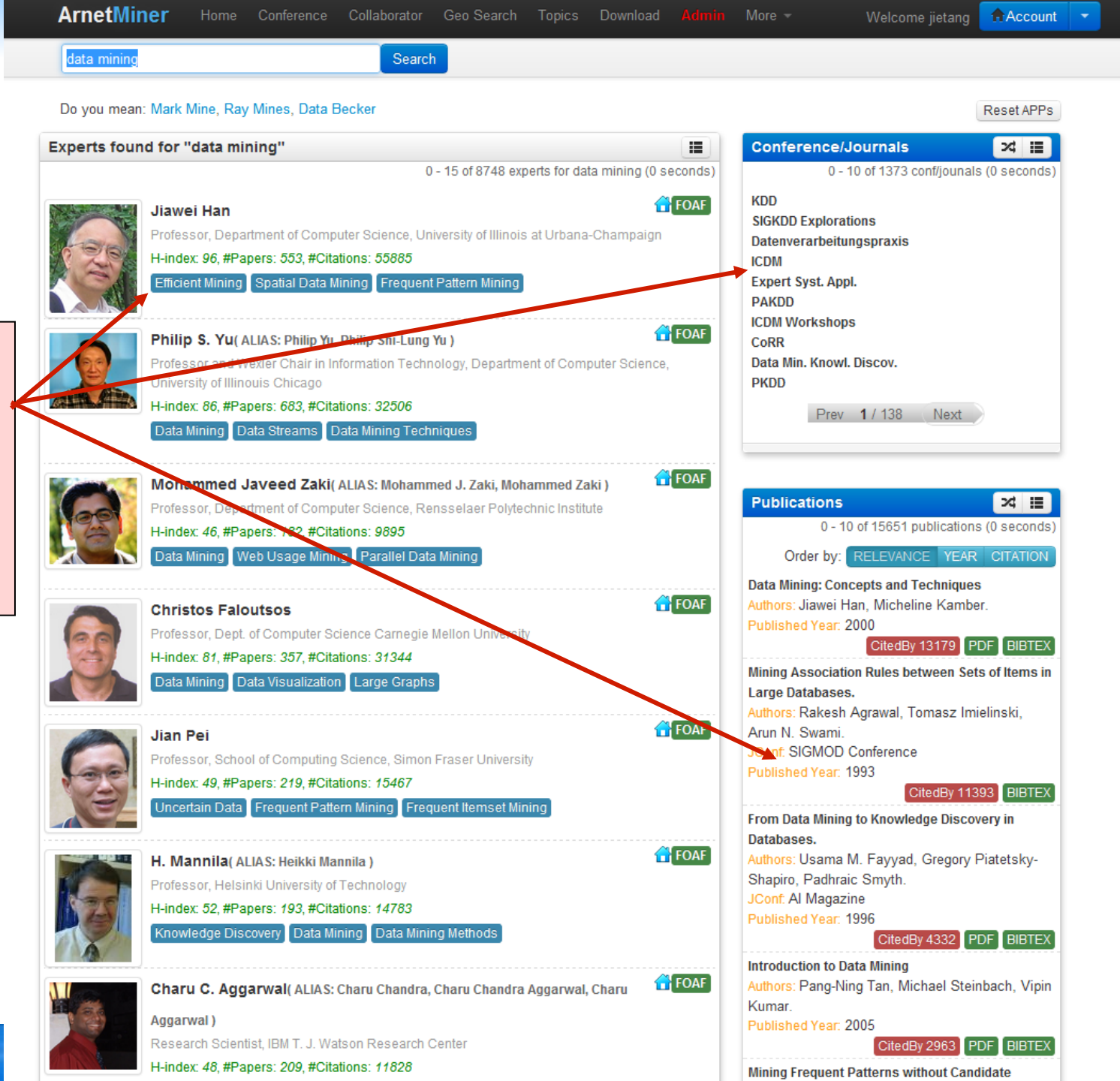
Research Interests

Publications



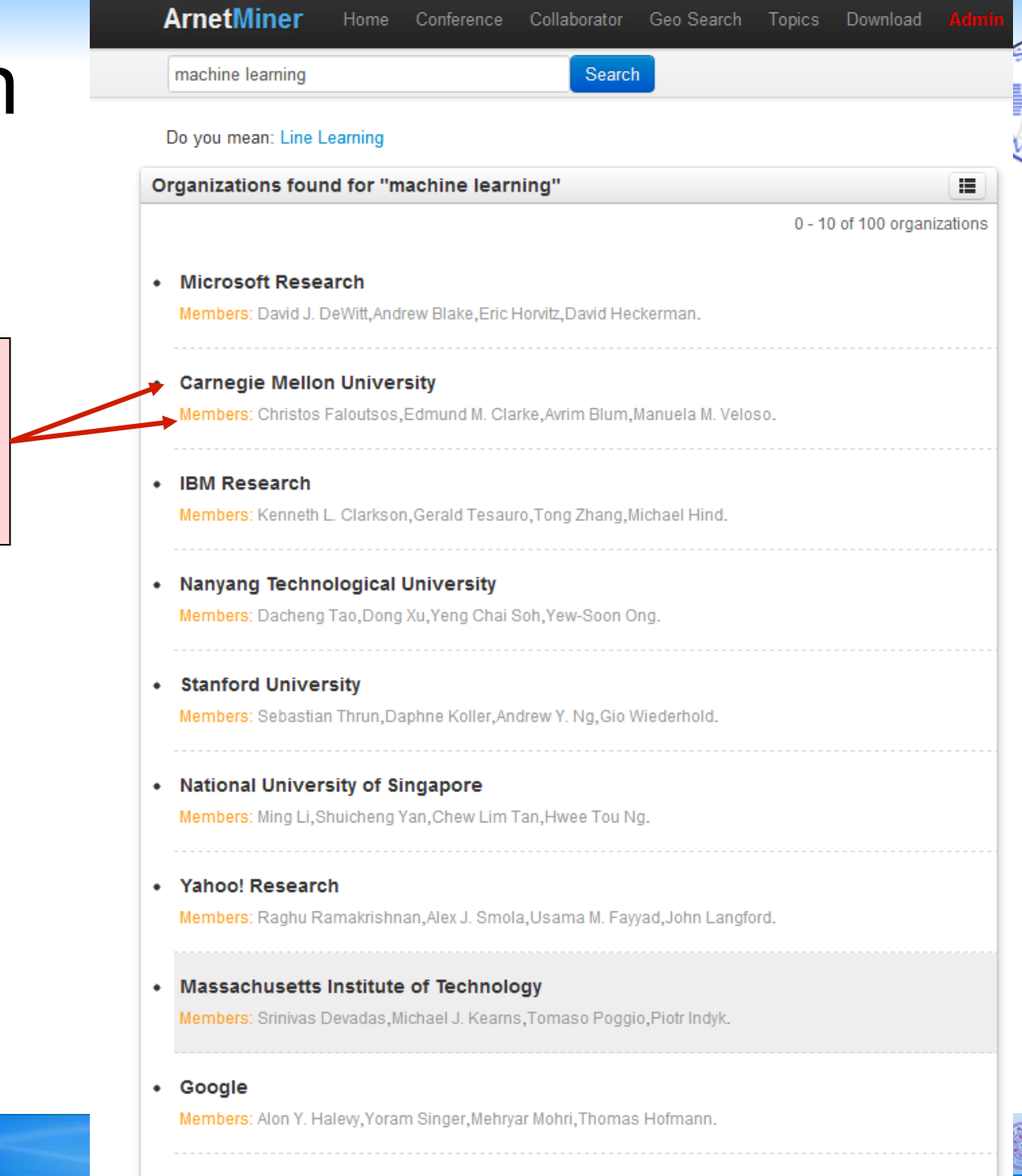
Expertise Search

Finding experts, expertise conferences, and expertise papers for “data mining”



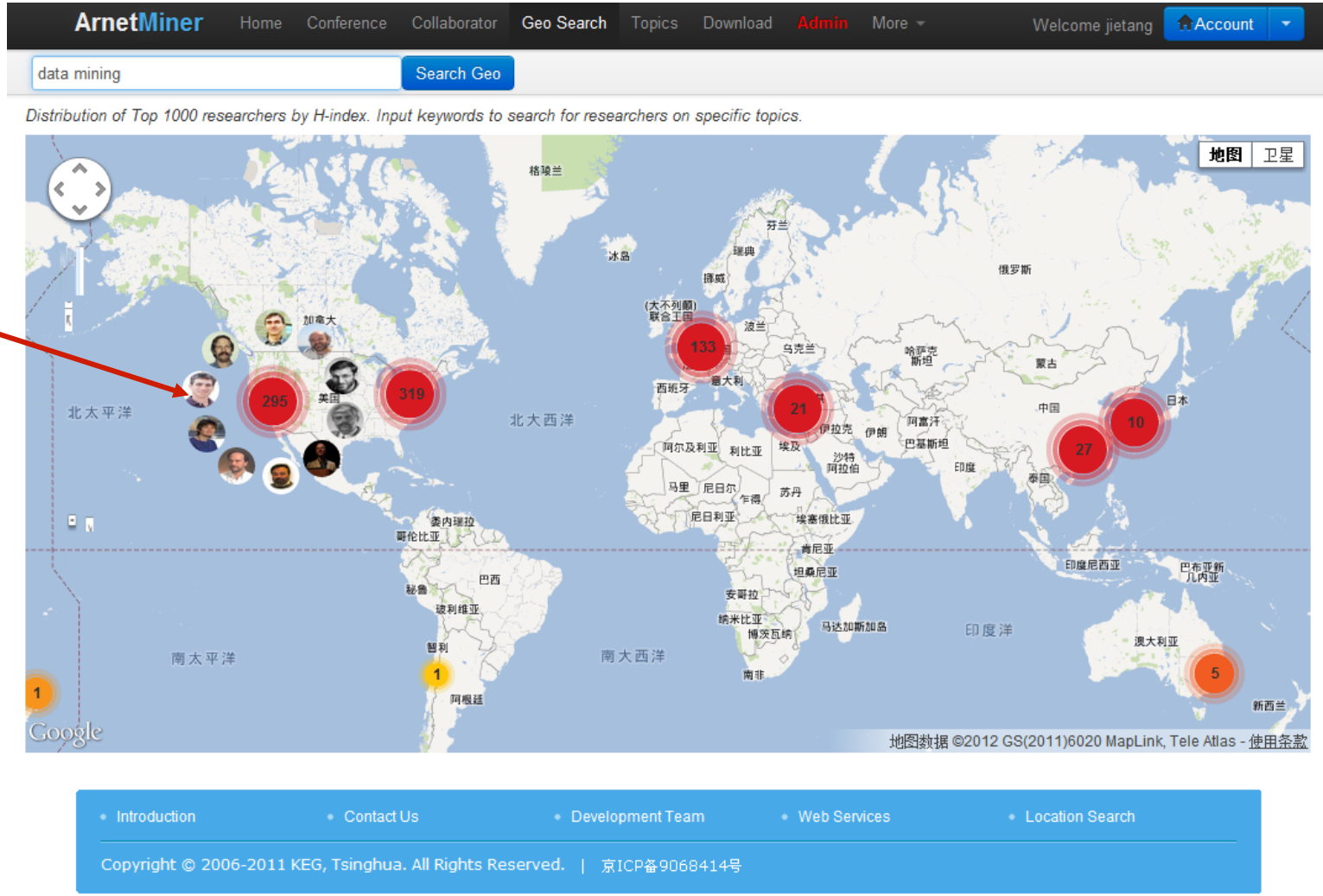
Organization Search

Ranking organizations on “machine learning”



Geographic search

Finding the most hot regions on "data mining"

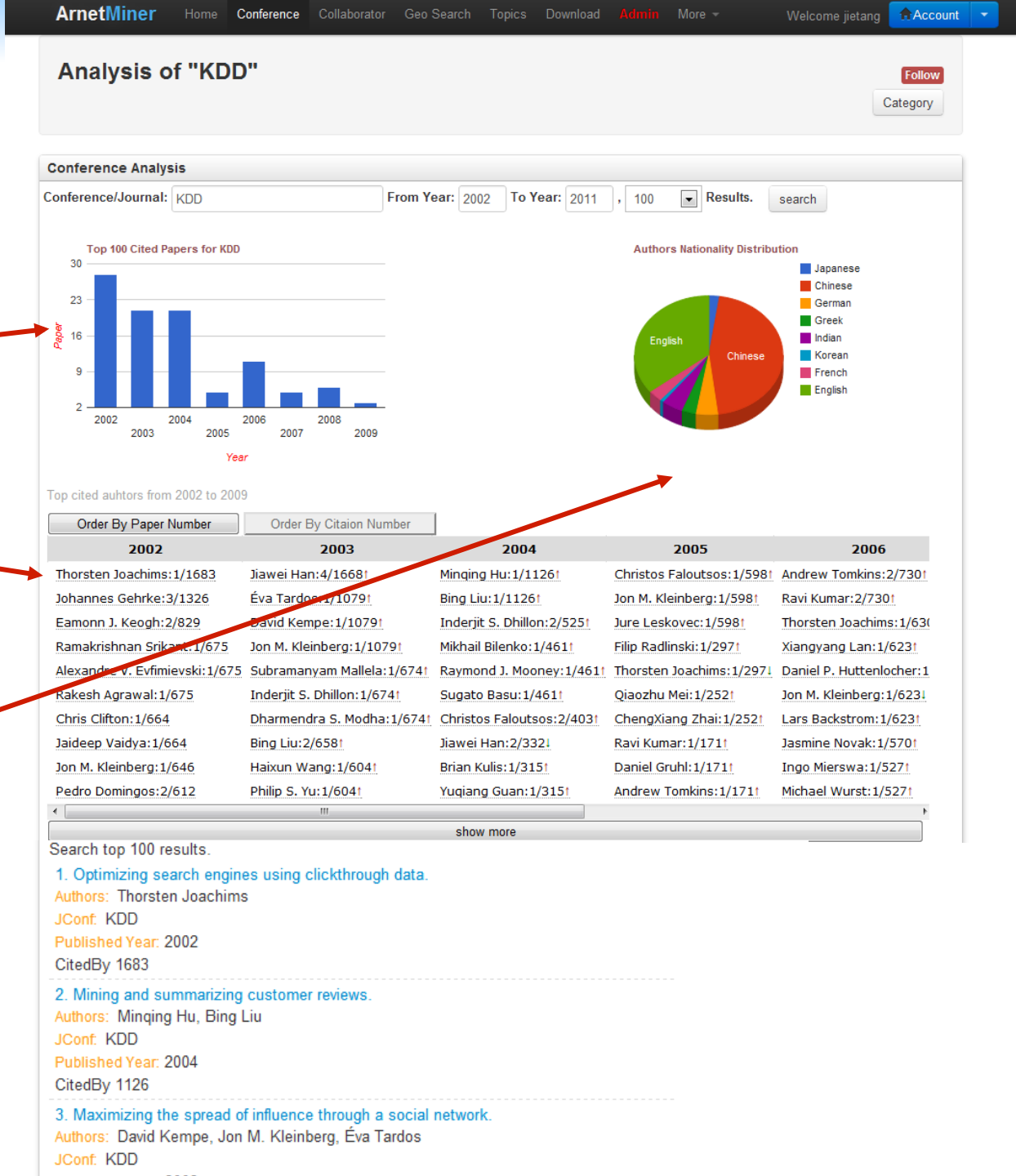


Conference Analysis

Which year is the most successful in the KDD's history?

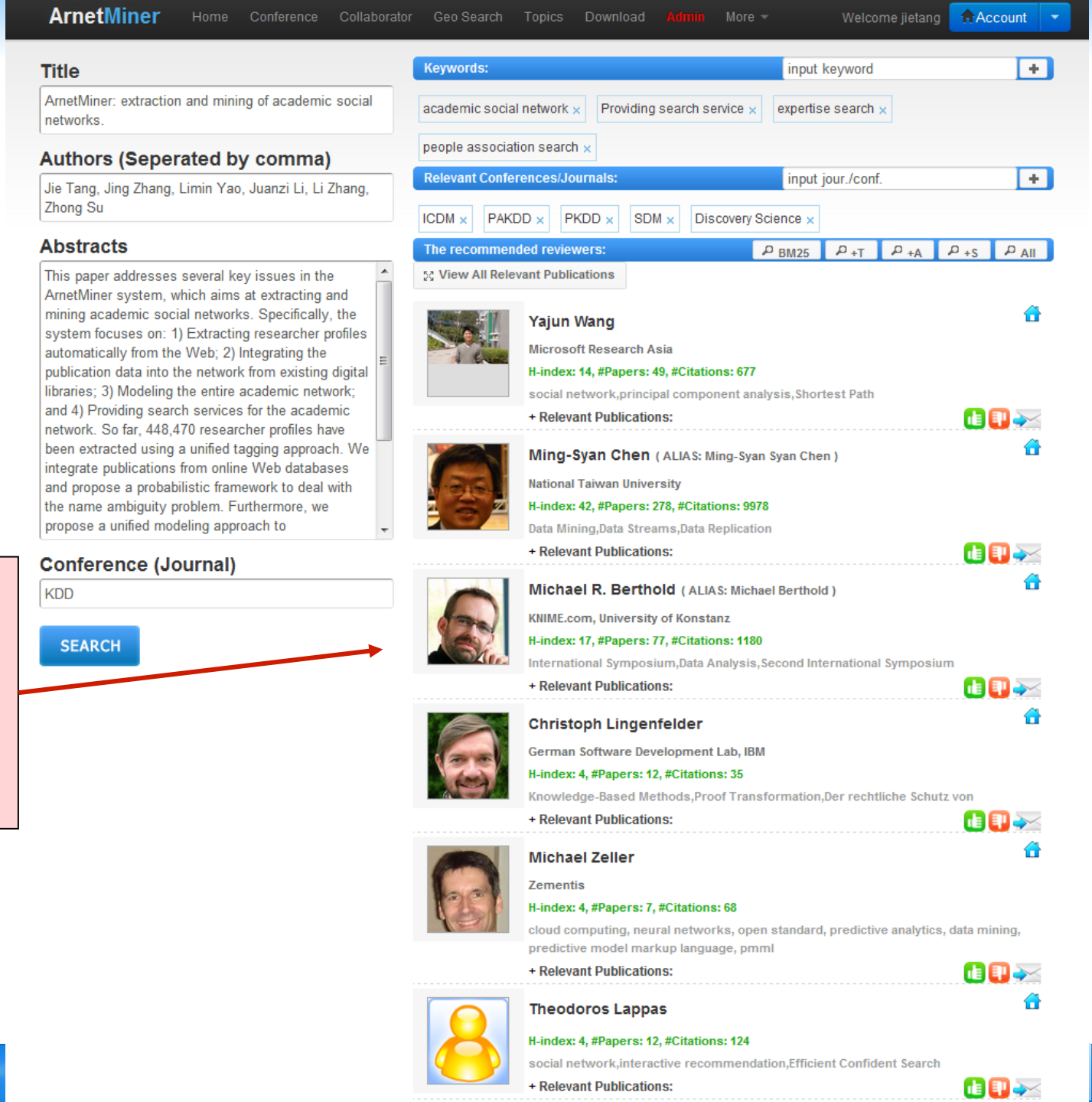
Who are the most cited authors?

What is author distribution for the highly cited KDD papers in the past years?



Reviewer Suggestion

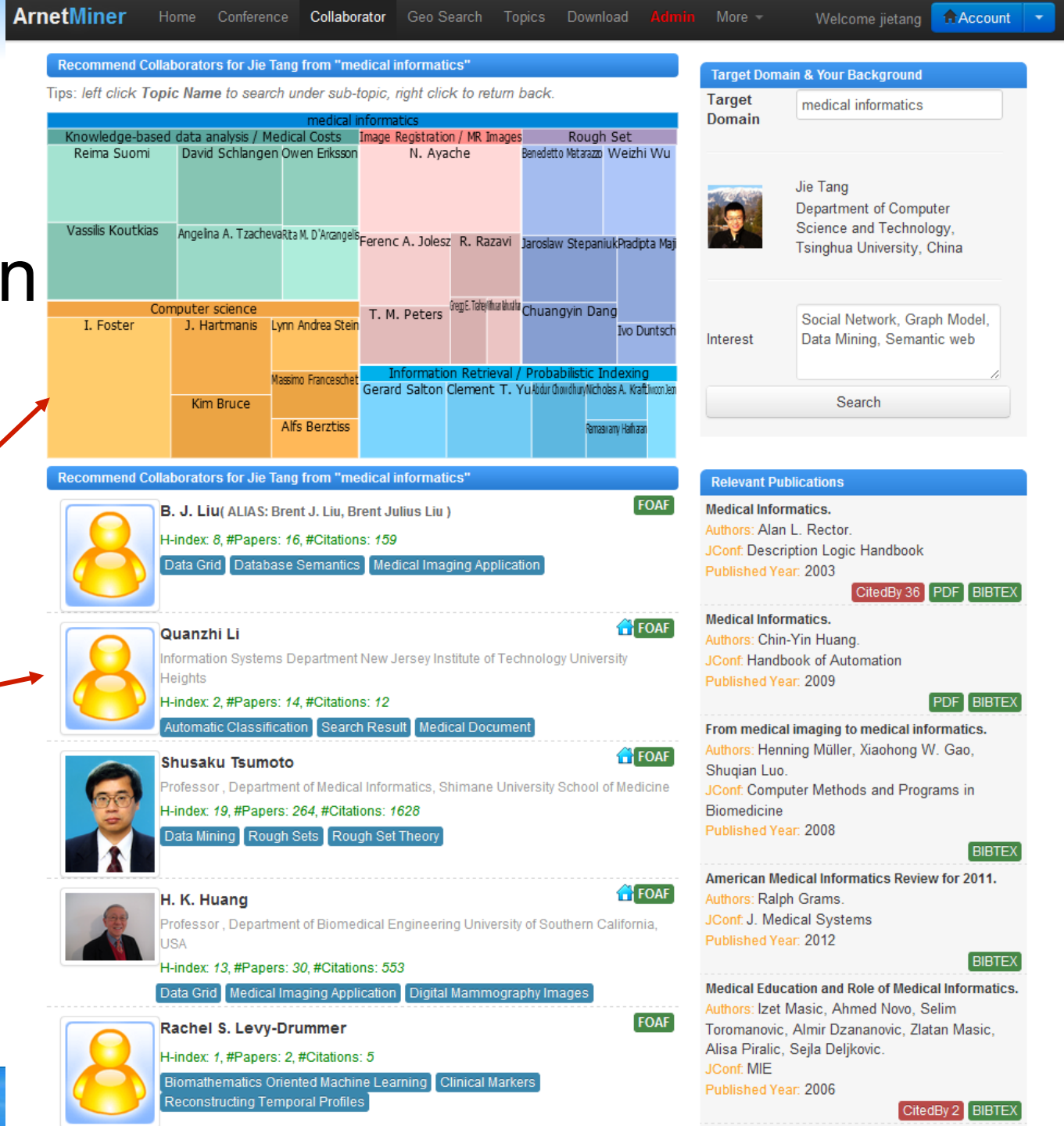
Inerest matching
COI avoiding
Load balancing
Forcast review quality



Cross-domain Collaboratinon Recommendation

What kinds of topics should you collaborate with the target domain?

Who are the best collaborators on each topic?



Topic Browser

Topic 83 — Semantic Web / Ontology Alignment



International Semantic Web Conference
EKAW
ESWC
Wissensmanagement

Topic 84 — Parallel Applications / Average-Case Performance Analysis



IPDPS
Euro-Par
CLUSTER
HiPC

Topic 85 — Service management



ISCC
NOMS
J. Network Syst. Manage.
Integrated Network Management

Topic 86 — Web Information Systems / Component-Based Software



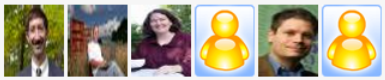
EUROMICRO
ICWE
Computers and Artificial Intelligence
DS-RT

Topic 87 — Power optimization / Synchronous Sequential Circuits



DAC
IEEE Trans. on CAD of Integrated Circuits and Systems
ICCAD
ASP-DAC

Topic 88 — Digital library / Information Access



Environmental Modelling and Software
JCDL
ECDL
ICADL

Topic 89 — Data mining



KDD
ICDM
PAKDD

200 topics have been discovered automatically from the academic network

Topic 89: Data mining

Hot Words

- data (0.0399)
- clustering (0.0148)
- discovery (0.0109)
- based (0.0088)
- patterns (0.0080)
- algorithm (0.0079)
- approach (0.0075)
- knowledge (0.0069)
- databases (0.0055)
- detection (0.0053)
- mining (0.0398)
- learning (0.0110)
- classification (0.0107)
- rules (0.0081)
- analysis (0.0080)
- association (0.0077)
- frequent (0.0070)
- efficient (0.0058)
- web (0.0055)
- text (0.0051)

Word Trend



Active Inventors

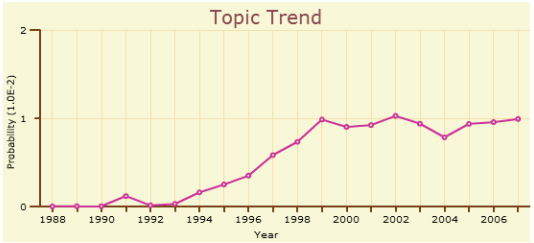


Jiawei Han, Philip S. Yu, Shusaku Tsumoto, Charu C. Aggarwal, Eamonn J. Keogh, Heikki Mannila, Christos Faloutsos, Mohammed Javeed Zaki, Einoshin Suzuki, Tao Li

Popular Phrases

- data mining
- binary data
- data mining application
- time series data
- Clustering Binary Data
- data stream
- concept-drifting data stream
- massive audio data stream
- Categorical Data Streams
- Data Mining Techniques

Topic Trend



Relevant Conferences

- KDD (0.1425)
- ICDM (0.1113)
- PAKDD (0.1096)
- PKDD (0.0668)
- SDM (0.0559)
- Discovery Science (0.0480)
- ICDM Workshops (0.0436)
- Intell. Data Anal. (0.0391)
- ADMA (0.0376)
- SIGKDD Explorations (0.0346)
- Data Min. Knowl. Discov. (0.0283)
- Knowl. Inf. Syst. (0.0261)
- IDA (0.0260)
- DaWaK (0.0221)
- MLDM (0.0206)
- CIDM (0.0187)
- IEEE Trans. Knowl. Data Eng. (0.0123)
- AusDM (0.0090)
- IJBIDM (0.0070)
- PAKDD Workshops (0.0067)

Relevant Publications

- Data Mining Techniques for Associations, Clustering and Classification
- On string classification in data streams
- Randomization methods in data mining
- Data Mining for the Web
- Mining of Evolving Data Streams with Privacy Preservation
- On demand classification of data streams
- Rule Discovery in Databases with Missing Values Based on Rough Set Model
- Knowledge Discovery and Data Mining: Towards a Unifying Framework
- A general model for clustering binary data
- Mining concept-drifting data streams using ensemble classifiers
- Mining Association Rules for Estimation and Prediction
- On Effective Conceptual Indexing and Similarity Search in Text Data
- Towards systematic design of distance functions

清華大學 Tsinghua University

Academic Performance Measure

Academic Statistics

Home | H-index | Citation | Uptrend | Activity | Longevity | Diversity | Sociability | New Star

Top 3 by H-index:

1 Anil K. Jain (Distinguished Professor, Michigan State University)



H-index: 112; Papers: 368; Citation: 66111
Homepage: http://www.cse.msu.edu/~jain/

Expertise: Object Recognition / Two-View Motion Estimation Face recognition / Image analysis Image segmentation / Face Recognition

2 H. Garcia (Professor, Departments of Computer Science and Electrical Engineering, Stanford University)



H-index: 107; Papers: 412; Citation: 45242
Homepage: http://www-db.stanford.edu/people/hector.html

Expertise: Real-Time Systems / Automated Software Test Data XML Data Database Systems Dynamic Networks / Extended Abstract

3 C. Papadimitriou (Professor, Computer Science Division University of California at Berkeley)



H-index: 100; Papers: 356; Citation: 54068
Homepage: http://www.cs.berkeley.edu/~christos/

Expertise: Mechanism design / Learning Stochastic Finite Automata Communication Complexity / Lower Bounds Approximation Algorithms / Perfect Graphs Finite Languages / Database Queries

Top 3 by Citation:

1 L. Zadeh (Professor Emeritus, Graduate School, Computer Science Division Department of Electrical Engineering and Computer Sciences University of California)



H-index: 34; Papers: 108; Citation: 68634
Homepage: http://www.cs.berkeley.edu/~zadeh/

Expertise: Fuzzy system Rough Set Intelligent Information Systems / Neural Network Learning

2 Anil K. Jain (Distinguished Professor, Michigan State University)



H-index: 112; Papers: 368; Citation: 66111
Homepage: http://www.cse.msu.edu/~jain/

Expertise: Object Recognition / Two-View Motion Estimation Face recognition / Image analysis Image segmentation / Face Recognition

3 Hari Balakrishnan (Professor, Department of EECS Massachusetts Institute of Technology)



H-index: 80; Papers: 142; Citation: 60955
Homepage: http://nms.csail.mit.edu/~hari/

Expertise: Wireless network / End to end Routing Behavior Sensor Networks / Outdoor wireless Wireless Networks / Priority Scheduling File System /

New Stars

Home | H-index | Citation | Uptrend | Activity | Longevity | Diversity | Sociability | New Star

Data mining (Topic:89) Show all topics

1 2 3 4 5 6 7 8 9 10 11 ...

1 Dijun Luo (PhD student, Computer Science and Engineering Department of University of Texas at Arlington)

New Star: 6.42; H-index: 4; Papers: 24; Citation: 65
Homepage: http://sites.google.com/site/dijunluo/

Expertise: Data mining Object Recognition / Two-View Motion Estimation Neural Networks

2 Weizhu Chen (RESEARCHER , Machine Learning group, Microsoft Research Asia)



New Star: 3.92; H-index: 9; Papers: 22; Citation: 141
Homepage: http://research.microsoft.com/en-us/people/wzchen/

Expertise: Data mining Learning Search Control Rules / Explanation-based Approach

3 Yan Liu (The Hong Kong Polytechnic University, Hong Kong, China)

New Star: 3.72; H-index: 4; Papers: 23; Citation: 45

Expertise: Data mining Software Engineering / Design Patterns Frameworks

4 Quanquan Gu (Ph.D. Candidate)



New Star: 3.61; H-index: 5; Papers: 26; Citation: 83
Homepage: http://www.cs.illinois.edu/homes/qgu3/

Expertise: Data mining

5 Nathan Nan Liu (PHD student , Hong Kong University of Science and Technology)



New Star: 3.59; H-index: 7; Papers: 15; Citation: 209
Homepage: http://www.cse.ust.hk/~nliu/

Expertise: Data mining Adaptive Mobile Applications / Pervasive Computing Systems

6 Ding Yuan

New Star: 3.54; H-index: 6; Papers: 8; Citation: 128
Homepage: http://www.cs.uiuc.edu/homes/dyuan3/

Expertise: Data mining Mobile Robot / Hybrid Control Grid Computing / Parallel System Architecture

7 Barna Saha

New Star: 3.54; H-index: 11; Papers: 22; Citation: 400

Expertise: Approximation Algorithms / Perfect Graphs Data mining

8 Chi Wang (PhD Candidate, Department of Computer Science and Technology, University of Illinois at Urbana-Champaign)



New Star: 3.54; H-index: 5; Papers: 11; Citation: 257
Homepage: http://www.cs.illinois.edu/homes/chiwang1/

AMiner (ArnetMiner)

- Academic Social Network Analysis and Mining system—Aminer (http://arnetminer.org)
  - Online since 2006
  - >1 million researcher profiles
  - >131 million requests
  - >2.35 Terabyte data
  - 100K IP access from 170 countries
  - 10% increase of visits per month
- Deep SN analysis, mining, and search



User Distribution

4.04 million IP from 220 countries/regions



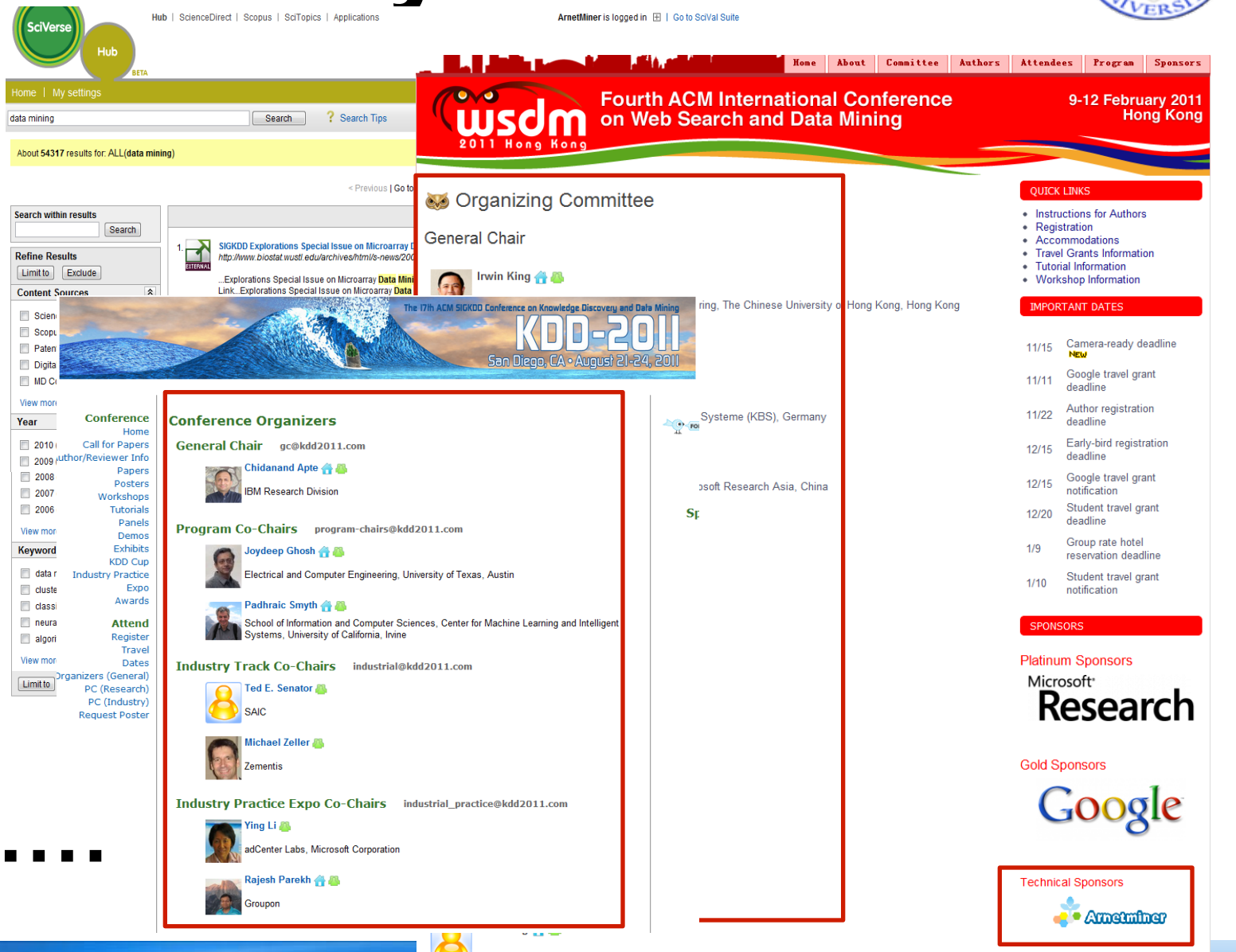
Top 10 countries

1. USA
2. China
3. Germany
4. India
5. UK
6. Canada
7. Japan
8. Spain
9. France
10. Italy

Widely used..

- **The largest publisher:**
  Elsevier
- **Conferences**
  KDD 2010
  KDD 2011
  KDD 2012
  WSDM 2011
  ICDM 2011
  ICDM 2012
  SocInfo 2011
  ICMLA 2011
  WAIM 2011
  etc.

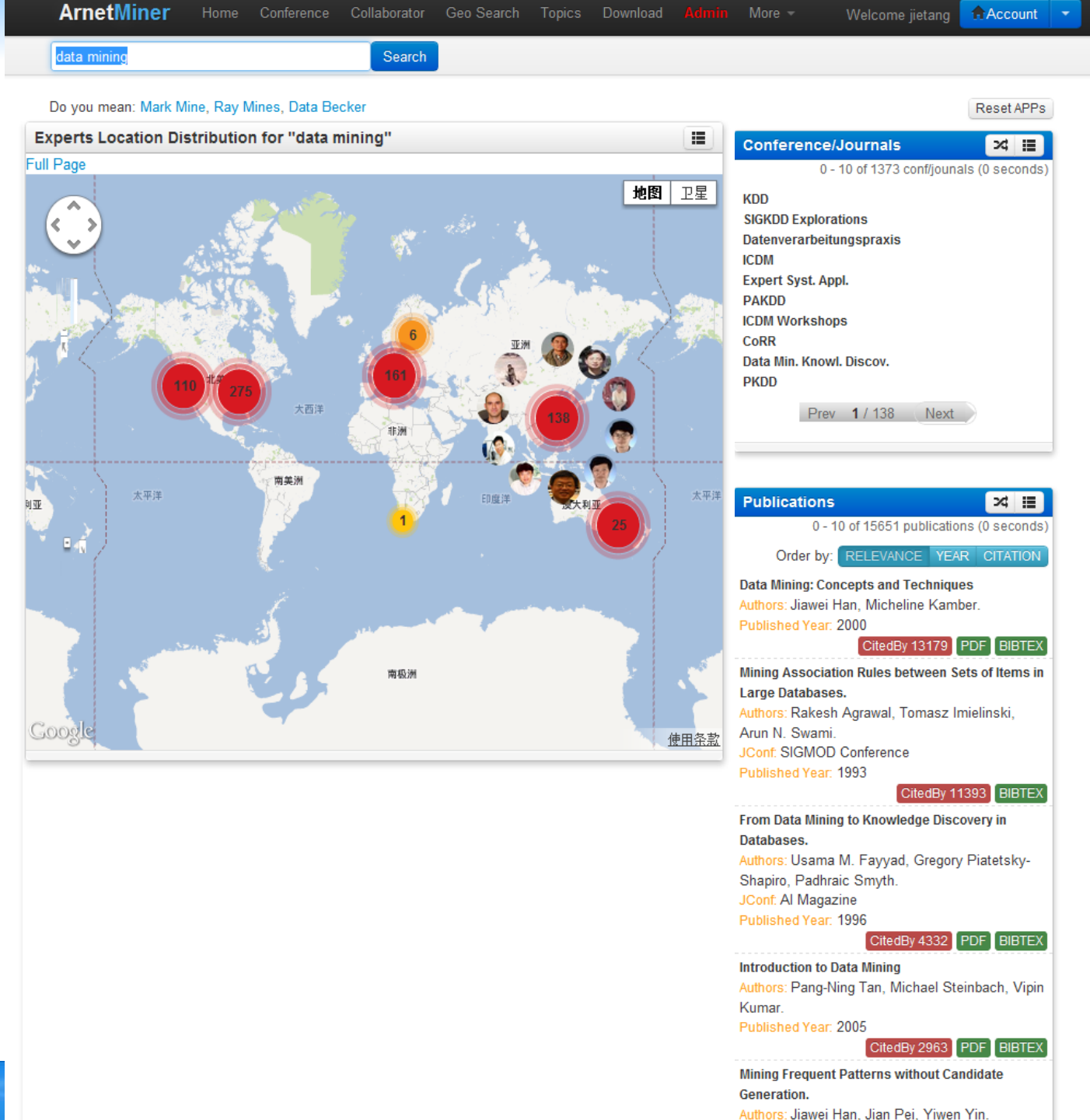
. . . . . .



ArnetApp Platform

---to deploy your apps on Arnetminer.org

To customize the search on Arnetminer.



Data Representation

Data Structures

- An example with *p* features

$$\boldsymbol{x}_i=[x_{i1}, x_{i2},\ldots, x_{ip}]$$

- Data (set) matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nf} & \cdots & x_{Np} \end{bmatrix}$$

Type of data

- Interval-scaled variables
- Binary variables
- Nominal, ordinal, and ratio variables
- Variables of mixed types

- To make things easy, let us say $x_i \in X = R^n$

Similarity and Dissimilarity between Data Instances

- Distances are normally used to measure the similarity or dissimilarity between two data instances
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

Similarity and Dissimilarity Between data Instances (Cont.)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

  - Properties
    - *d(i,j)* != 0
    - *d(i,i)* = 0
    - *d(i,j)* = *d(j,i)*
    - *d(i,j)* < *d(i,k)* + *d(k,j)*
- Or more general Mahalanobis distance $(x_i - x_j)^T A (x_i - x_j)$, where *A* is a covariance matrix.
- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

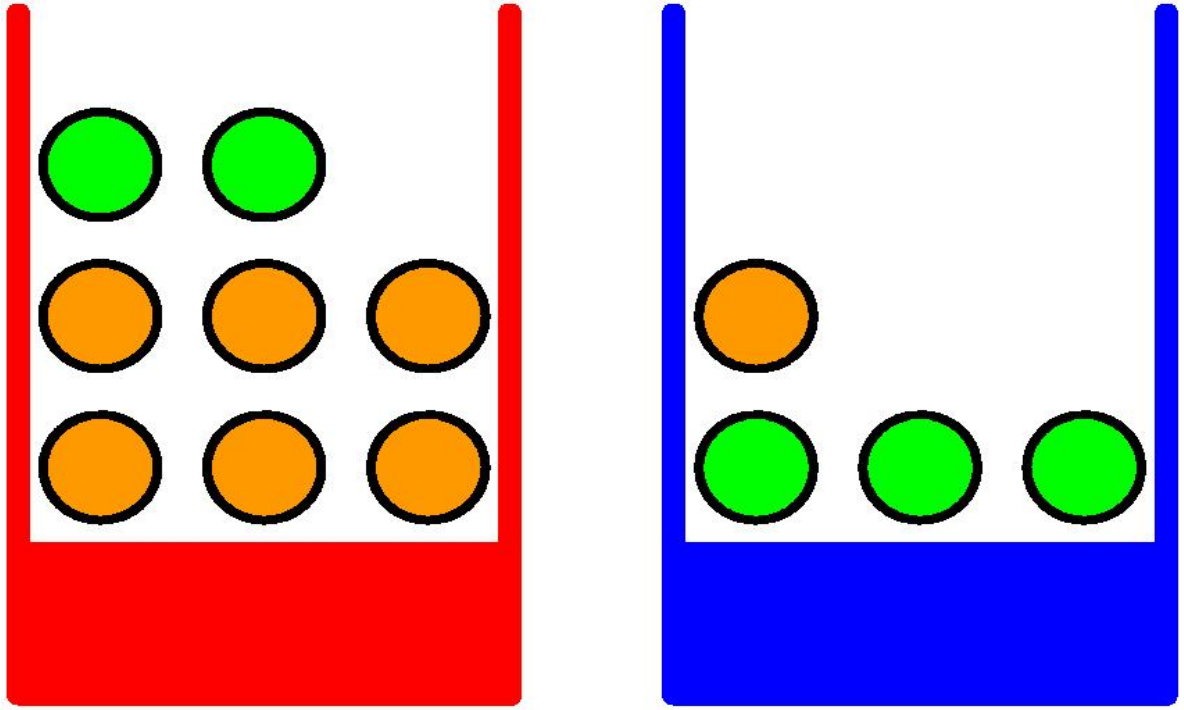
Cosine Similarity

- Given two vectors (data instances), their similarity can be calculated by cosine similarity

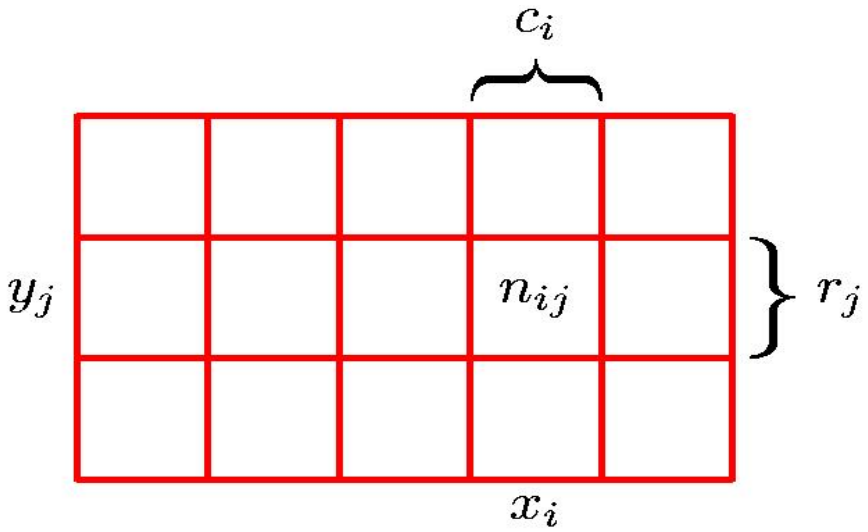
$$sim(x_i, x_j) = \frac{\sum_{k=1,\ldots,p} x_{ik} \times x_{jk}}{\sqrt{\sum_{k=1,\ldots,p} x_{ik}^2} \times \sqrt{\sum_{k=1,\ldots,p} x_{jk}^2}}$$

Probability Basis

Apples and Oranges



Probability Basis



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

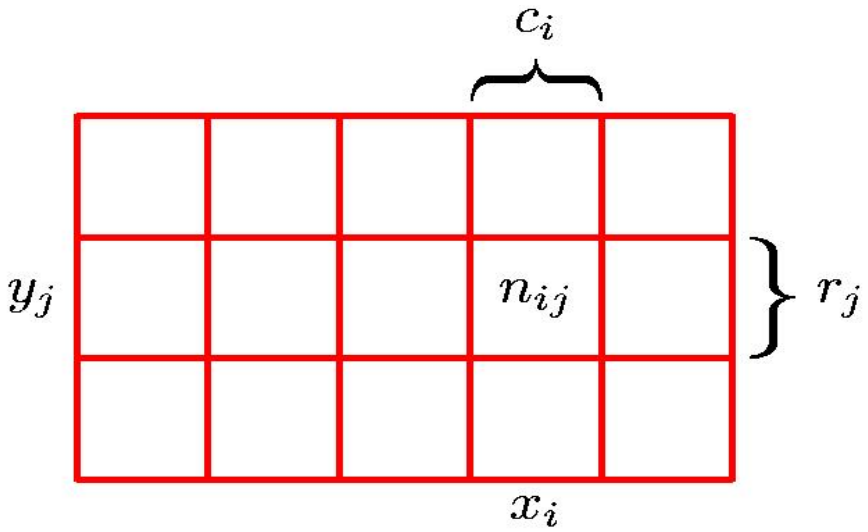
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Basis



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i)p(X = x_i)$$

The Rules of Probability

Sum Rule $$p(X) = \sum_{Y} p(X, Y)$$

Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

Posterior ≈ likelihood × prior

The State of ML

Machine Learning History



pre-computer age
thoughts on symbolic reasoning

1950s & 1960s: symbolic reasoning, representation, search

1960s & 1970s: success on NLP systems

1970s & 1980s: expert systems

since 1990s
Great growth in machine learning

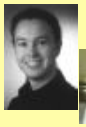
The State of Machine Learning

Classification models

Decision tree 
Bayesian classifier
Perceptron
Neural networks

Maximal Margin

SVM (Vapnik)



Maximal margin sequential learning

M3N network (Ben Taskar)


Deep learning

Deep belief networks




Topic models

PLSI, LDA, etc.





Sequential learning

HMM

Sequential learning

MEMM, CRF, voted perceptron




Graphical models

Factor graph, Exponential model






For more: http://arnetminer.org/expertisesearch.do?keyword=Machine%20Learning

Thanks!

Jie Tang, DCST
http://keg.cs.tsinghua.edu.cn/jietang/
http://arnetminer.org
Email: jietang@tsinghua.edu.cn

What is Machine Learning?

Classification Problem

- Given a training set $S=\{(x_1, y_1),(x_2, y_2),\ldots,(x_m, y_m)\}$, and $x_i \in X=R^n$, $y_i \in Y=\{1,-1\}$, $i=1,2,\ldots,m$
- To learn a function $g(x)$, which can best fit the data

$$y = f(x)$$

- and make the decision function $f(x)=\mathrm{sgn}(g(x))$ can classify new input x, or the value of y.

$$e.g.,\ f(x) = \begin{cases} 1, & g(x) > 0 \\ -1, & g(x) < 0 \end{cases}$$

A linear model

- A linear classifier

$$y_i = \text{sign}(\mathbf{w} \cdot x_i) = \text{sign}(\sum_{k=1}^{p} w_k \cdot x_{ik})$$

where **w** is a vector of parameters we need to learn.

Optimization Framework

- Given a training set $S=\{(x_1, y_1),(x_2, y_2),\ldots,(x_m, y_m)\}$, and $x_i \in X=R^n$, $y_i \in Y=\{1,-1\}$, $i=1,2,\ldots,m$
- Let us learn the parameters **w** by fitting a optimization framework

$$\min_{\mathrm{w}} \sum_{i=1}^{m} \left\| \mathrm{w}^{\mathrm{T}} x_i - y_i \right\|^2$$

- Or if we consider a maximum entropy model, we have

$$\max_{\mathrm{w}} \sum_{i=1}^{m} \log p(y_i \mid x_i, \mathrm{w}) - \lambda \left\| \mathrm{w} \right\|^2$$