



AMiner.org

—Deep Analysis and Mining for Academic Social Networks
(<http://aminer.org>)

Jie Tang

Knowledge Engineering Group
Department of Computer Science and Technology
Tsinghua University



The era of big and complex data...

- Web trend
 - Data centric → User centric
 - Offline sparse social → Online dense social
 - Large-scale data mining → Big data deep analytic
- Tech trend
 - Formal text → Informal text
 - User modeling → Collective intelligence
 - Keyword-based modeling → Semantic-based modeling
 - Macro-level analysis → Micro-level analysis
 - ...



Why AMiner.org (Arnetminer)?

The image shows a collage of academic search interfaces from various platforms. Red arrows point from the text "Academic search is treated as document search, but ignore semantics" to the Google Scholar search bar and the CiteSeer search results page. Red arrows also point from the text "The information need is not only about publication..." to the CiteSeer search results page and the CiteULike interface.

Google Scholar BETA

data mining

Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar All articles - Recent articles

CiteSeer Find: data mining

Documents Citations

Searching for PHRASE data mining

Restrict to: Header Title Order by

8612 documents found. Only retrieve

Wrappers for Feature Subset Selection

citeulike

SAGE JOURNALS Online

Free Trial to 485+ SAGE journals until May 31, 2008

Click here for more information...

Sponsored link

ANNUAL REVIEWS

INSIGHTFUL RESEARCH STARTS HERE

Sponsored link

A free online journal library

Some recent publications

- Highly Cited Cell
- Clinical Trials by Silvia Hayman, Widespread poster
- in Cervical
- famous
- 072. Elisabeth Müller-Holzner, Frederic R. Santer, Andreas J. Jansen-Pohl, Werner Zwierschke
- algorithm algorithms
- alignment analysis and architecture attention
- bacteria bayesian
- bioinformatics biology
- brain cancer cell
- classification clustering
- collaboration
- communication commun
- complexity control culture
- data database design

"Academic search is treated as document search, but ignore semantics"

"The information need is not only about publication..."

Examples – Expertise search

Researcher A



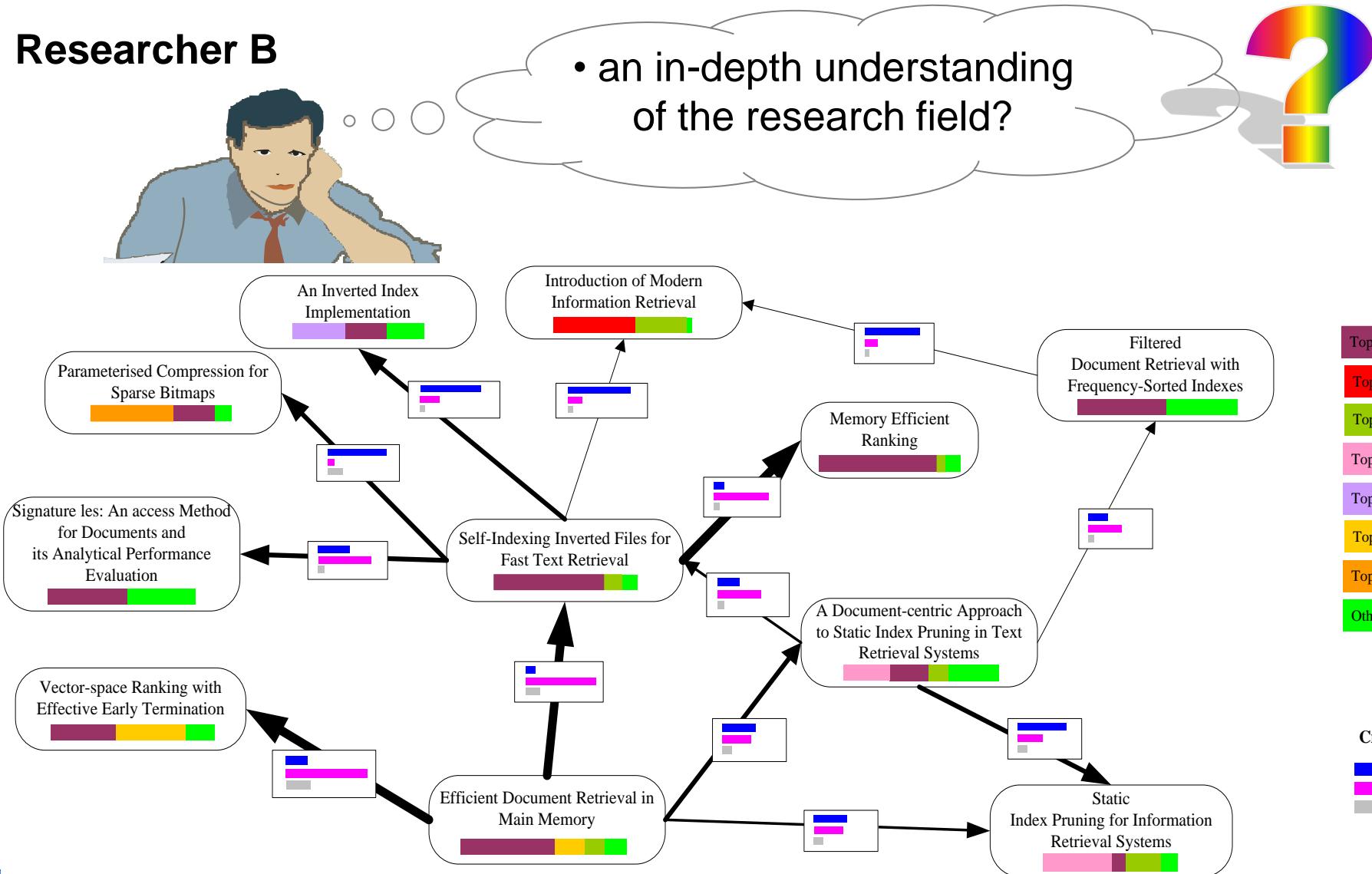
- When starting a work in a new research topic;
- Or brainstorming for novel ideas.

- Who are experts in this field?
- What are the top conferences in the field?
- What are the best papers?
- What are the top research labs?



Examples – Citation network analysis

Researcher B



Examples – Conference Suggestion

Which conference
should we submit the
paper?



Researcher C



Latent Dirichlet Co-Clustering

M. Mahdi Shafiei and Evangelos E. Milios
Faculty of Computer Science, Dalhousie University
6050 University Ave., Halifax, Canada
shafiei@cs.dal.ca , eem@cs.dal.ca

authors

Abstract

We present a generative model for simultaneously clustering documents and terms. Our model is a four-level hierarchical Bayesian model, in which each document is modeled as a random mixture of document topics, where each topic is a distribution over some segments of the text. Each of these segments in the document can be modeled as a mixture of word topics where each topic is a distribution over words. We present efficient approximate inference techniques based on Markov Chain Monte Carlo method and a Moment-Matching algorithm for empirical Bayes parameter estimation. We report results in document modeling, document and term clustering, comparing to other topic models, Clustering and Co-Clustering algorithms including Latent Dirichlet Allocation (LDA), Model-based Overlapping Clustering (MOC), Model-based Overlapping Co-Clustering (MOCC) and Information-Theoretic Co-Clustering (ITCC).

1 Introduction

Finding the appropriate representation model for text data has been one of the main issues for the data mining community since it started to look at the problem of processing text automatically. The “bag-of-words” representation is the basic and most widely used representation method for textual data [19]. In this approach, the order of words at which they appear in documents are ignored and only the word frequencies are taken into account. But this approach has been criticized for several reasons. Among those, it provides a relatively high dimensional representation of data (equal to the dictionary size) which causes curse of dimensionality problem [19]. Furthermore, it does not consider synonymy and polysemy relations of words in natural language. It has been also criticized of losing information due to its ignorance of word order. Various preprocessing steps such as removing stop-words and stemming have been used to reduce dimensionality, create and select

better features.

To overcome the high dimensionality issue of the bag-of-words representation, several dimension reduction methods have been proposed. Feature selection methods select a subset of words to reduce the dimensionality. Feature transformation methods try to tackle not only the high dimensionality problem of “bag-of-words” representation, but indirectly consider synonymy and polysemy as well. Latent Semantic Indexing (LSI) [6] is one of these approaches which uses singular value decomposition to identify a linear subspace in the original space of features. It is believed that the resulting new features also capture the two mentioned properties of natural language - polysemy and synonymy.

But the problem with most cartesian space representation approaches for text like LSI is their inability to provide interpretable components. Despite some work on interpreting the dimensions generated by these methods [5], these approaches are still far from providing a natural interpretation in the case of text. Topic models, on the other hand, are a class of statistical models in which the semantic properties of words and documents are expressed in terms of probabilistic topics. Probabilistic topic modeling as a way of representing the content of words and documents has the distinct advantage that each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms. The major difference between cartesian space methods like LSI and statistical topic models is that LSI family methods claim that words and documents can be represented as points in the Euclidean space whereas for the topic models, this is not the case.

One common assumption among most statistical models for language is still the *bag-of-words* assumption. In these models, no assumption is made about the order of words. In other words, while this family of methods tries to deal with the two first issues of bag-of-words representation, high dimensionality and ignoring polysemy and synonymy properties, it still keeps the “bag-of-words” assumption intact. Recently, there has been increased research interest in models sensitive to this kind of information [11].

content

Examples – Reviewer Suggestion

KDD Committee
Conference/Journal



Paper content

Who are **best matching** reviewers for each paper?

Who will **accept** the invitation to review the paper





Outline

- ArnetMiner: Academic Social Network
- Core Techniques
 - Knowledge Acquisition
 - Modeling and Heterogeneous Ranking
 - Social Network Analysis



AMiner.org

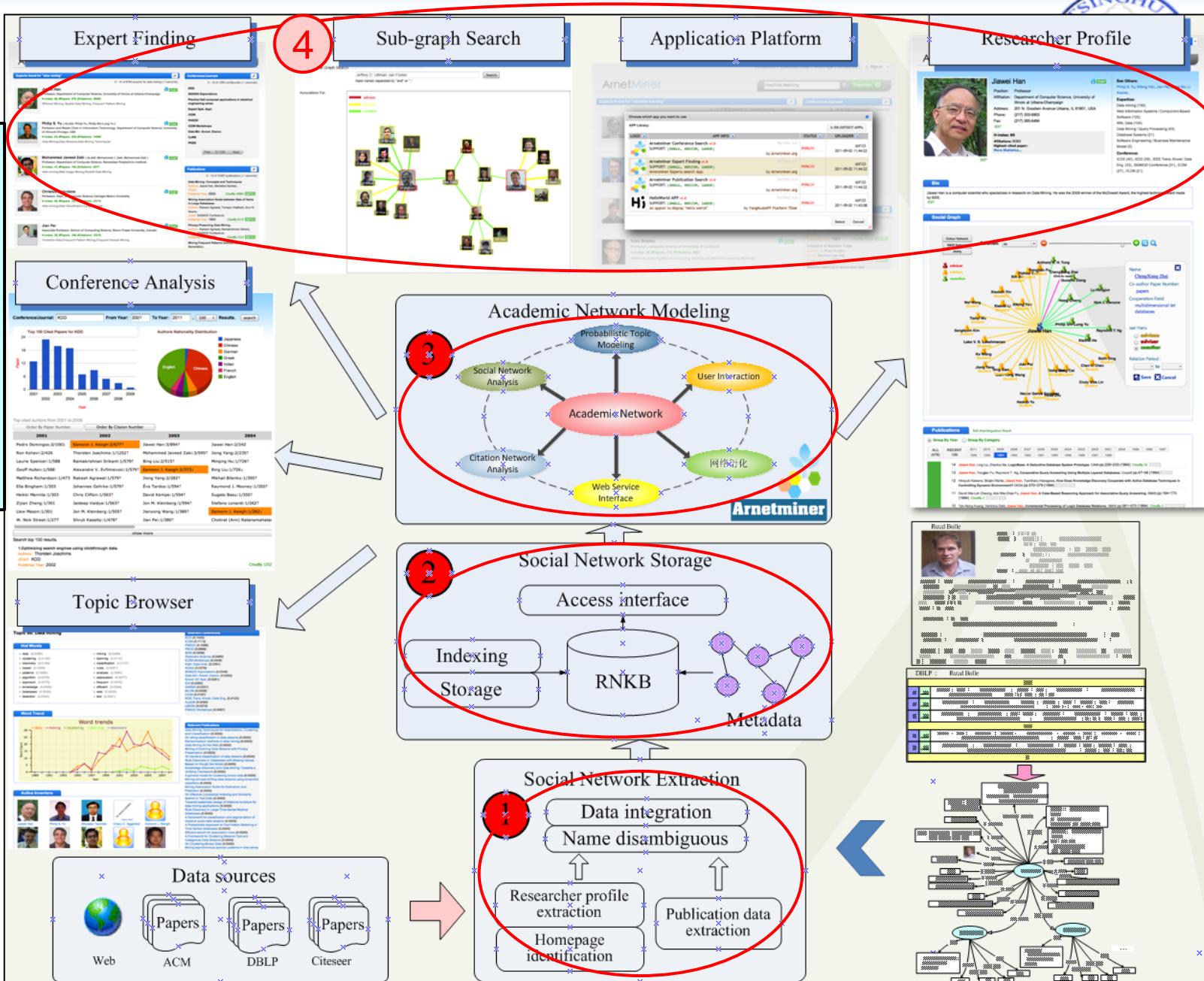
- Academic research social network analysis and mining system

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

ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin More Account

Search Experts Search

Jiawei Han

Position: Professor
Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign
Address: 201 N. Goodwin Avenue, Urbana, IL 61801, USA.
Phone: (217) 333-6903
Fax: (217) 265-6494
Email: han@cs.uiuc.edu
Links: [Home](#) [ORCID](#)

STATISTIC
H-index: 96 Uptrend: 30.46 Diversity: 0.71
#Papers: 553 Activity: 32.04 Sociability: 726.64
#Citations: 55885 Longevity: 26 [More Statistics...](#)

[Edit Profile](#)

[Facebook](#) [Twitter](#) [LinkedIn](#) [Google Scholar](#) [Scopus](#) [Publons](#) [ORCID](#) 18

Bio

Jiawei Han is a computer scientist who specializes in research on Data Mining. He was the 2009 winner of the McDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

Research Interest

Efficient Mining Spatial Data Mining Frequent Pattern Mining

[Edit Interest](#)

+ Show Temporal Interests (Do you want to see the change of his/her research interests?)

Education

Phd University Univ. Wisconsin-Madison

Phd Major Computer Science

Phd Date

[Edit](#)

Research Interests

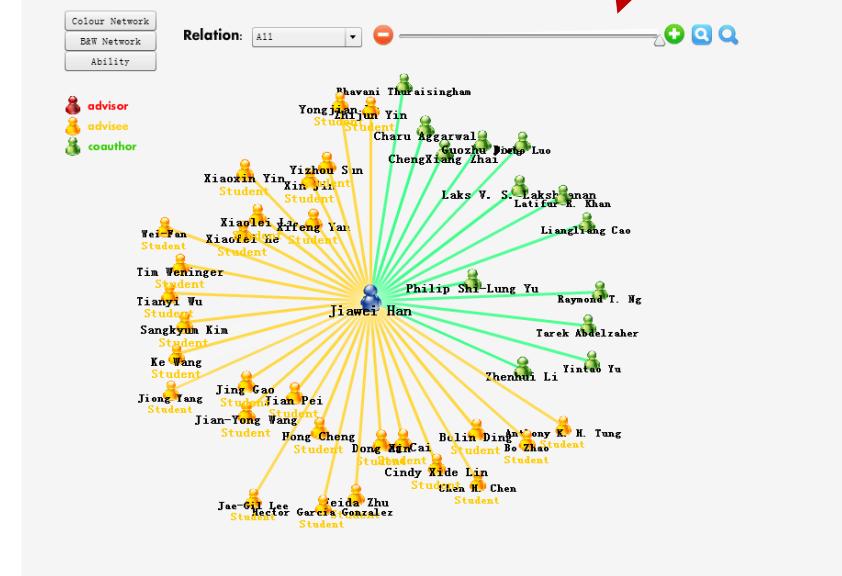
Publications

Basic Info.

Citation statistics

Social Network

Social Graph



Publications

Group By Year Group By Category

ALL	RECENT	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994
560	560																			

560 Query-Driven Discovery of Semantically Similar Substructures in Heterogeneous Networks

Xiao Yu, Yizhou Sun, Peixiang Zhao, Jiawei Han
2012, KDD, pp.1500-1503

BIBTEX PDF

559 Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks

Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, Xiao Yu
2012, KDD, pp.1348-1356

BIBTEX PDF

558 Mining periodic behaviors of object movements for animal and biological sustainability studies.

Zhenhai Li, Jiawei Han, Bolin Ding, Roland Keys
2012, Data Min. Knowl. Discov., pp.355-386

BIBTEX

Expertise Search

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

ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin More Welcome jietang Account Reset APPs

data mining Search

Do you mean: Mark Mine, Ray Mines, Data Becker

Experts found for "data mining" 0 - 15 of 8748 experts for data mining (0 seconds)

Jiawei Han Professor, Department of Computer Science, University of Illinois at Urbana-Champaign H-index: 96, #Papers: 553, #Citations: 55865 Efficient Mining Spatial Data Mining Frequent Pattern Mining FOAF

Philip S. Yu (ALIAS: Philip Yu, Philip Shui-Lung Yu) Professor and Wexler Chair in Information Technology, Department of Computer Science, University of Illinois Chicago H-index: 86, #Papers: 683, #Citations: 32506 Data Mining Data Streams Data Mining Techniques FOAF

Mohammed Javeed Zaki (ALIAS: Mohammed J. Zaki, Mohammed Zaki) Professor, Department of Computer Science, Rensselaer Polytechnic Institute H-index: 46, #Papers: 722, #Citations: 9895 Data Mining Web Usage Mining Parallel Data Mining FOAF

Christos Faloutsos Professor, Dept. of Computer Science Carnegie Mellon University H-index: 81, #Papers: 357, #Citations: 31344 Data Mining Data Visualization Large Graphs FOAF

Jian Pei Professor, School of Computing Science, Simon Fraser University H-index: 49, #Papers: 219, #Citations: 15467 Uncertain Data Frequent Pattern Mining Frequent Itemset Mining FOAF

H. Mannila (ALIAS: Heikki Mannila) Professor, Helsinki University of Technology H-index: 52, #Papers: 193, #Citations: 14783 Knowledge Discovery Data Mining Data Mining Methods FOAF

Charu C. Aggarwal (ALIAS: Charu Chandra, Charu Chandra Aggarwal, Charu Aggarwal) Research Scientist, IBM T. J. Watson Research Center H-index: 48, #Papers: 209, #Citations: 11828 FOAF

Conference/Journals 0 - 10 of 1373 conf/journals (0 seconds)

KDD SIGKDD Explorations Datenverarbeitungspraxis ICDM Expert Syst. Appl. PAKDD ICDM Workshops CoRR Data Min. Knowl. Discov. PKDD Prev 1 / 138 Next

Publications 0 - 10 of 15651 publications (0 seconds)

Data Mining: Concepts and Techniques Authors: Jiawei Han, Micheline Kamber. Published Year: 2000 CitedBy 13179 PDF BIBTEX

Mining Association Rules between Sets of Items in Large Databases. Authors: Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami. JConf: SIGMOD Conference Published Year: 1993 CitedBy 11393 PDF BIBTEX

From Data Mining to Knowledge Discovery in Databases. Authors: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. JConf: AI Magazine Published Year: 1996 CitedBy 4332 PDF BIBTEX

Introduction to Data Mining Authors: Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Published Year: 2005 CitedBy 2963 PDF BIBTEX

Mining Frequent Patterns without Candidate

Organization Search

Ranking organizations
on “machine learning”

ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin

machine learning

Do you mean: Line Learning

Organizations found for "machine learning"

0 - 10 of 100 organizations

- Microsoft Research
Members: David J. DeWitt, Andrew Blake, Eric Horvitz, David Heckerman.
- Carnegie Mellon University
Members: Christos Faloutsos, Edmund M. Clarke, Avrim Blum, Manuela M. Veloso.
- IBM Research
Members: Kenneth L. Clarkson, Gerald Tesauro, Tong Zhang, Michael Hind.
- Nanyang Technological University
Members: Dacheng Tao, Dong Xu, Yeng Chai Soh, Yew-Soon Ong.
- Stanford University
Members: Sebastian Thrun, Daphne Koller, Andrew Y. Ng, Yoav Singer.
- National University of Singapore
Members: Ming Li, Shuicheng Yan, Chew Lim Tan, Hwee Tou Ng.
- Yahoo! Research
Members: Raghu Ramakrishnan, Alex J. Smola, Usama M. Fayyad, John Langford.
- Massachusetts Institute of Technology
Members: Srinivas Devadas, Michael J. Kearns, Tomaso Poggio, Piotr Indyk.
- Google
Members: Alon Y. Halevy, Yoram Singer, Mehryar Mohri, Thomas Hofmann.

SINGHUA
KEG
VIVERSITY

Tsinghua University

Geographic search



The figure is a world map titled "Distribution of Top 1000 researchers by H-index". It shows the density of researchers across different countries, with larger circles indicating higher concentrations. A red arrow points to a cluster in North America, specifically around the United States and Canada. The map also includes labels for major oceans and some countries.

Distribution of Top 1000 researchers by H-index. Input keywords to search for researchers on specific topics.

data mining

Search Geo

ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin More Welcome jietang Account

地图 卫星

格陵兰

冰岛

芬兰

瑞典

挪威

波兰

乌克兰

(大不列颠)联合王国

意大利

西班牙

葡萄牙

阿尔及利亚

利比亚

埃及

沙特阿拉伯

伊拉克

伊朗

阿富汗

巴基斯坦

印度

中国

蒙古

日本

泰国

印度尼西亚

巴布亚新几内亚

澳大利亚

新西兰

南非

马达加斯加

博茨瓦纳

纳米比亚

安哥拉

坦桑尼亚

肯尼亚

尼日利亚

乍得

苏丹

埃塞俄比亚

马里

尼日尔

乍得

尼日利亚

乍得

苏丹

埃塞俄比亚

肯尼亚

坦桑尼亚

安哥拉

纳米比亚

博茨瓦纳

马达加斯加

南非

印度洋

南大西洋

南太平洋

北大西洋

北太平洋

1

Google

133

295

319

21

10

27

5

1

地图数据 ©2012 GS(2011)6020 MapLink, Tele Atlas - 使用条款

• Introduction • Contact Us • Development Team • Web Services • Location Search

Copyright © 2006-2011 KEG, Tsinghua. All Rights Reserved. | 京ICP备9068414号

Conference Analysis

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

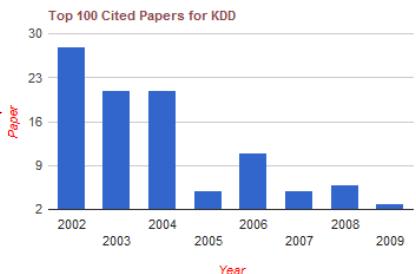
Who are the most cited authors?

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

Analysis of "KDD"

Conference Analysis

Conference/Journal: KDD From Year: 2002 To Year: 2011 , 100 Results. search



Authors Nationality Distribution



Top cited authors from 2002 to 2009

Order By Paper Number	Order By Citation Number	2002	2003	2004	2005	2006
Thorsten Joachims:1/1683	Jiawei Han:4/1668	Mingqiu Hu:1/1126	Christos Faloutsos:1/598	Andrew Tomkins:2/730		
Johannes Gehrke:3/1326	Éva Tardos:3/1079	Bing Liu:1/1126	Jon M. Kleinberg:1/598	Ravi Kumar:2/730		
Eamonn J. Keogh:2/829	David Kempe:1/1079	Inderjit S. Dhillon:2/525	Jure Leskovec:1/598	Thorsten Joachims:1/630		
Ramakrishnan Srikanth:1/675	Jon M. Kleinberg:1/1079	Mikhail Bilenko:1/461	Filip Radlinski:1/297	Xiangyang Lan:1/623		
Alexander V. Evfimievski:1/675	Subramanyam Mallela:1/674	Raymond J. Mooney:1/461	Thorsten Joachims:1/297	Daniel P. Huttenlocher:1		
Rakesh Agrawal:1/675	Inderjit S. Dhillon:1/674	Sugato Basu:1/461	Qiaozhu Mei:1/252	Jon M. Kleinberg:1/623		
Chris Clifton:1/664	Dharmendra S. Modha:1/674	Christos Faloutsos:2/403	ChengXiang Zhai:1/252	Lars Backstrom:1/623		
Jaideep Vaidya:1/664	Bing Liu:2/658	Brian Kulis:1/315	Ravi Kumar:1/171	Jasmine Novak:1/570		
Jon M. Kleinberg:1/646	Haixun Wang:1/604	Yuqiang Guan:1/315	Daniel Gruhl:1/171	Ingo Mierswa:1/527		
Pedro Domingos:2/612	Philip S. Yu:1/604		Andrew Tomkins:1/171	Michael Wurst:1/527		

show more

Search top 100 results.

1. Optimizing search engines using clickthrough data.

Authors: Thorsten Joachims

JConf: KDD

Published Year: 2002

CitedBy 1683

2. Mining and summarizing customer reviews.

Authors: Mingqiu Hu, Bing Liu

JConf: KDD

Published Year: 2004

CitedBy 1126

3. Maximizing the spread of influence through a social network.

Authors: David Kempe, Jon M. Kleinberg, Éva Tardos

JConf: KDD

Reviewer Suggestion

Interest matching
COI avoiding
Load balancing
Forecast review quality

Title
 ArnetMiner: extraction and mining of academic social networks.

Authors (Separated by comma)
 Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, Zhong Su

Abstracts

This paper addresses several key issues in the ArnetMiner system, which aims at extracting and mining academic social networks. Specifically, the system focuses on: 1) Extracting researcher profiles automatically from the Web; 2) Integrating the publication data into the network from existing digital libraries; 3) Modeling the entire academic network; and 4) Providing search services for the academic network. So far, 448,470 researcher profiles have been extracted using a unified tagging approach. We integrate publications from online Web databases and propose a probabilistic framework to deal with the name ambiguity problem. Furthermore, we propose a unified modeling approach to

Conference (Journal)
 SEARCH

Keywords: +

academic social network Providing search service expertise search

people association search

Relevant Conferences/Journals: +

ICDM PAKDD PKDD SDM Discovery Science

The recommended reviewers: BM25 +T +A +S All

[View All Relevant Publications](#)

 **Yajun Wang**
 Microsoft Research Asia
 H-index: 14, #Papers: 49, #Citations: 677
 social network,principal component analysis,Shortest Path
[+ Relevant Publications](#)

 **Ming-Syan Chen** (ALIAS: Ming-Syan Chen)
 National Taiwan University
 H-index: 42, #Papers: 278, #Citations: 9978
 Data Mining,Data Streams,Data Replication
[+ Relevant Publications](#)

 **Michael R. Berthold** (ALIAS: Michael Berthold)
 KNIME.com, University of Konstanz
 H-index: 17, #Papers: 77, #Citations: 1180
 International Symposium,Data Analysis,Second International Symposium
[+ Relevant Publications](#)

 **Christoph Lingenfelder**
 German Software Development Lab, IBM
 H-index: 4, #Papers: 12, #Citations: 35
 Knowledge-Based Methods,Proof Transformation,Der rechtliche Schutz von
[+ Relevant Publications](#)

 **Michael Zeller**
 Zementis
 H-index: 4, #Papers: 7, #Citations: 68
 cloud computing, neural networks, open standard, predictive analytics, data mining, predictive model markup language, pmml
[+ Relevant Publications](#)

 **Theodoros Lappas**
 H-index: 4, #Papers: 12, #Citations: 124
 social network,interactive recommendation,Efficient Confident Search
[+ Relevant Publications](#)

16

Cross-domain Collaboration Recommendation

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

Who are the best
collaborators on each
topic?

Recommend Collaborators for Jie Tang from "medical informatics"

Tips: left click Topic Name to search under sub-topic, right click to return back.

medical informatics					
Knowledge-based	data analysis / Medical Costs	Image Registration / MR Images	Rough Set		
Reima Suomi	David Schlangen Owen Eriksson	N. Ayache	Benedetto Metzger	Weizhi Wu	
Vassilis Koutkias	Angelina A. Tzacheva Rita M. D'Angelis	Ferenc A. Jolesz R. Razavi	Jaroslaw Stepienuk Pradipita Majumder		
Computer science			T. M. Peters	E. Tadevosyan Chuangyin Dang	Ivo Dutschke
I. Foster	J. Hartmanis Lynn Andrea Stein	Massimo Franceschet	Gerard Salton Clement T. Yu Abdur Chowdhury Nicholas A. Kraft	Ulf Konnerup-Jensen	
	Kim Bruce	Alfs Berztiss		Ramashray Halder	
Information Retrieval / Probabilistic Indexing					

Recommend Collaborators for Jie Tang from "medical informatics"

 **B. J. Liu** (ALIAS: Brent J. Liu, Brent Julius Liu) 

H-index: 8, #Papers: 16, #Citations: 159
[Data Grid](#) [Database Semantics](#) [Medical Imaging Application](#)

 **Quanzhi Li** 

Information Systems Department New Jersey Institute of Technology University Heights
H-index: 2, #Papers: 14, #Citations: 12
[Automatic Classification](#) [Search Result](#) [Medical Document](#)

 **Shusaku Tsumoto** 

Professor, Department of Medical Informatics, Shimane University School of Medicine
H-index: 19, #Papers: 264, #Citations: 1628
[Data Mining](#) [Rough Sets](#) [Rough Set Theory](#)

 **H. K. Huang** 

Professor, Department of Biomedical Engineering University of Southern California, USA
H-index: 13, #Papers: 30, #Citations: 553
[Data Grid](#) [Medical Imaging Application](#) [Digital Mammography Images](#)

 **Rachel S. Levy-Drummer** 

H-index: 1, #Papers: 2, #Citations: 5
[Biomathematics Oriented Machine Learning](#) [Clinical Markers](#) [Reconstructing Temporal Profiles](#)

Target Domain & Your Background

Target Domain medical informatics

Jie Tang
Department of Computer Science and Technology, Tsinghua University, China

Interest Social Network, Graph Model, Data Mining, Semantic web

Search

Relevant Publications

Medical Informatics.
Authors: Alan L. Rector.
JConf: Description Logic Handbook
Published Year: 2003
[CitedBy 36](#) [PDF](#) [BIBTEX](#)

Medical Informatics.
Authors: Chin-Yin Huang.
JConf: Handbook of Automation
Published Year: 2009
[PDF](#) [BIBTEX](#)

From medical imaging to medical informatics.
Authors: Henning Müller, Xiaohong W. Gao, Shujian Luo.
JConf: Computer Methods and Programs in Biomedicine
Published Year: 2008
[BIBTEX](#)

American Medical Informatics Review for 2011.
Authors: Ralph Grams.
JConf: J. Medical Systems
Published Year: 2012
[BIBTEX](#)

Medical Education and Role of Medical Informatics.
Authors: Izet Masic, Ahmed Novo, Selim Toromanovic, Almir Dzananovic, Zlatan Masic, Alisa Piralic, Sejla Deljkovic.
JConf: MIE
Published Year: 2006
[CitedBy 2](#) [BIBTEX](#)

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 (0.0086)
- » 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, Churn C. Aggarwal, Eamonn J. Keogh, Heikki Mannila, Christos Faloutsos, Mohammed Javeed Zaki, Einoshin Suzuki, Tao Li

Popular Phrases

» data mining	» data stream
» binary data	» concept-drifting data stream
» data mining application	» massive audio data stream
» time series data	» Categorical Data Streams
» Clustering Binary Data	» 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)
- UBIDM (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



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: 45244
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 networks / 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
Homepage: [http://www.hongkongpolytechnic.edu.hk/~yanliu/](#)

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/~nlui/>

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
Homepage: [http://www.cs.illinois.edu/~homes/barna/](#)

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://arxiv.org>)

- Online since 2006
- >1 million researcher profiles
- >131 million requests
- >2.35 Terabyte data
- 100K IP access from 170 countries per month
- 10% increase of visits per month

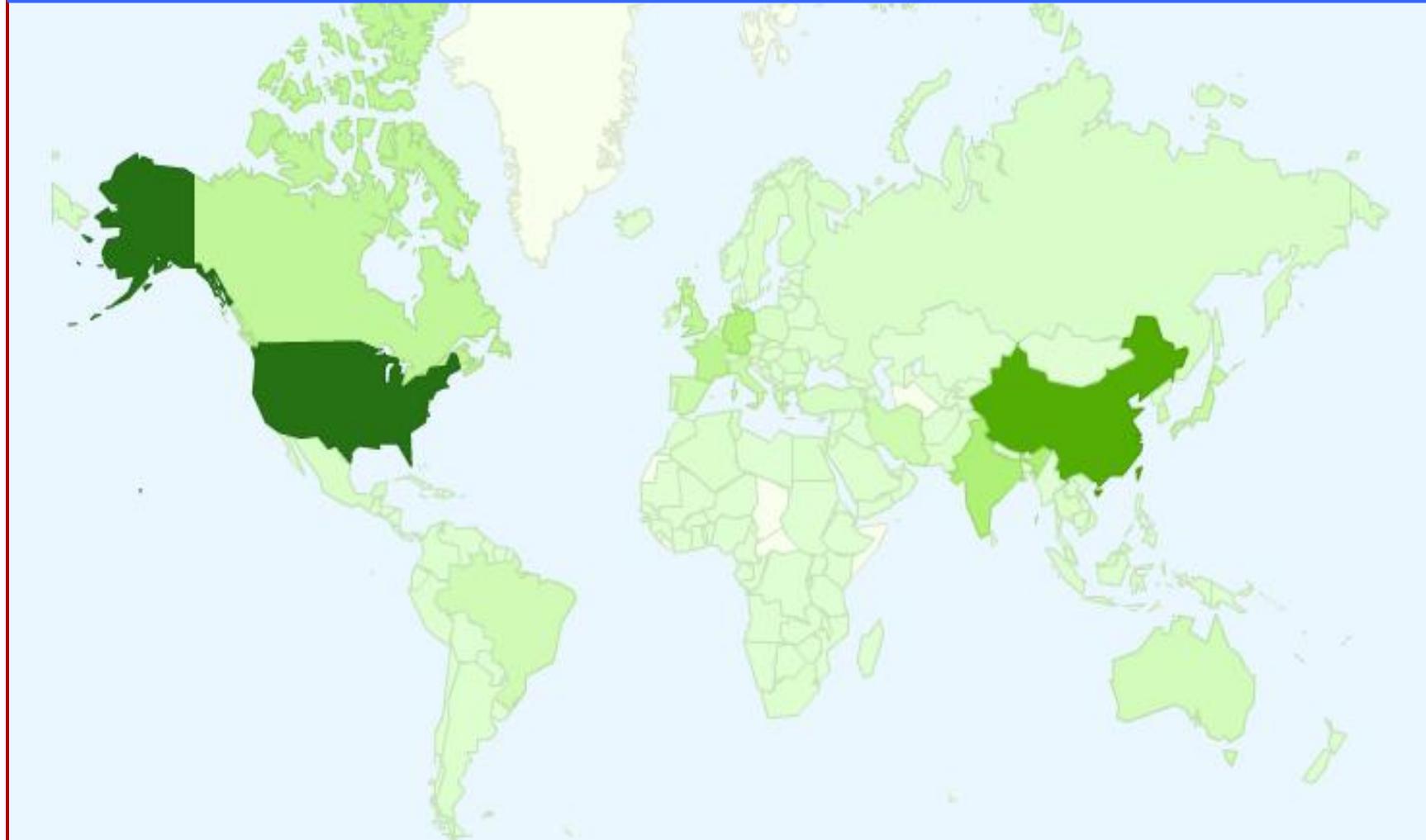
- Deep SN analysis, mining, and search

The screenshot shows two main sections of the ArnetMiner website:

- Researcher Profile:** A detailed profile page for "Jiawei Han". It includes a photo, basic information (Position: Professor, Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign), and statistics (H-index: 96, #Papers: 553, #Citations: 55885). Below this is a "Bio" section and a "Research Interest" section listing "Efficient Mining", "Spatial Data Mining", and "Frequent Pattern Mining".
- Social Graph:** A network diagram showing connections between various researchers. Nodes represent individuals, and edges represent relationships like "coauthor" or "collaborator". A central node is Jiawei Han, with many other names like Tony X. He, Christos Faloutsos, Jian Pei, and H. Mannila visible.

User Distribution

4.04 million IP from 220 countries/regions





User Distribution

4.04 million IP from 220 countries/regions

Top 10 countries

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



Outline

- ArnetMiner: Academic Social Network
- Core Techniques
 - Knowledge Acquisition
 - Modeling and Heterogeneous Ranking
 - Social Network Analysis

CT1: Knowledge Acquisition from Social Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)

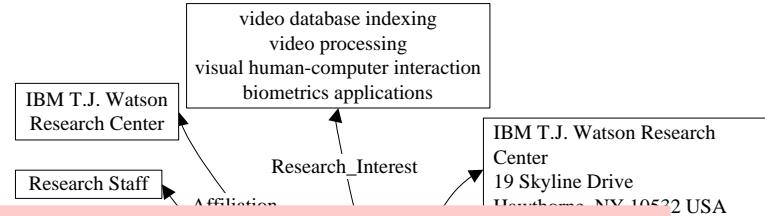




Ruud Bolle 1

Office: 1S-D58
Letters: IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598 USA
Packages: IBM T.J. Watson Research Center
19 Skyline Drive

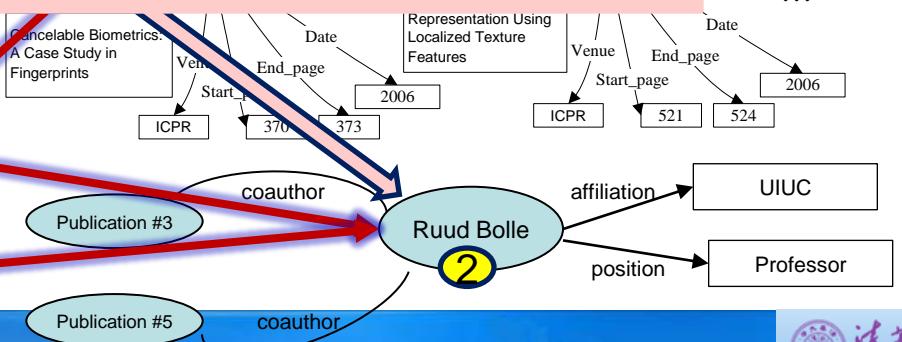
Contact Information



Two questions:

- How to accurately extract the researcher profile information from the Web?
- How to integrate the information from different sources?

DBLP: R	
50	EE 1 Nalini K. Ratha, Jonathan Connell, Ruud M. Bolle, Sharat Chikkerur: Cancelable Biometrics: A Case Study in Fingerprints. ICPR (4) 2006: 370-373
49	EE 2 Sharat Chikkerur, Sharath Pankanti, Alan Jea, Nalini K. Ratha, Ruud M. Bolle: Fingerprint Representation Using Localized Texture Features. ICPR (4) 2006: 521-524
43	EE 1 Andrew Senior, Arun Hampapur, Ying-li Tian, Lisa Brown, Sharath Pankanti, Ruud M. Bolle: Appearance models for occlusion handling. Image Vision Comput. 24(11): 1233-1243 (2006)
2005	
47	EE 1 Ruud M. Bolle, Jonathan H. Connell, Sharath Pankanti, Nalini K. Ratha, Andrew W. Senior: The Relation between the ROC Curve and the CMC. AutoID 2005: 15-20
46	EE 2 Sharat Chikkerur, Venu Govindaraju, Sharath Pankanti, Ruud M. Bolle, Nalini K. Ratha: Novel Approaches for Minutiae Verification in Fingerprint Images. WACV. 2005: 111-116
...	





Researcher Profile Database[1]

ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin More Account Welcome jietang FOAF Account

Search Experts Search

Jiawei Han

Position: Professor
Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign
Address: 140 N Goodwin Avenue, Urbana, IL 61801, USA
Phone: (217) 333-6932
Fax: (217) 265-6494
Email: han@cs.uiuc.edu
Links: [ORCID](#) [Google Scholar](#)

STATISTIC
H-index: 96 Uptrend: 30.46 Diversity: 0.71
#Papers: 553 Activity: 32.04 Sociability: 726.64
#Citations: 55885 Longevity: 26 [More Statistics...](#)

Bio
Jiawei Han is a computer scientist who specializes in research on Data Mining. He was the 2009 winner of the McDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

Research Interest
[Efficient Mining](#) [Spatial Data Mining](#) [Frequent Pattern Mining](#) [Edit Bio](#)

Education
Phd University [Edit](#)

M. I. Jordan

FOAF Follow

ALIAS: Michael I. Jordan, Michael Jordan, Michael Irwin Jordan

Position: Professor
Affiliation: Department of EECS Department of Statistics University of California, Berkeley
Address: University of California, Berkeley EECS Department 731 Soda Hall #1776 Berkeley, CA 94720-1776
Phone: (510) 642-3806
Fax: (510) 642-5775
Email: jordan@stat.berkeley.edu
Links: [ORCID](#) [Google Scholar](#) [Edit Profile](#)

STATISTIC
H-index: 75 Uptrend: 7.2 Diversity: 0.03
#Papers: 242 Activity: 11.12 Sociability: 331.69
#Citations: 44312 Longevity: 23 [More Statistics...](#)

H. Garcia

FOAF Follow

ALIAS: H. Garcia Molina, H. Garcia-Molina, Hector Garcia Molina, Hector Garcia Molina

Position: Professor
Affiliation: Departments of Computer Science and Electrical Engineering, Stanford University
Address: Department of Computer Science Stanford University Gates Hall 4A, Room 434 Stanford, CA 94305-9040 USA
Phone: (650) 723-0885
Fax: (650) 725-2588
Email: hector@cs.stanford.edu

See Others:
Andreas Paepcke Cauthor-Count: 32 H-index: 37
Jennifer Widom Cauthor-Count: 24 H-index: 79
D. Barbara Cauthor-Count: 26 H-index: 0

Expertise:
Data (115) Database Systems (60) Time Systems / Automated Ware Test Data (30) Mining (23) Mobile Robot / Hybrid Control (22) General library / Information Access [Edit Profile](#)

Conference:
KDD (49) ICDE (40) IEEE Trans. Knowl. Data Eng. (36) SIGMOD Conference (32) VLDB (21)

ArnetMiner Home Conference Collaborator

Search Experts

Scott

Position:
Affiliation:
Address:
Phone:
Links: [Edit](#)

STATISTIC
H-index: 96 Uptrend: -4.04 Diversity: 0.22
#Papers: 195 Activity: 4.86 Sociability: 407.19
#Citations: 57908 Longevity: 25 [More Statistics...](#)

Expertise:
Wireless network / End-to-end Routing Behavior (80) ATM Networks (21)

Research Interest
[Database Systems](#) [Data Management](#) [Data Warehousing](#) [Edit Interest](#)

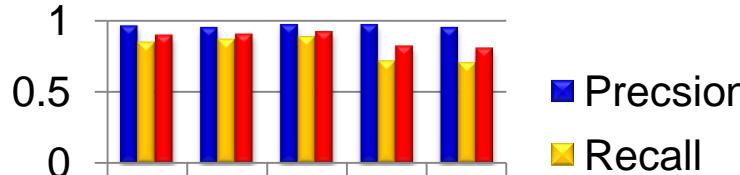
Monterrey, Mexico, in 1974. From Stanford University, Stanford, California, he received in 1975 a MS in electrical engineering and a PhD in computer science in 1979. He holds an honorary PhD from ETH Zurich (2007). Garcia-Molina is a Fellow of the Association for Computing Machinery and of the American Academy of Arts and Sciences; is a member of the National Academy of Engineering; received the 1999 ACM SIGMOD Innovations Award; is a Venture Advisor for Onset Ventures, and is a member of the Board of Directors of Oracle.

[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

RiMOM^[1-3]-A Tool for Semantic Integration (OAEI'06-09)

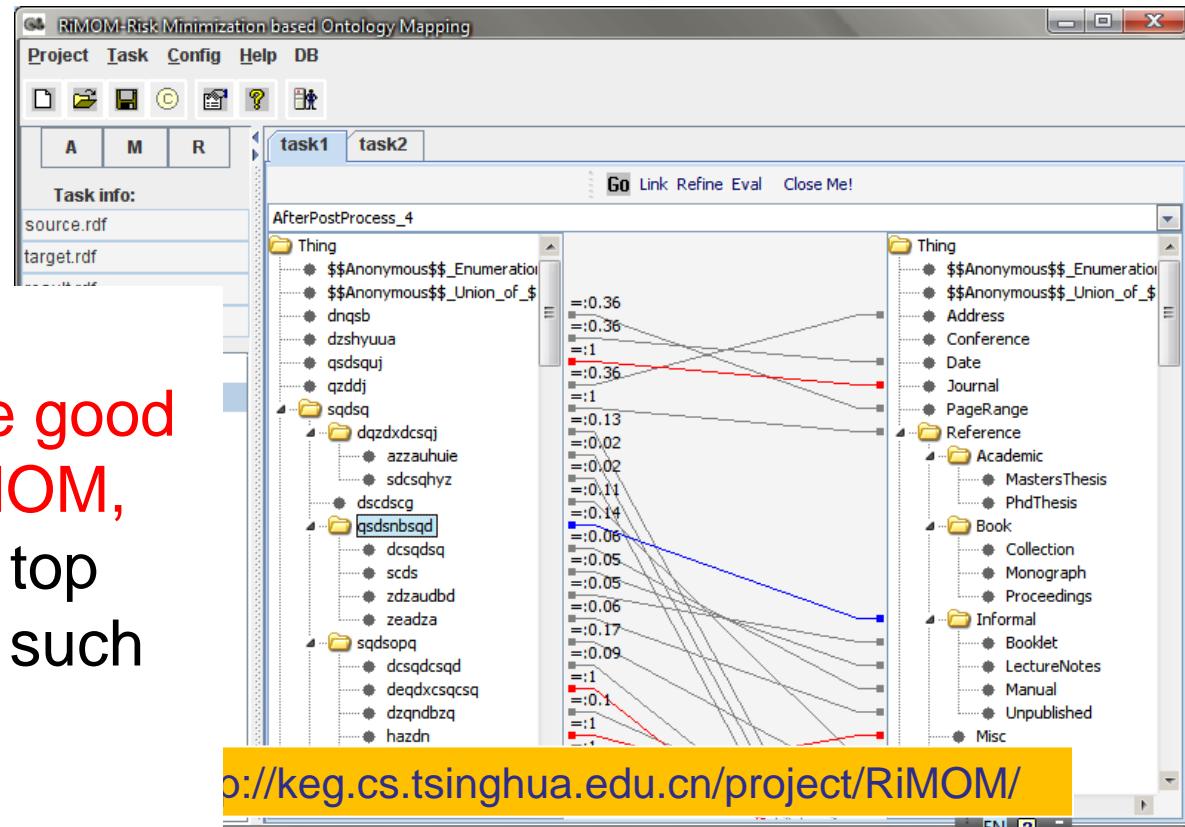


Benchmark Results



Chair Message:

“I’m really surprised by the good results of these years RiMOM, you can compete with the top systems that make use of such background knowledge.”



- [1] J. Li, J. Tang, Y. Li, and Q. Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE Transaction on Knowledge and Data Engineering (TKDE). August 2009 (vol. 21 no. 8) pp. 1218-1232.
- [2] Q. Zhong, H. Li, J. Li, G. Xie, J. Tang, and L. Zhou. A Gauss Function based Approach for Unbalanced Ontology Matching. SIGMOD’09, pages 669-680, 2009.
- [3] Z. Wang, J. Li, Z. Wang, and J. Tang. Cross-lingual Knowledge Linking Across Wiki Knowledge Bases. WWW’12, pages 459-468, 2012.



Name Disambiguation^[1,2]

Jing Zhang

List of publications from the [DBLP Bibliography Server](#) - [FAQ](#)

[Coauthor Index](#) - Ask others: [ACM DL](#) - [ACM Guide](#) - [CiteSeer](#) - [CSB](#) - [Google](#)

Name	Affiliation
Jing Zhang (26)	Shanghai Jiao Tong Univ.
	Yunnan Univ.
	Tsinghua Univ.
	Alabama Univ.
	Univ. of California, Davis
	Carnegie Mellon University
	Henan Institute of Education

2007		
83	EE	Jing Zhang, Guizhong Liu: Hyperspectral images lossless compression by a novel three-dimensional wavelet coding
82	EE	Jing Peng, Dongqing Yang, Changjie Tang, Jing Zhang, Jianjun Hu: CACS: A Novel Classification Algorithm Base
81	EE	Jing Zhang, Xi Chen, Ming Li: Computing Exact p-Value for Structured Motif. <i>CPM 2007</i> : 162-172
80	EE	Jing Zhang, Jie Tang, Juan-Zi Li: Expert Finding in a Social Network. <i>DASFAA 2007</i> : 1066-1069
79	EE	Guojun Chen, Jing Zhang, Xiaoli Xu, Yuan Yin: Real-Time Visualization of Tire Tracks in Dynamic Terrain with LC
78	EE	Jing Zhang, Hai Huang: Federate Job Mapping Strategy in Grid-Based Virtual Wargame Collaborative Environme
77	EE	Maria Wimmer, Michael Goul, Jing Zhang: Minitrack: E-Government Information and Knowledge Management. <i>H</i>
76	EE	Kai Kang, Jing Zhang, Baoshan Xu: Optimizing the Selection of Partners in Collaborative Operation Networks. <i>IC</i>
75	EE	Lingshuang Shao, Jing Zhang, Yong Wei, Junfeng Zhao, Bing Xie, Hong Mei: Personalized QoS Prediction for We
74	EE	Benyong Liu, Jing Zhang, Xiaowei Chen: Adaptive Training of a Kernel-Based Representative and Discriminative 1
73	EE	Jilong Wang, Jing Zhang: Federation Based Solution for Peer-to-Peer Network Management. <i>International Confer</i>
72	EE	Jing Zhang, Fanhai Shi, Jianhua Wang, Yuncai Liu: 3D Motion Segmentation from Straight-Line Optical Flow. <i>M</i>

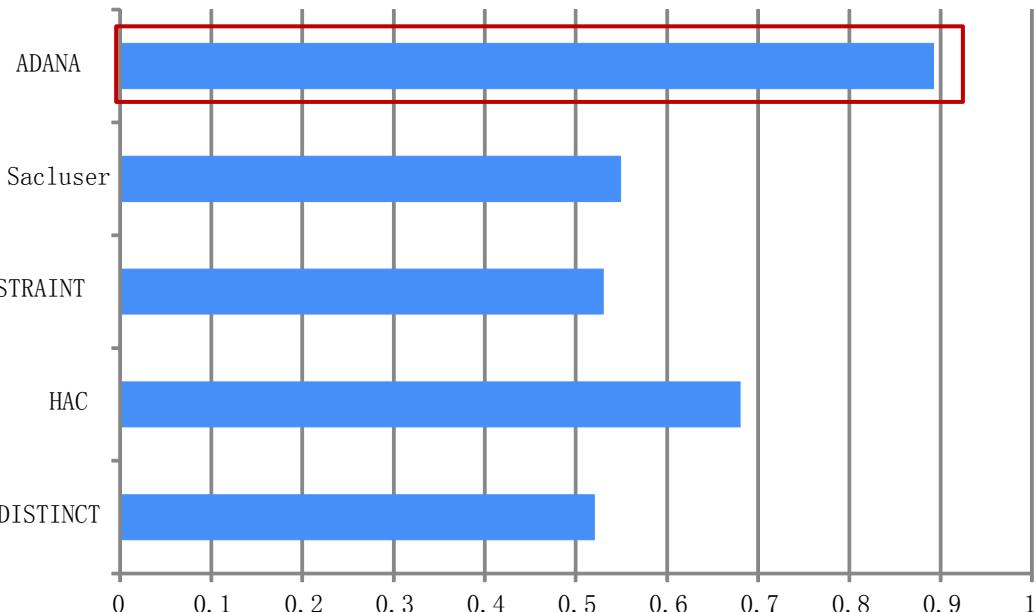
- How to perform the assignment automatically?
- How to estimate the person number?

[1] J. Tang, A.C.M. Fong, B. Wang, and J. Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *IEEE Transaction on Knowledge and Data Engineering (TKDE)*, Volume 24, Issue 6, 2012, Pages 975-987.

[2] X. Wang, J. Tang, H. Cheng, and P. S. Yu. ADANA: Active Name Disambiguation. *ICDM'11*, pages 794-803, 2011.



Disambiguation Performance



ArnetMiner Home Conference Collaborator Geo Search Topics Download Admin More Welcome jietang Account

Name Disambiguation Results for 5 "Hang Li". Show all Hang Li. 1/5

Search Experts Search

Hang Li FOAF Follow

See Others:

- Tie-Yan Liu Coauthor-Count: 30 H-index: 25
- Tao Qin Coauthor-Count: 13 H-index: 16
- Naoki Abe Coauthor-Count: 13 H-index: 22

Position: Senior Researcher and Research Manager
Affiliation: Information Retrieval and Mining Group, Microsoft Research Asia
Address: No. 5, Da Ling Street, Haidian District, Beijing China, 100080
Phone: (86-10)58963177
Fax: (86-10)82868529
Email: hangli@microsoft.com

Links: [Edit Profile](#)

STATISTIC

H-index:	36	Uptrend:	2.98	Diversity:	1.35
#Papers:	121	Activity:	6.77	Sociability:	204.88
#Citations:	3708	Longevity:	17	More Statistics...	

EDIT

[f](#) [t](#) [m](#) [y](#) [n](#) [g](#) [d](#) [s](#) [c](#) [v](#) 0

Bio

I am senior researcher and research manager of Information Retrieval and Mining Group at Microsoft Research Asia. I am also adjunct professor of Peking University, Nanjing University, Xian Jiaotong University, and Nankai University. I am co-director of the MS joint lab at PKU and member of the machine learning lab at NJU.

I joined Microsoft Research in June 2001. Prior to that, I worked at the Research Laboratories of NEC Corporation.

I obtained a B.S. in Electrical Engineering from Kyoto University in 1988 and a M.S. in Computer Science from Kyoto University in 1990. I earned my Ph.D. in Computer Science from the University of Tokyo in 1998.

I am interested in statistical learning, information retrieval, data mining, and natural language processing.

EDIT BIO

Research Interest

Information Retrieval Text Classification Web Search [Edit Interest](#)

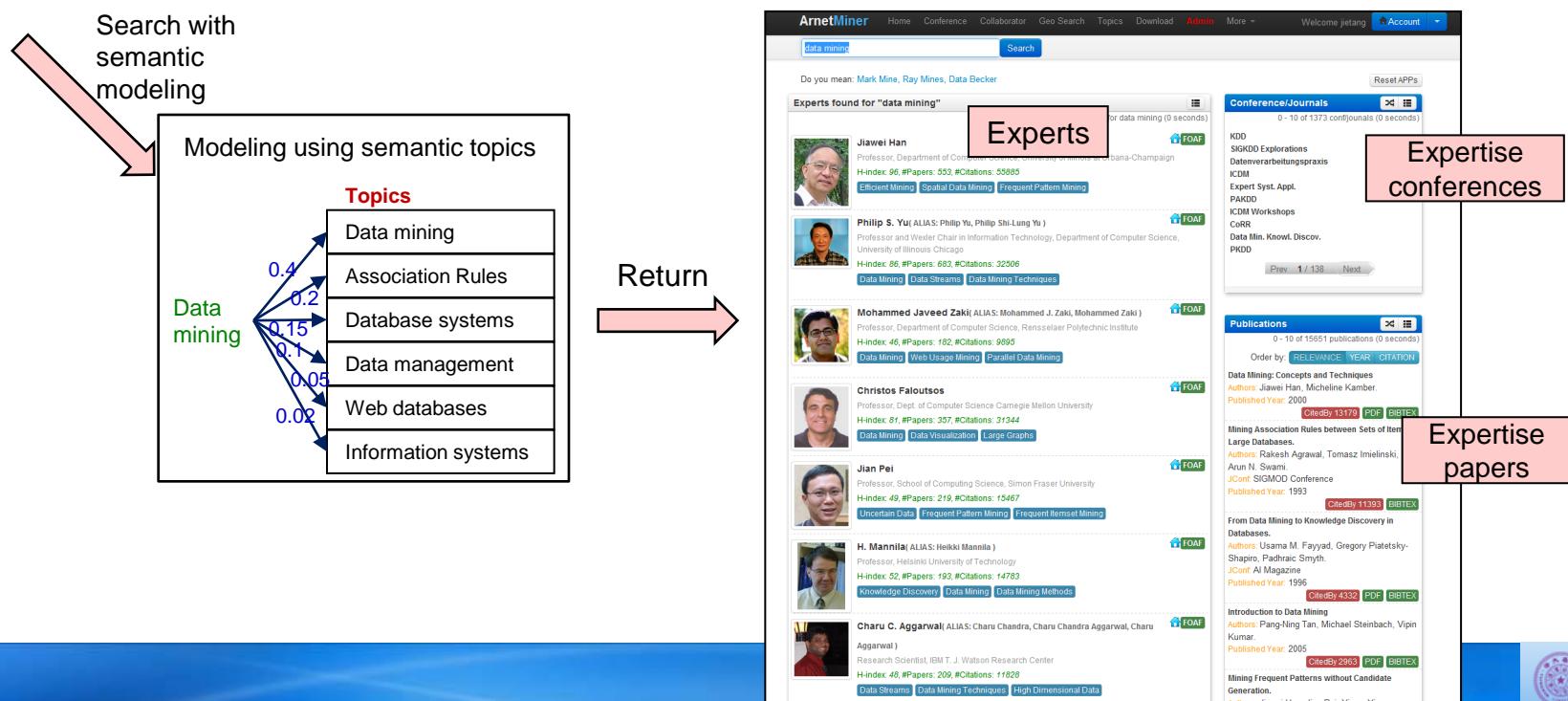
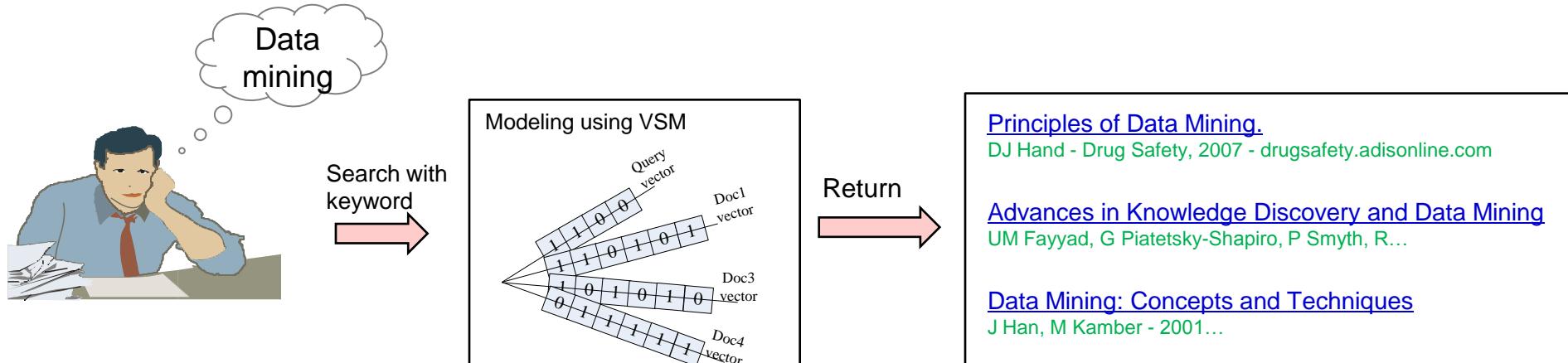


Outline

- ArnetMiner: Academic Social Network
- Core Techniques
 - Knowledge Acquisition
 - **Modeling and Heterogeneous Ranking**
 - Social Network Analysis

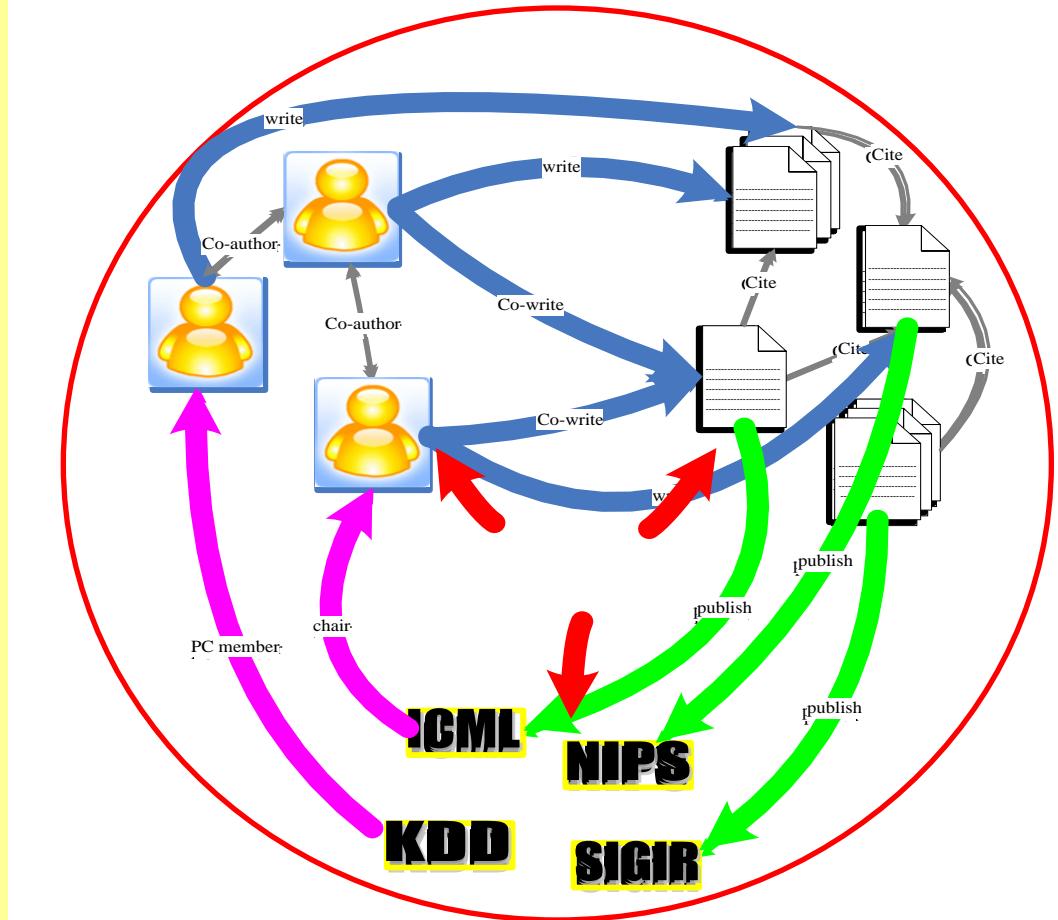
CT2: Topic-based Heterogeneous Ranking

(KDD'08, KDD'12, Machine Learning)

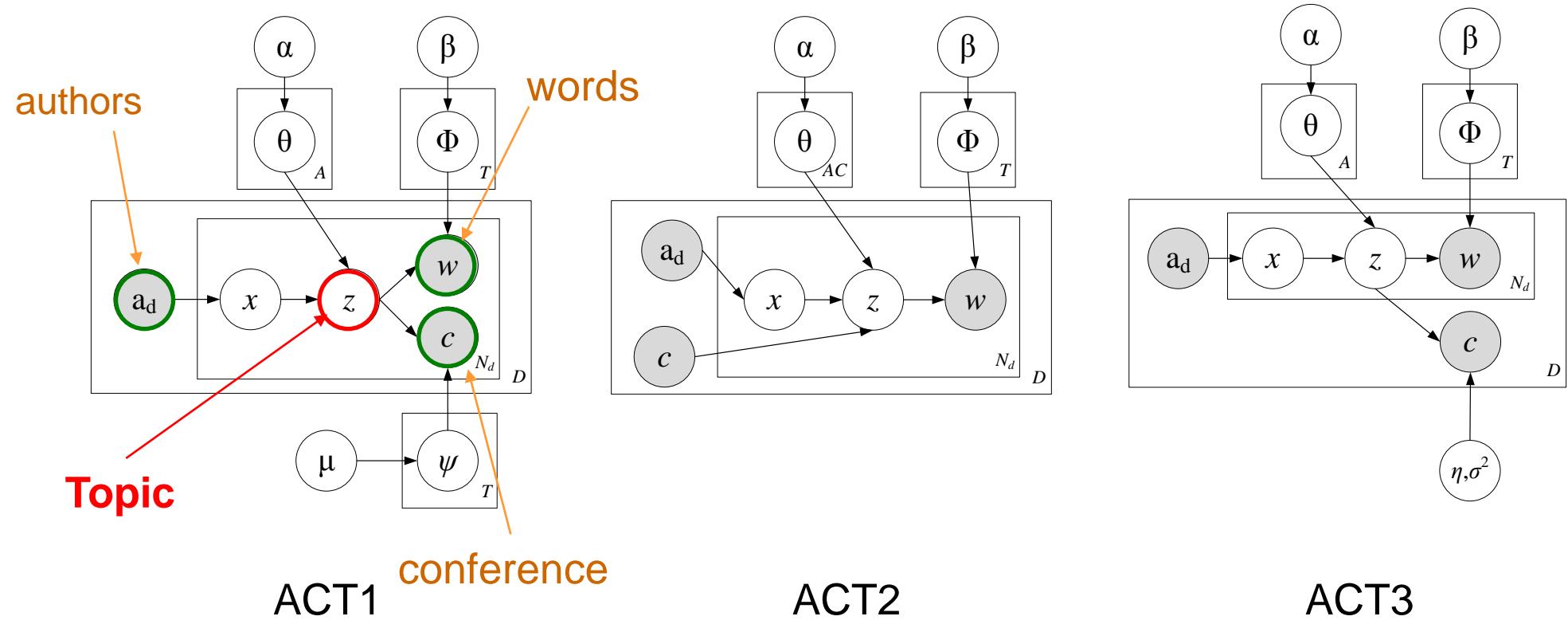


Challenges

1. How to model the heterogeneous academic network?
 2. How to capture the link information for ranking objects in the academic network?
- Topic



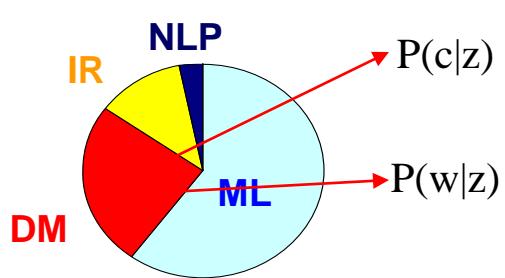
(1) Modeling the Academic Network^[1]



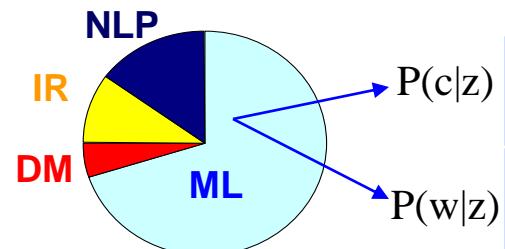
Author-Conference-Topic Model

Generative Story of ACT1 Model

- Generative process



Shafiei



Miliros

Paper

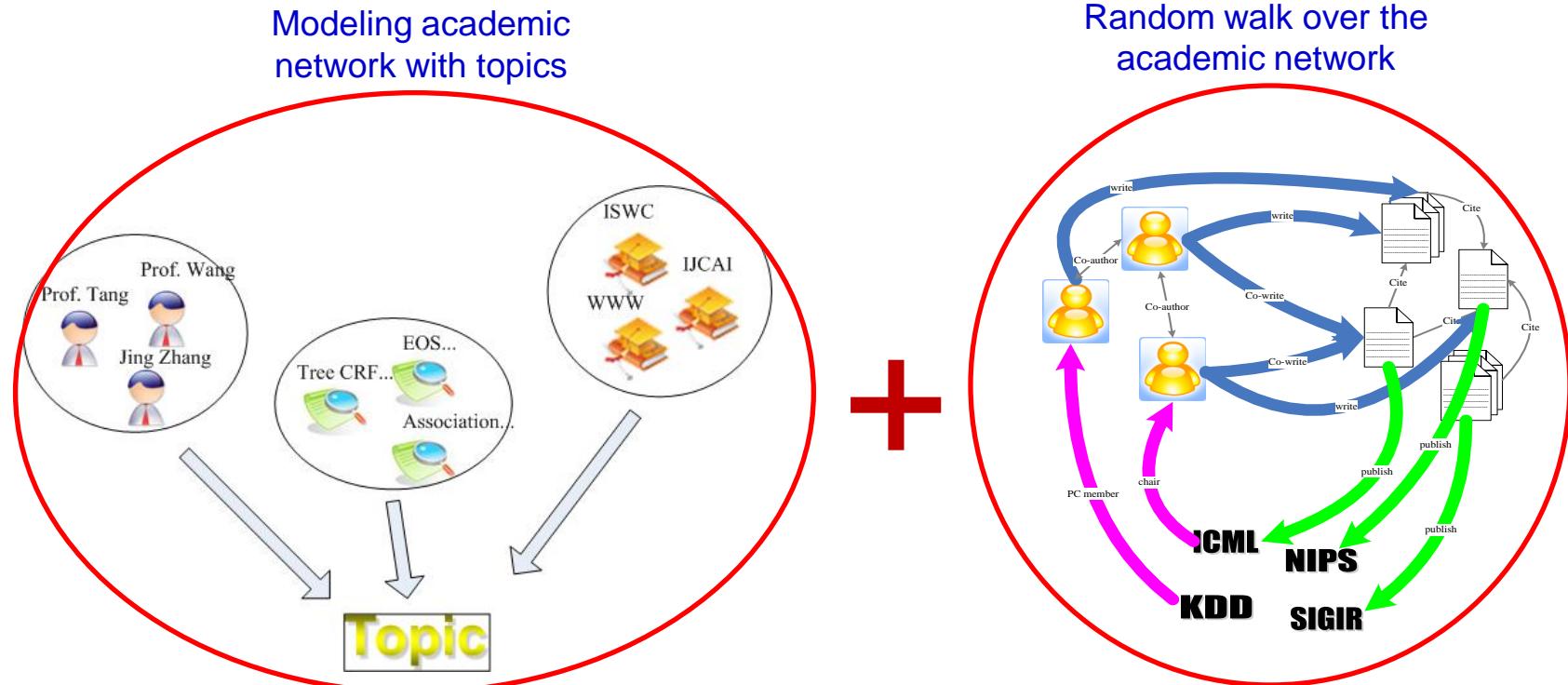
Latent Dirichlet Co-clustering

Shafiei and Miliros

ICDM **NIPS**

We present a generative model for **clustering documents and terms**. Our model is a four **hierarchical bayesian model**. We present efficient **inference techniques** based on **Markow Chain Monte Carlo**. We report results in document modeling, document and terms clustering ...

(2) Integrating with Random Walk^[1]



Author-Conference-Topic
Model [Tang et al., 08]

[1] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic Level Expertise Search over Heterogeneous Networks. Machine Learning Journal, Volume 82, Issue 2 (2011), Pages 211-237.



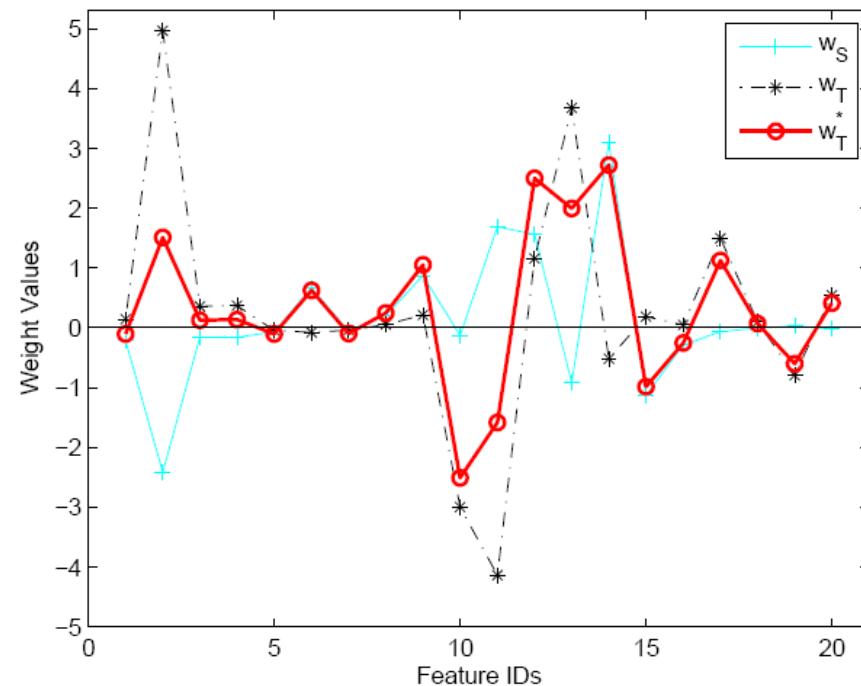
Experimental Results

- Data sets
 - Homogeneous Data
 - LETOR 2.0: TREC2003, TREC2004, and OHSUMED
 - Heterogeneous Data
 - Academic network consisting of 14,134 authors, 10,716 papers, and 1,434 conferences.
 - Heterogeneous Tasks
 - Expert finding vs. Bole search
- Baselines
 - RSVM
 - Language model

Results on Heterogeneous Data

Table Performances of different approaches for expert finding.

Approach	MAP	N@1	N@3	N@5	N@10
Libra	.5104	.4800	.4634	.4467	.4978
Rexa	.4621	.4512	.4332	.4236	.4798
DCVIM	.5004	.5071	.5020	.5024	.5202



Features	Weights
S1	2.7201
L10	-2.5080
H2	2.5018
H3	1.9956
H1	-1.5827
L2	1.5122
S4	1.1284
L9	1.0525
S2	-0.9863
L6	0.6276



Results on Heterogeneous Tasks

- Expert finding verse Bole search (finding best supervisor)
- To obtain ground truth of bole for each query
 - We sent emails to 50 senior researchers and 50 junior researchers (91.6% are post doc or graduates)
 - Average their feedbacks

Table Results on heterogeneous tasks.

Approach	P@5	P@10	P@15	MAP	N@5	N@10
RSVM	.7714	.8429	.8285	.7756	.5545	.5947
RSVMt	.8000	.8286	.8476	.7837	.5923	.5999
MTR SVM	.8000	.8286	.8476	.7875	.6140	.6075
HCDRank	.8285	.7857	.8571	.7971	.6189	.6112
Language model	.6250	.6875	.6500	.6726	.3343	.3809



Outline

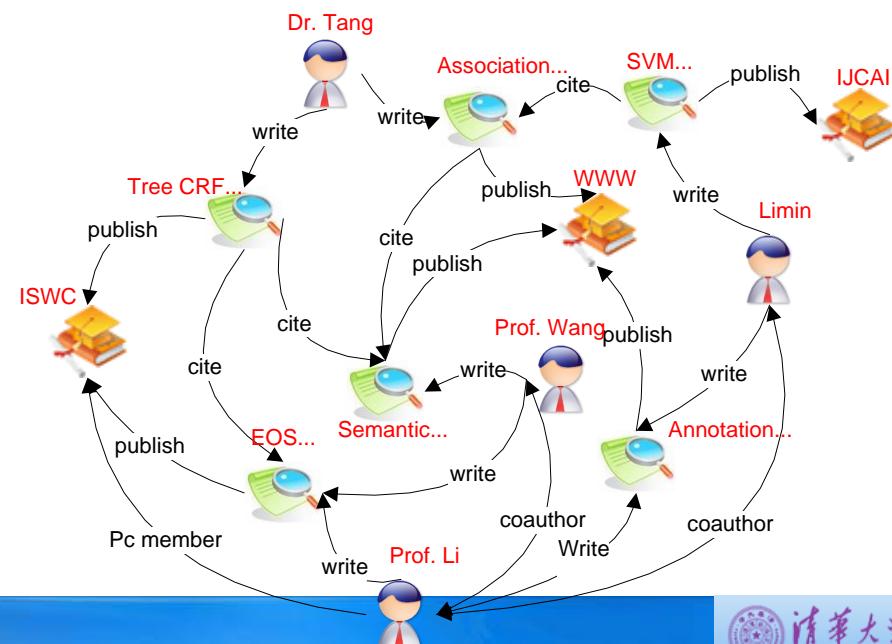
- ArnetMiner: Academic Social Network
- Core Techniques
 - Knowledge Acquisition
 - Modeling and Heterogeneous Ranking
 - **Social Network Analysis**

CT3: Social Network Analysis

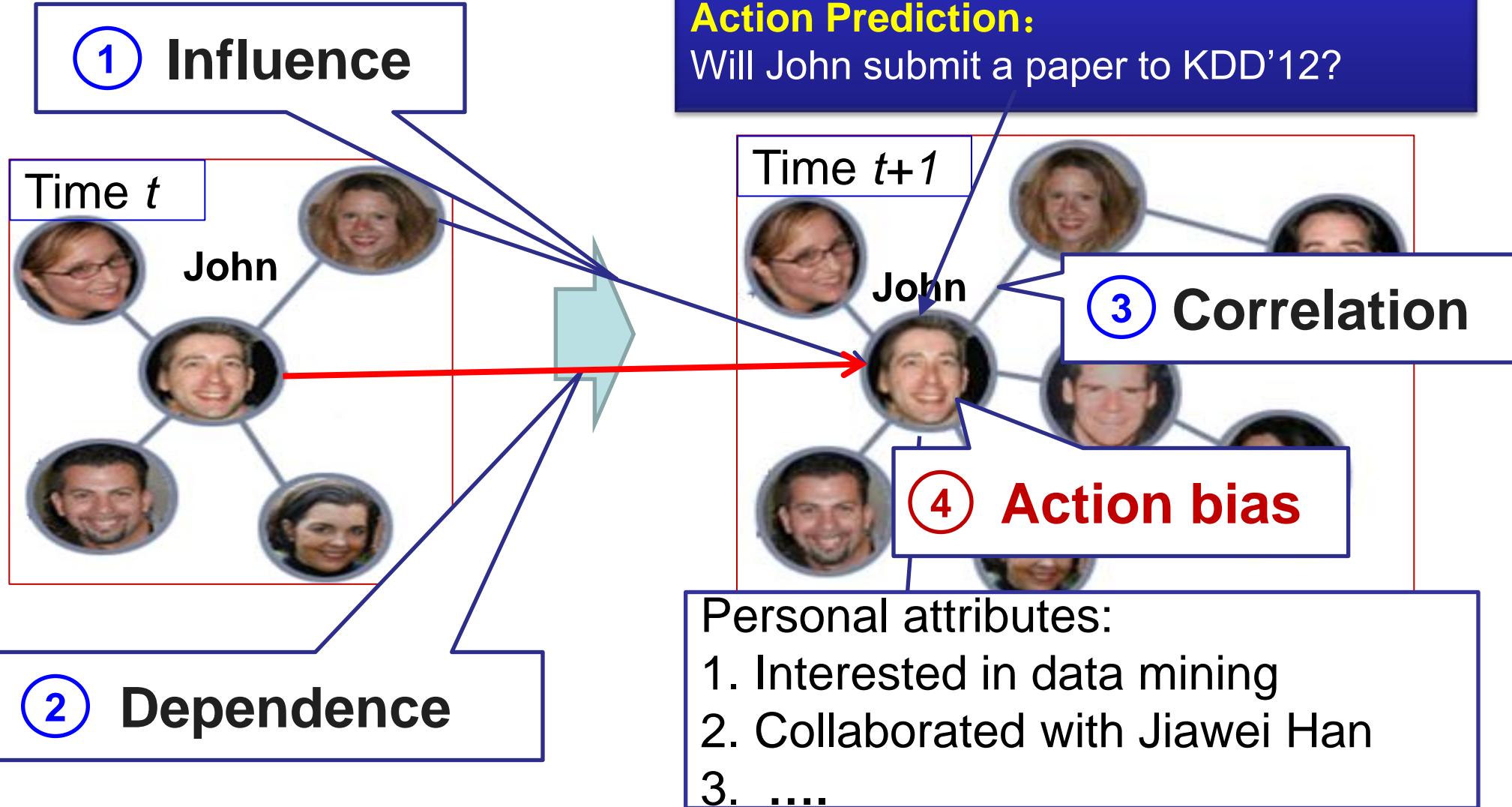
(KDD'09, '10, '11, DMKD'12, JIS)



- **User modeling:** can we model and forecast users' behaviors?
- **Influence:** how to quantify the influence between users?
- **Tie:** what is the relationship between users?
- **Community:** which (core) communities determine the evolution of the network structure?



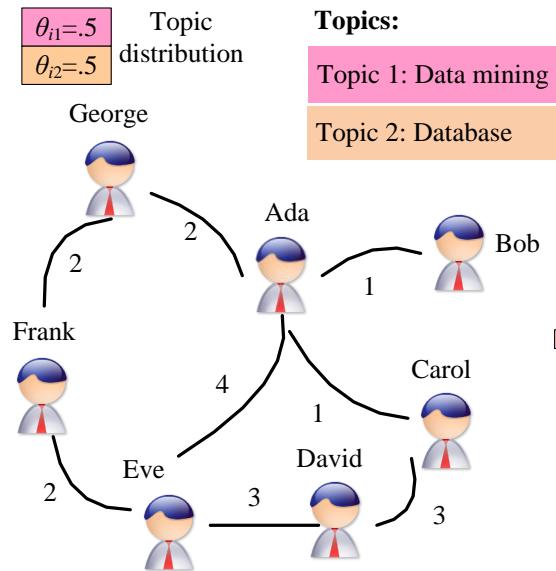
(1) Social Action Modeling and Prediction^[1]



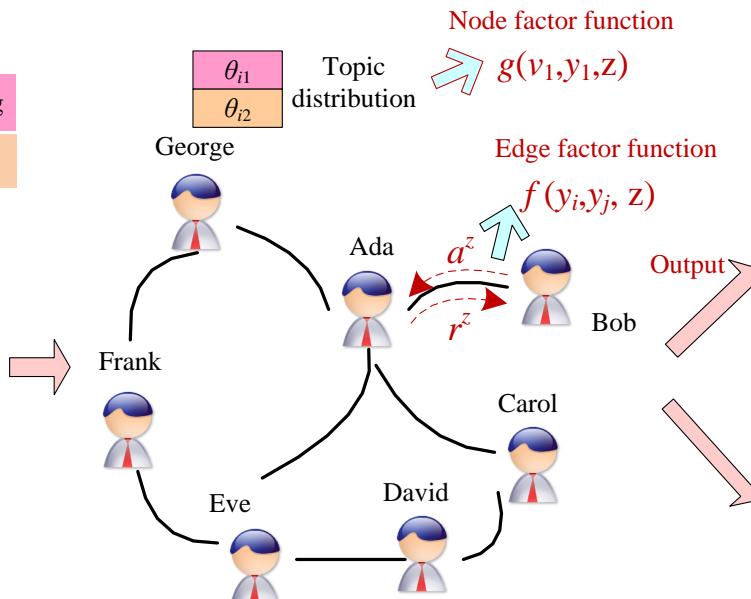
[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. SIGKDD'10, pages 1049-1058, 2010.

(2) Social Influence Analysis [1]

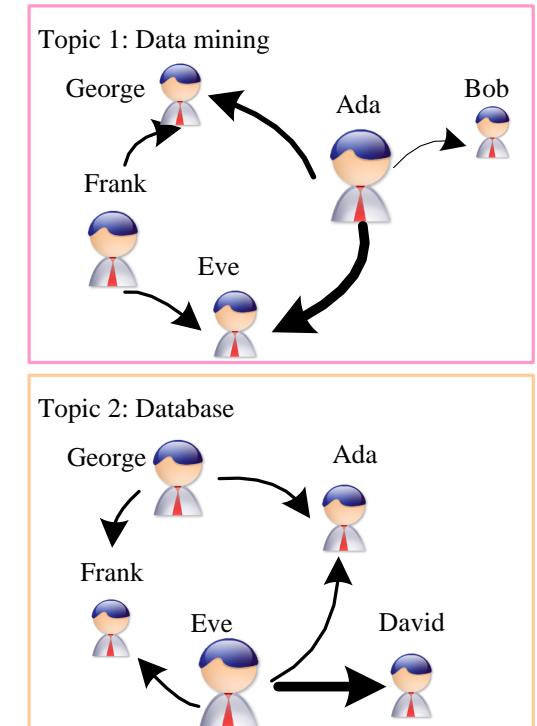
Input: coauthor network



Social influence analysis



Output: topic-based social influences



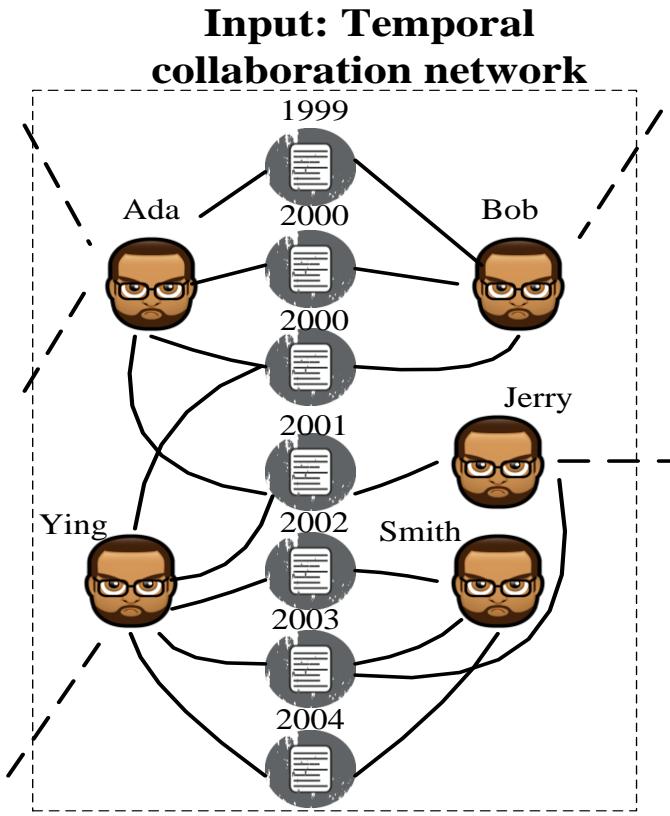
Several key challenges:

- How to differentiate the social influences from different topics?
- How to incorporate both structure and content into a unified model?
- How to estimate the model on real-large networks?

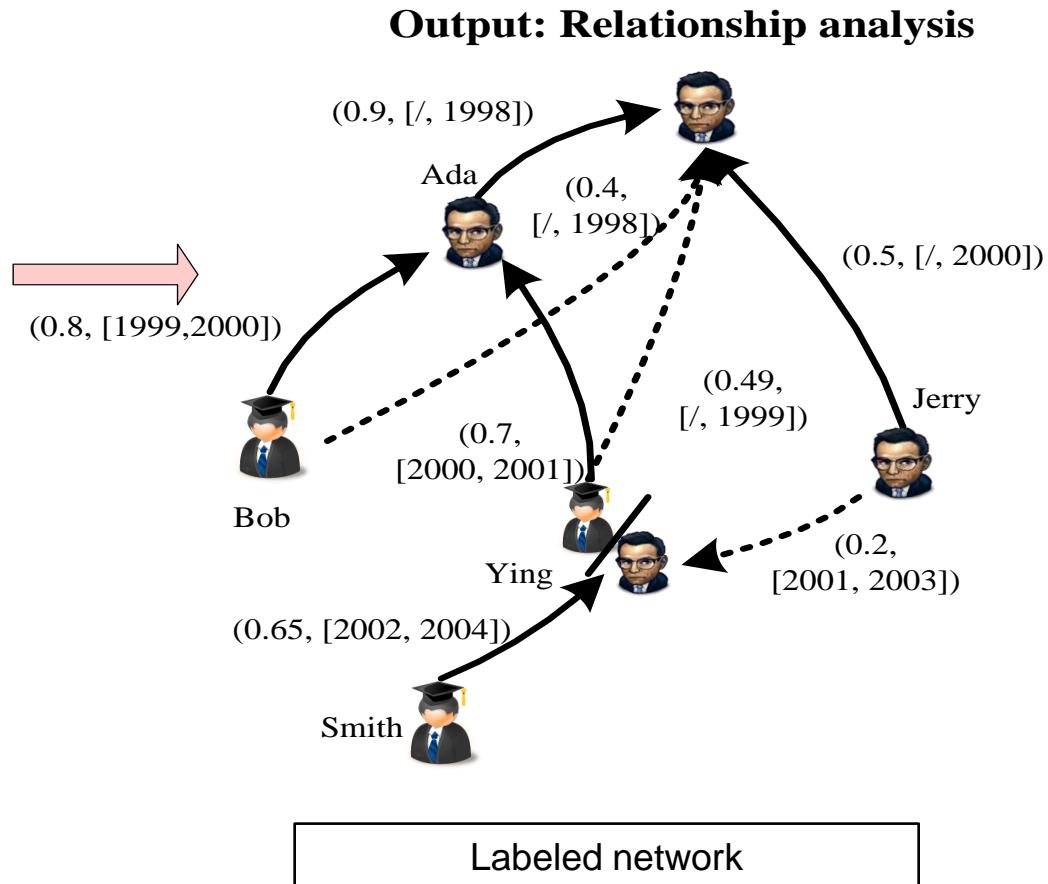
Influential nodes on different topics

Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell, C. Lee Giles, Terrence J. Sejnowski, Samy Bengio, Daphne Koller, Luc De Raedt, Vasant Honavar, Floriana Esposito, Bernhard Scholkopf
	Database System	Gerhard Weikum, John Mylopoulos, Michael Stonebraker, Barbara Pernici, Philip S. Yu, Sharad Mehrotra, Wei Sun, V. S. Subrahmanian, Alejandro P. Buchmann, Kian-Lee Tan, Jiawei Han
	Information Retrieval	Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Frieder, Alan F. Smeaton, Rong Jin
	Web Services	Yan Wang, Liang-jie Zhang, Schahram Dustdar, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah
	Semantic Web	Wolfgang Nejdl, Daniel Schwabe, Steffen Staab, Mark A. Musen, Andrew Tomkins, Juliana Freire, Carole A. Goble, James A. Hendler, Rudi Studer, Enrico Motta
	Bayesian Network	Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets
Citation	Data Mining	Fast Algorithms for Mining Association Rules in Large Databases, Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Discovery of Multiple-Level Association Rules from Large Databases, Interleaving a Join Sequence with Semijoins in Distributed Query Processing
	Machine Learning	Object Recognition with Gradient-Based Learning, Correctness of Local Probability Propagation in Graphical Models with Loops, A Learning Theorem for Networks at Detailed Stochastic Equilibrium, The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, A Unifying Review of Linear Gaussian Models
	Database System	Mediators in the Architecture of Future Information Systems, Database Techniques for the World-Wide Web: A Survey, The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles, Fast Algorithms for Mining Association Rules in Large Databases
	Web Services	The Web Service Modeling Framework WSMF, Interval Timed Coloured Petri Nets and their Analysis, The design and implementation of real-time schedulers in RED-linux, The Self-Serv Environment for Web Services Composition
	Web Mining	Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Fast Algorithms for Mining Association Rules in Large Databases, The OO-Binary Relationship Model: A Truly Object Oriented Conceptual Model, Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations, Improving Fault Tolerance and Supporting Partial Writes in Structured Coterie Protocols for Replicated Objects
	Semantic Web	FaCT and iFaCT, The GRAIL concept modelling language for medical terminology, Semantic Integration of Semistructured and Structured Data Sources, Description of the RACER System and its Applications, DL-Lite: Practical Reasoning for Rich DLs

(3) Mining Advisor-Advisee Relationship^[1]



Dynamic collaborative network



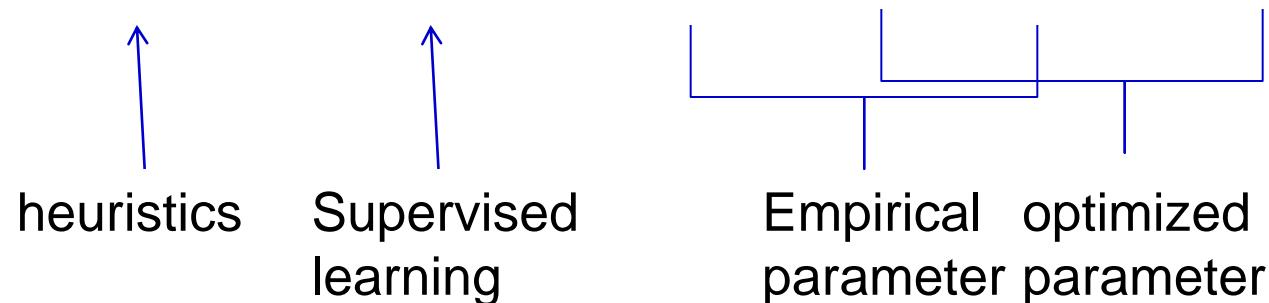
Labeled network

Output: potential types of relationships and their probabilities:
 $(\text{type}, \text{prob}, [\text{s_time}, \text{e_time}])$

Results

- DBLP data: 654, 628 authors, 1,076,946 publications, years provided.
- Ground truth: MathGenealogy Project; AI Genealogy Project; Faculty Homepage

Datasets	RULE	SVM	IndMAX		Model 1	
TEST1	69.9%	73.4%	75.2%	78.9%	80.2%	84.4%
TEST2	69.8%	74.6%	74.6%	79.0%	81.5%	84.3%
TEST3	80.6%	86.7%	83.1%	90.9%	88.8%	91.3%



Results

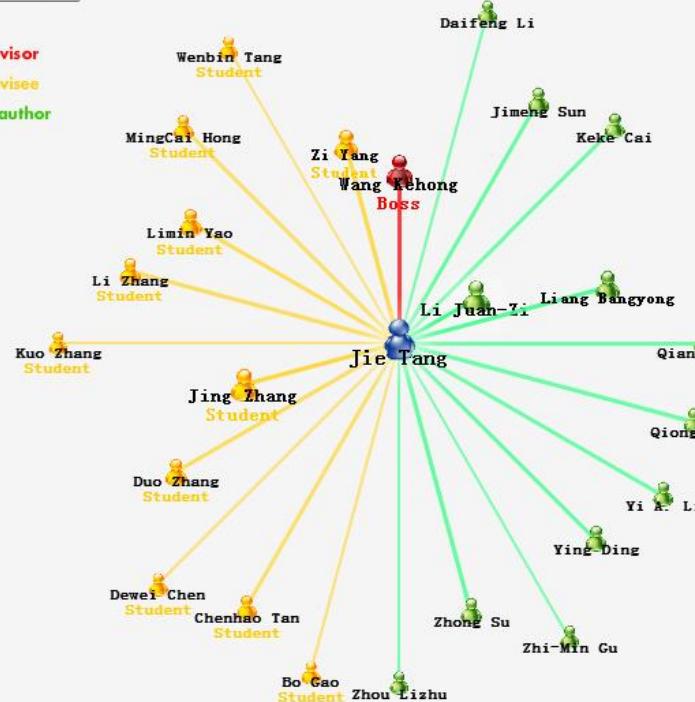
Social Graph

Colour Network
B&W Network
Ability

Relation: All



advisor
advisee
coauthor



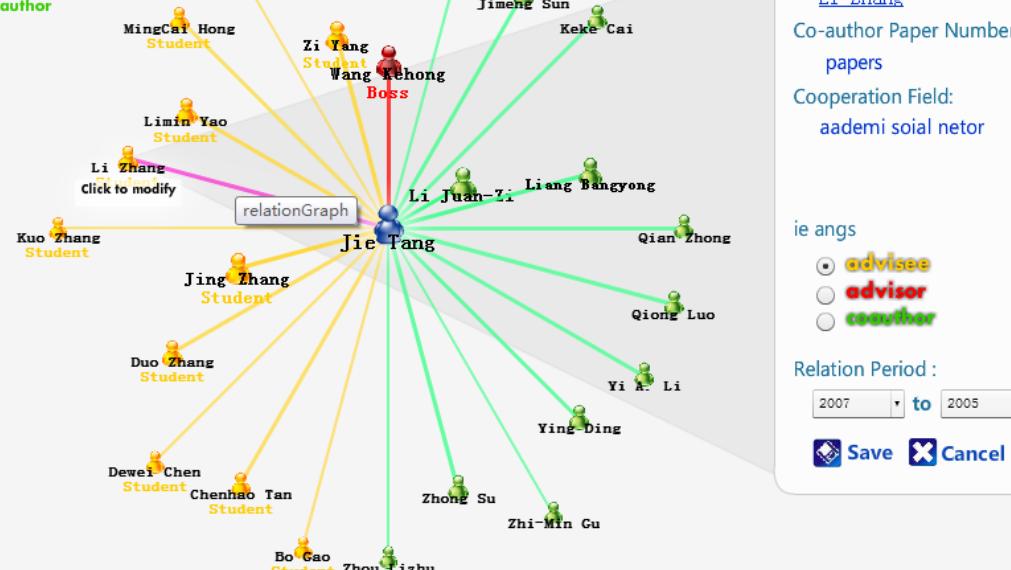
Social Graph

Colour Network
B&W Network
Ability

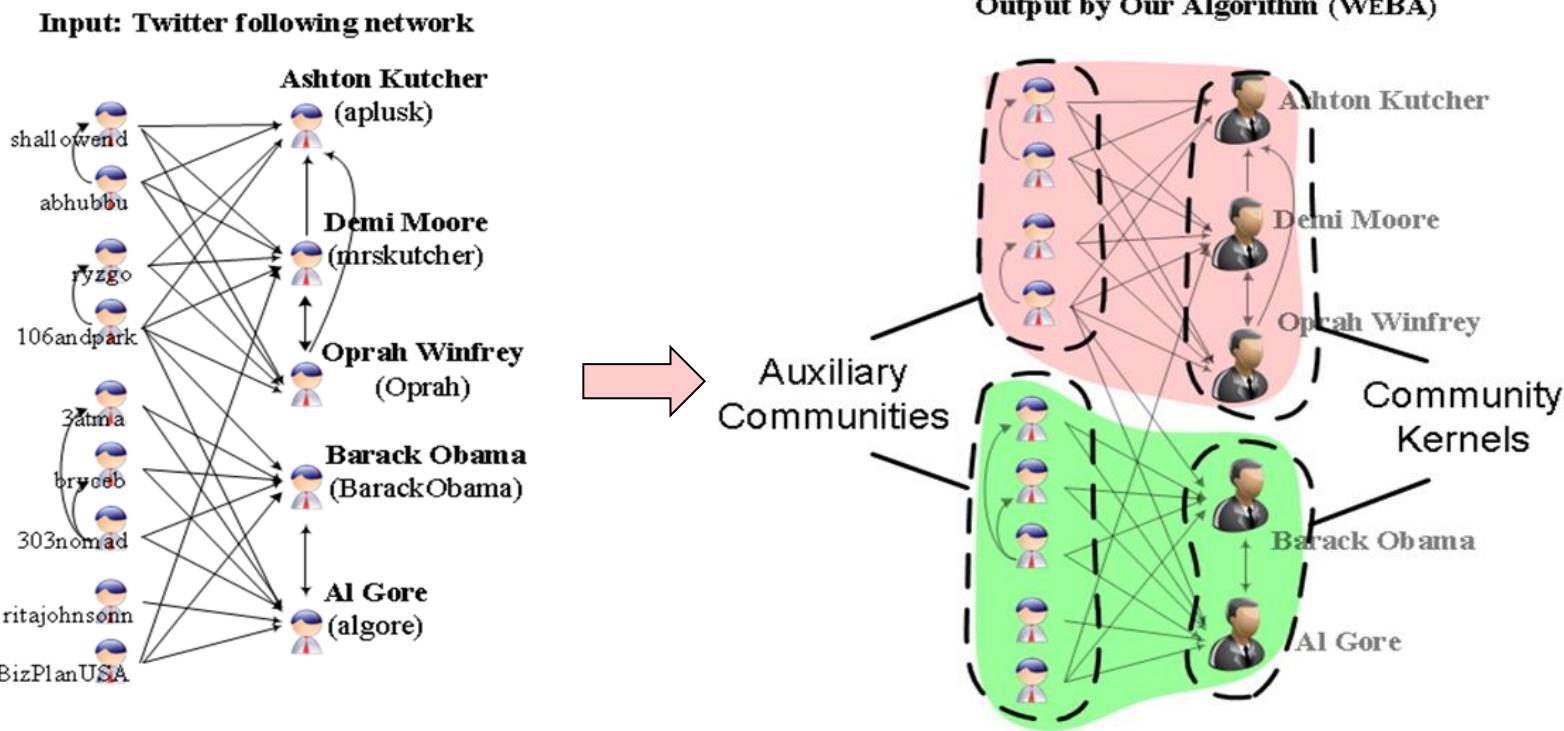
Relation: All



advisor
advisee
coauthor



(4) Community Kernel[1]



Pareto Principle: Less than 1% of the Twitter users (e.g. entertainers, politicians, writers) produce 50% of its content, while the others (e.g. fans, followers, readers) have much less influence and completely different social behavior.

Kernel users and ordinary users exhibit very different behaviors.

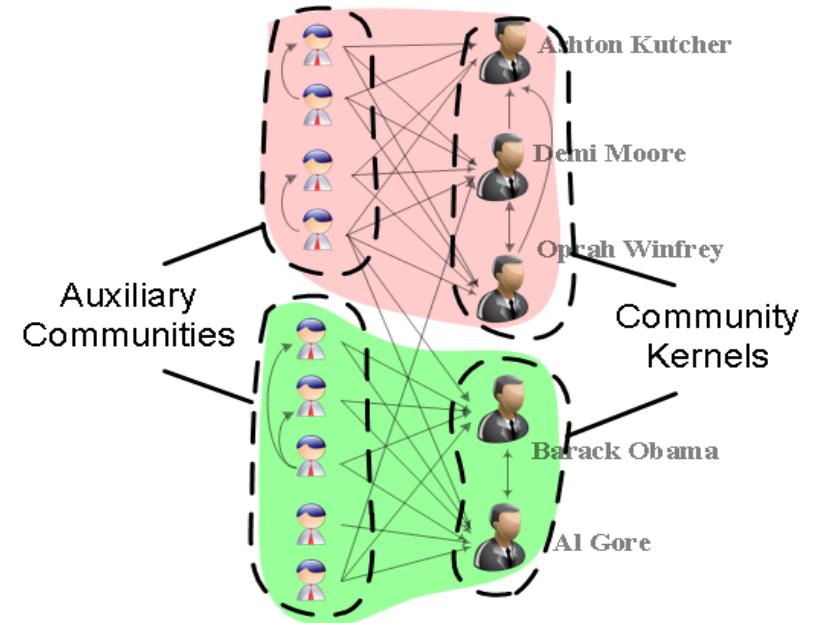
Approach (Greedy & WeBA)

Challenges

- How to identify kernel members, and
- How to determine the structural of community kernels.

Formalize the problem into an optimization problem.

- Proposed two new algorithms
- Greedy
 - Maximum cardinality search(MCS)
 - Runs in linear time
- Global Relaxation
 - Random walk(Annealing)
 - Theoretical error bound
 - Almost linear time.



Results on Coauthor & Wikipedia

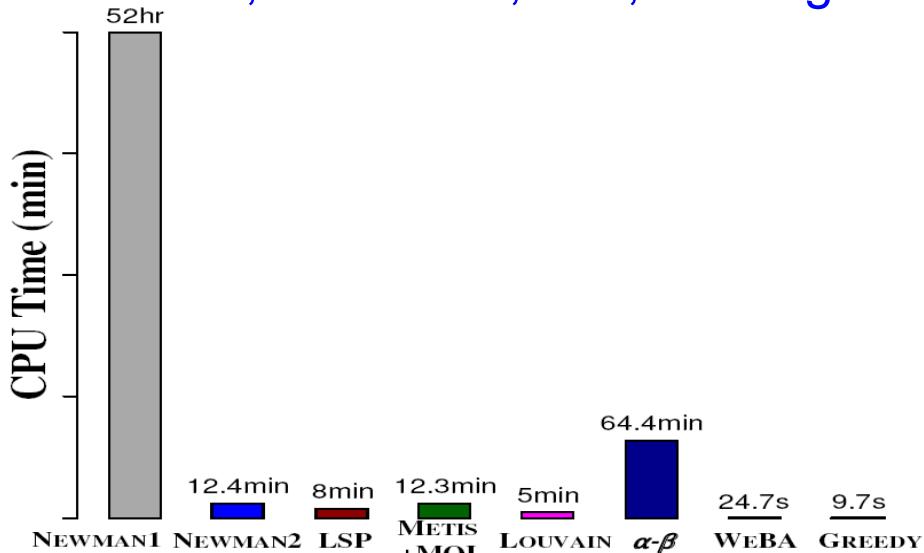


- On average, WEBA improves Precision by 340% (wiki) and 70% (coauthor), and improves Recall by 130% (wiki) and 41% (coauthor).

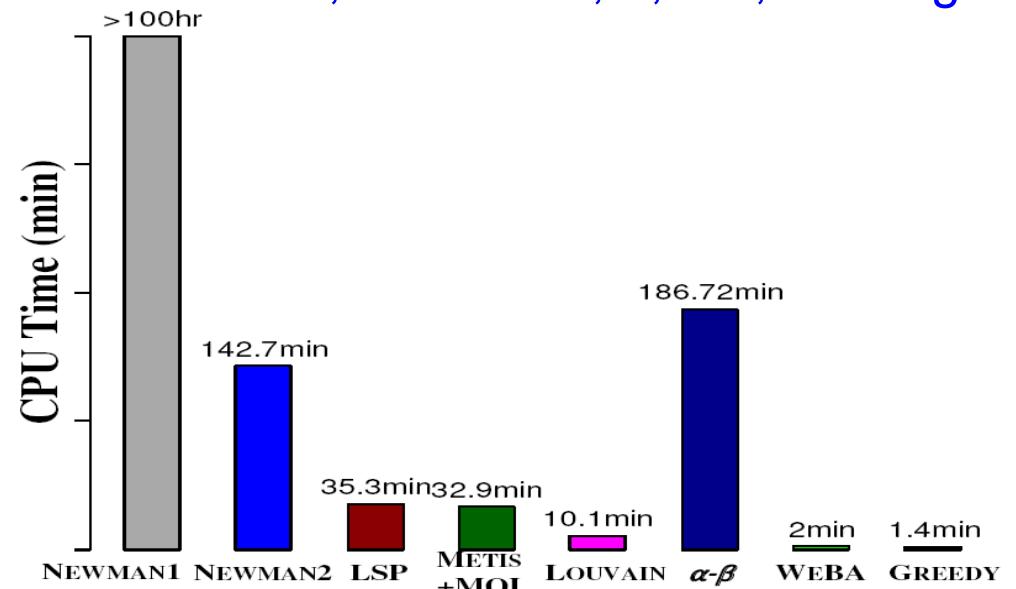
	Precision						Recall					
	wiki		coauthor				wiki		coauthor			
	Talk	User	AI	...	NC	Average	Talk	User	AI	...	NC	Average
LSP	0.061	0.085	0.502	...	0.342	0.573	0.171	0.315	0.458	...	0.398	0.561
d-LSP	0.051	0.091	0.528	...	0.504	0.617	0.427	0.273	0.519	...	0.463	0.609
p-LSP	0.046	0.082	0.678	...	0.403	0.641	0.442	0.237	0.337	...	0.491	0.574
METIS+MQI	0.049	0.012	0.847	...	0.055	0.488	0.062	0.361	0.089	...	0.077	0.379
LOUVAIN	0.063	0.122	0.216	...	0.272	0.437	0.388	0.348	0.184	...	0.19	0.343
NEWMAN1	0.033	0.203	0.4	...	0.259	0.431	0.099	0.077	0.306	...	0.174	0.311
NEWMAN2	0.039	0.085	0.298	...	0.613	0.463	0.029	0.075	0.364	...	0.467	0.335
$\alpha\text{-}\beta$	0.324	0.336	0.443	...	0.747	0.626	0.422	0.427	0.602	...	0.568	0.654
WEBA	0.456	0.46	0.852	...	0.837	0.911	0.589	0.57	0.577	...	0.582	0.664
GREEDY	0.334	0.403	0.83	...	0.746	0.752	0.432	0.499	0.545	...	0.56	0.659

EFFICIENCY — TWITTER & COAUTHOR

465,023 nodes, 833,590 edges



822,415 nodes, 2,928,360 edges





Arnetminer Today

— A brief summary



ArnetMiner's History

Date	Version	New Features
2006/5	V0.1	Profile extraction, person/paper/conf. search
2006/8	V1.0	Rewritten all codes in Java.
2007/7	V2.0	Survey search, research interest, association search
2008/4	V3.0	Query understanding, New search GUI, log analysis
2008/11	V4.0	Graph search, topic mining, NSFC/NSF
2009/4	V5.0	Bole/course search, profile editing, open resources, #citation
2009/12	V6.0	Academic statistics, user feedbacks, refined ranking
2010/5	V7.0	Name disambiguation, reviewer assignment, supervisor suggestion, open API
2010/7	V8.0	ArnetApp Platform
2011/7	V II	AMiner, location search, conference analysis
2012	V 1.0	New UI, cross-domain collaboration



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.

SciVerse Hub BETA

Home | My settings

data mining

About 54317 results for ALL(data mining)

Search within results

Refine Results

Limit to Exclude

Content Sources

Scien Scopu Paten Digi MD Cr

View more

Year

2010 Call for Papers Author/Reviewer Info Papers Posters Workshops Tutorials Panels Demos Exhibits KDD Cup Industry Practice Expo Awards

2009

2008

2007

2006

View more

Keyword

data r cluste class neura algor

Industry Practice

Expo Awards

Attend Register Travel Dates Organizers (General) PC (Research) PC (Industry) Request Poster

Conference Organizers

General Chair gc@kdd2011.com

Chidanand Apte IBM Research Division

Program Co-Chairs program-chairs@kdd2011.com

Joydeep Ghosh Electrical and Computer Engineering, University of Texas, Austin

Padhraic Smyth School of Information and Computer Sciences, Center for Machine Learning and Intelligent Systems, University of California, Irvine

Industry Track Co-Chairs industrial@kdd2011.com

Ted E. Senator SAIC

Michael Zeller Zementis

Industry Practice Expo Co-Chairs industrial_practice@kdd2011.com

Ying Li adCenter Labs, Microsoft Corporation

Rajesh Parekh Groupon

ArnetMiner is logged in | Go to SciVal Suite

Home | About | Committee | Authors | Attendees | Program | Sponsors

Fourth ACM International Conference on Web Search and Data Mining

9-12 February 2011 Hong Kong

QUICK LINKS

- Instructions for Authors
- Registration
- Accommodations
- Travel Grants Information
- Tutorial Information
- Workshop Information

IMPORTANT DATES

11/15	Camera-ready deadline New
11/11	Google travel grant deadline
11/22	Author registration deadline
12/15	Early-bird registration deadline
12/15	Google travel grant notification
12/20	Student travel grant deadline
1/9	Group rate hotel reservation deadline
1/10	Student travel grant notification

SPONSORS

Platinum Sponsors

Microsoft Research

Gold Sponsors

Google

Technical Sponsors

ArnetMiner

The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
KDD-2011
San Diego, CA • August 21-24, 2011

The Chinese University of Hong Kong, Hong Kong

data mining

Search

Do you mean: [Mark Mine](#), [Ray Mines](#), [Data Becker](#)

Reset APPs

Experts Location Distribution for "data mining"

Full Page

地图 卫星



Conference/Journals

0 - 10 of 1373 conf/journals (0 seconds)

KDD

SIGKDD Explorations

Datenverarbeitungspraxis

ICDM

Expert Syst. Appl.

PAKDD

ICDM Workshops

CoRR

Data Min. Knowl. Discov.

PKDD

Prev 1 / 138 Next

Publications

0 - 10 of 15651 publications (0 seconds)

Order by: RELEVANCE YEAR CITATION

Data Mining: Concepts and Techniques

Authors: Jiawei Han, Micheline Kamber.

Published Year: 2000

CitedBy 13179 PDF BIBTEX

Mining Association Rules between Sets of Items in Large Databases.

Authors: Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami.

JConf: SIGMOD Conference

Published Year: 1993

CitedBy 11393 PDF BIBTEX

From Data Mining to Knowledge Discovery in Databases.

Authors: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth.

JConf: AI Magazine

Published Year: 1996

CitedBy 4332 PDF BIBTEX

Introduction to Data Mining

Authors: Pang-Ning Tan, Michael Steinbach, Vipin Kumar.

Published Year: 2005

CitedBy 2963 PDF BIBTEX

Mining Frequent Patterns without Candidate Generation.

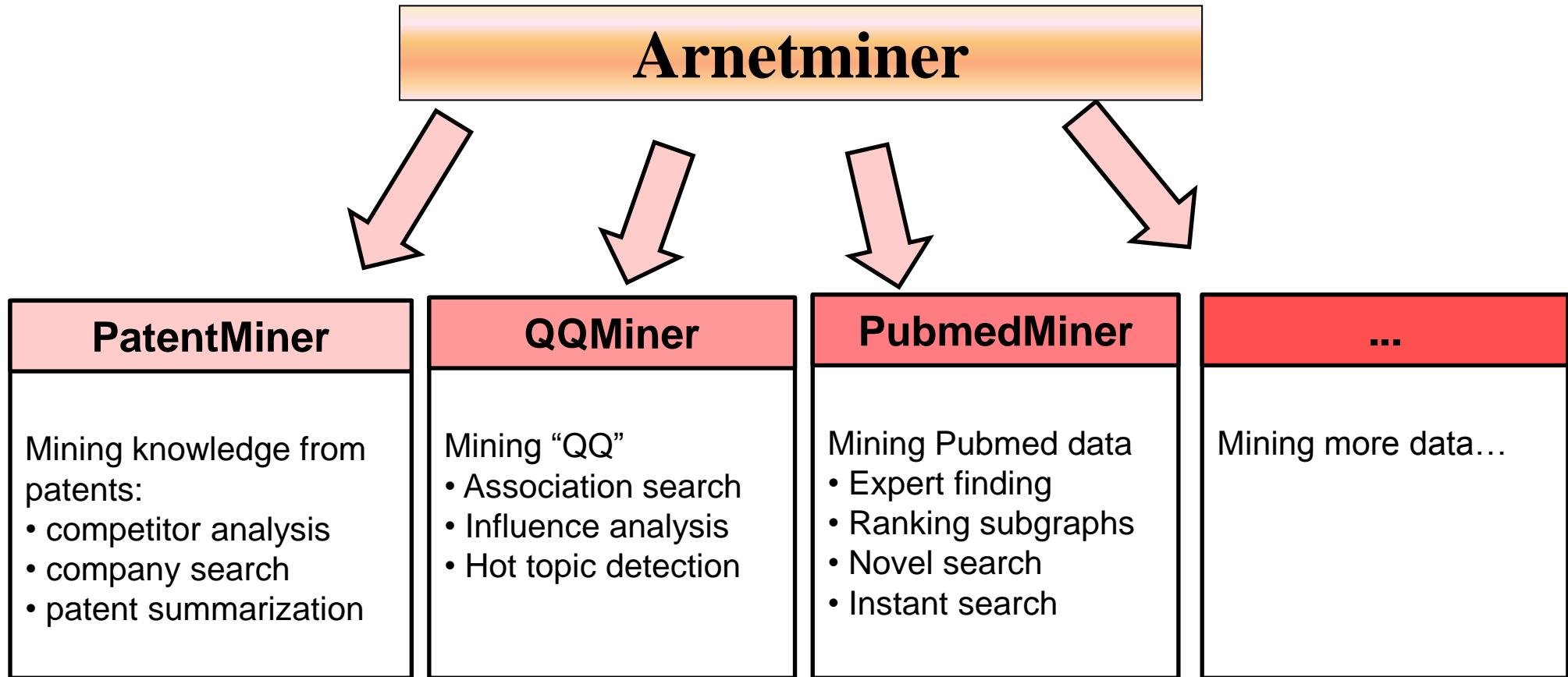
Authors: Jiawei Han, Jian Pei, Yiwen Yin.

ArnetApp Platform

---to deploy your apps on Arnetminer.org

To customize the search on Arnetminer.

Arnetminer as a platform...





PatentMiner



Patent Search

PatentMiner

Summary of "data mining":

Result Categories (U.S. Class)

Tips: left click to focus, right click to return.

All results

Close

Mining or in situ disintegration of hard material

Hard material disintegrating machines

Processes

Cutter tooth or
tooth head

Automatic control;
signaling or indicating

Hydraulic and earth engineering

Earth treatment or control

In situ conversion of
solid to fluid

Related sub
topics

Patents on
“data mining”

Found for "data mining"

relevance date authority

Found 3211 patents, used time 0.477 seconds

Extensible data mining framework

Inventors: Raman S. Iyer, Ioan Bogdan Crivat, C. James MacLennan, Scott C. Oveson, Rong J. Guan, ZhaoHui Tang, Pyungchul Kim, Irina G. Gorbach

Company: Microsoft Corporation

Issued Date: 2008-06-03

The subject disclosure pertains to extensible data mining systems, means, and methodologies. For example, a data mining system is disclosed that supports plug-in or integration of non-native mining algorithms, perhaps provided by third parties, such ...

Patient data mining

Inventors: R. Bharat Rao, Sathyakama Sandilya, Christopher Jude Amies, Radu Stefan Niculescu, Arun Kumar Goel, Thomas R. Warrick

Company: Siemens Medical Solutions USA, Inc.

Issued Date: 2009-11-10

The present invention provides a data mining framework for mining high-quality structured clinical information. The data mining framework includes a data miner that mines medical information from a computerized patient record (CPR) based on domain-sp ...

Clustering module for data mining

Inventors: Marcos M. Campos

Company: Oracle International Corporation

Issued Date: 2006-08-15

A system, software module, and computer program product for performing clustering based data mining that improved performance in model building, good integration with the various databases throughout the enterprise, flexible specification and adjustm ...

GUI guide for data mining

Inventors: Robert E. Medi

Company: International Business Machines Corporation

Issued Date: 2000-08-22

Inventors # query related / all

William R. Kennedy	# patents: 38/46
John M. Kennedy	# patents: 37/43
Gerhard Merten	# patents: 23/64
Walter Weirich	# patents: 21/49
Rakesh Agrawal	# patents: 21/171
John C. Stankus	# patents: 19/51
Kunibert Becker	# patents: 17/22
Raymond L. Wright	# patents: 17/20
Maurice K. LeBegue	# patents: 16/30
James J. Fallon	# patents: 14/24

Prev 1 / 451 Next

Inventor

Companies # query related / all

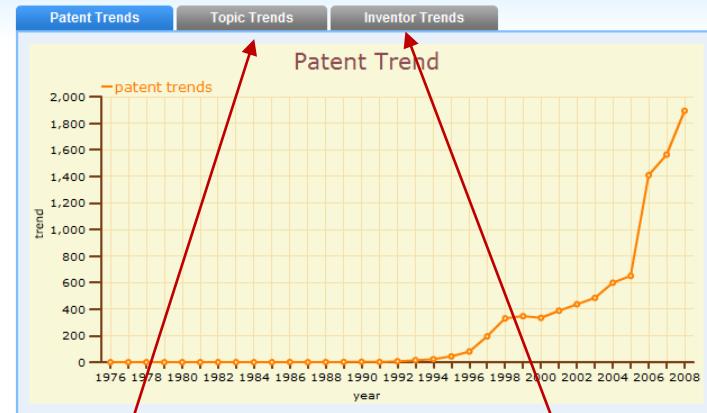
International Business Machines Corporation	# patents: 152/60180
Gewerkschaft Eisenhütte Westfalia	# patents: 147/349
The United States of America as represented by the Secretary of the Navy	# patents: 140/23714
Coal Industry	# patents: 59/283
Gebr. Eickhoff, Maschinenfabrik und Eisengiesserei m.b.H.	# patents: 54/76
Minnesota Mining and Manufacturing Company	# patents: 52/12395

Company



Microsoft Corporation

Trend Analysis:



Company Name:
Microsoft Corporation
Patent Count: 13885
Website:
microsoft.com



Competitors:

Global Evolution Topic Topic Evolution

Heath Company
patents: 34

General Motors Corporation
patents: 10384

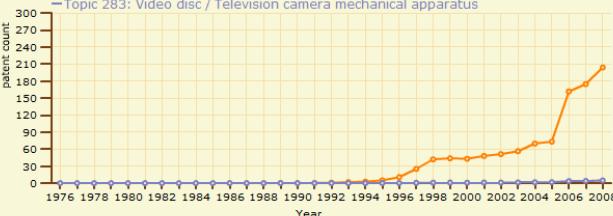
Patent Trends

Topic Trends

Inventor Trends

Patent Trend

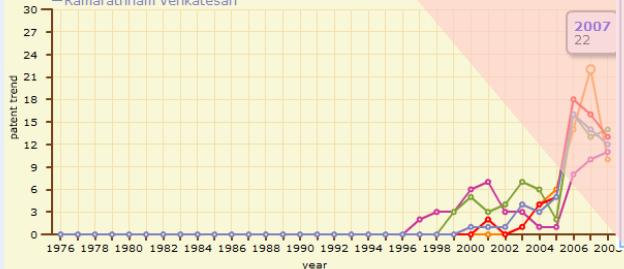
- Topic 220: Database management system / Layering cache
- Topic 72: Computer system / Web site
- Topic 71: Manufacturing method / Optical information recording medium
- Topic 53: Rotary bearing assembly / Polymerization catalyst
- Topic 283: Video disc / Television camera mechanical apparatus



Patent Trends

Top 5 Inventor Trend

- Hong Jiang Zhang
- John R Doucet
- Heung Yeung Shum
- Paul England
- Ramarathnam Venkatesan



Competitors:

Global Evolution Topic Topic Evolution

Find query-related competitor evolution for "Microsoft Corporation"

mobile phone

Research In Motion Limited

patents: 887



Silverbrook Research Pty Ltd

patents: 2831



1996–2000

Nokia Corporation

patents: 3869



International Business Machines Corporation

patents: 56920



Hewlett-Packard Development Company, L.P.

patents: 11035



Research In Motion Limited

patents: 887



Silverbrook Research Pty Ltd

patents: 2831



2001–2005

Nokia Corporation

patents: 3869



International Business Machines Corporation

patents: 56920



Hewlett-Packard Development Company, L.P.

patents: 11035



Nokia Mobile Phones Ltd.

patents: 663



NEC Corporation

patents: 21655



Samsung Electronics Co., Ltd.

patents: 24091



Sony Ericsson Mobile Communications AB

patents: 423



International Business Machines Corporation

patents: 56920



PatentMiner Today

* Patent data:

- > 3.8M patents
- > 2.4M inventors
- > 400K companies
- > 10M citation relationships

* Journal data:

- > 2k journal papers
- > 3.7k authors

The crawled data is increasing to
>300 Gigabytes.

Top Browser | My reviews | Welcome! guest | Sign Out

PatentMiner

Summary of "data mining":
1 - 10 of 6313 patents for 'data mining' (0.5160 seconds)

Patent Found for "data mining"

Parallel object-oriented decision tree system
Inventors: Chandrika Kamath, Eric Corrado, et al.
Company: The Regents of the University of California
Issued Date: 2009-02-28
[0 reviews]

A decision tree system that uncovers patterns, associations, anomalies, and other statistically significant structures in data by reading and displaying data files, extracting methods.

Microsoft
Microsoft Corporation
Founded: Albuquerque, New Mexico April 4, 1975
Industry: Computer software Consumer electronics Digital distribution Computer hardware Video games IT consulting Online advertising Retail stores Automotive three M's...

Visual presentation technique for data mining system
Inventors: Yuchun Lee, Ruby Kennedy
Company: Unica Technologies, Inc.
Issued Date: 2001-07-31
[0 reviews]

A method for presenting measurements of modeling performance. The method displays a lift chart on an output display device. The three M's...

Method and apparatus for providing and expressing
Inventors: David J. Balasian, Elina Khurjan
Company: Affinexus, Inc.
Issued Date: 2001-02-06
[0 reviews]

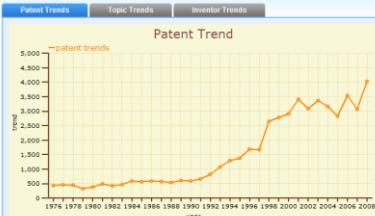
According to the invention, a system and method for organizing mining. A database model is provided which may organize information analysis of expression.

Unstructured data in a mining model language

International Business Machines Corporation

Trend Analysis:
Patent Trends **Topic Trends** **Inventor Trends**

Patent Trend



Year

Patent Count: 59669

Website: <http://ibm.com>

[0 reviews]

Competitors:

- Hitachi, Ltd. # patents: 36805
- Fujitsu Limited # patents: 23160
- Intel Corporation # patents: 18661
- Hewlett-Packard Development Company, L.P. # patents: 21012
- Motorola, Inc. # patents: 18452
- Sun Microsystems, Inc. # patents: 7291
- Microsoft Corporation

Hot Words

- computer (0.0110)
- storage (0.0052)
- management (0.0048)
- interface (0.0043)
- access (0.0040)
- program (0.0079)
- multiple (0.0049)
- user (0.0047)
- memory (0.0042)
- providing (0.0038)

Topic 220: Database management system / Layering cache

Topic Trends



Year

Company Trend

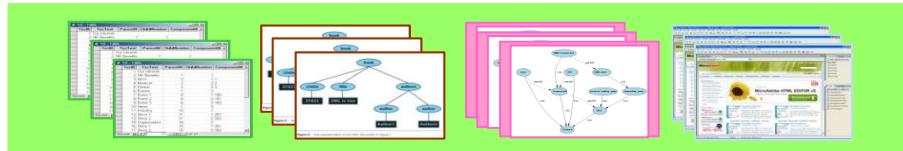
Top 5 Company Trend



Year



Opportunity: exploiting social network and semantic web in the real-world



Web, relational data,
ontological data,
social data

Data Mining and Social Network techniques

Scientific Literature

Users cover >180 countries
>600K researcher
>3M papers

Arnetminer.org
(NSFC, 863)

Social search & mining

Social extraction
Social mining

IBM US, Tencent
IBM CRL

Advertisement

Advertisement
Recommendation

Sohu

Mobile Context

Mobile search & recommendation

Nokia

Energy trend analysis

Energy product
Evolution
Techniques
Trend

Oil Company

Large-scale Mining

Scalable algorithms
for message tagging
and community
Discovery

Google

国家核高基项目 (NSFC)

自然科学基金重点 (NSFC Key)

科技信息资源内容监测与分析服务平台 (中国科技部信息情报研究所)

Search, browsing, complex query, integration, collaboration, trustable analysis, decision support, intelligent services,



Representative Publications

- Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. **Machine Learning Journal**, 2011.
- Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM **TKDD**, 2010.
- Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE **TKDE**, 2012.
- Juanzi Li, Jie Tang, Yi Li, Qiong Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE **TKDE**, 2009. (**Top 6** cited papers among TKDE 2009's papers)
- Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. Using Bayesian Decision for Ontology Mapping. JoWS, 2006. (if =3.41) (**Top 10** cited papers in JWS's history)
- Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. Social Context Summarization. **SIGIR'11**.
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. **KDD'12** (Full Presentation & Best Poster Award)
- Jie Tang, Bo Wang, Yang Yang, Po Hu, Yanting Zhao, Xinyu Yan, Bo Gao, Minlie Huang, Peng Xu, Weichang Li, and Adam K. Usadi. PatentMiner: Topic-driven Patent Analysis and Mining. **KDD'12**.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. **KDD'11**.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. **KDD'10**.
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. **KDD'09**. (**Top 4** cited papers among KDD 2009's papers)
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. **KDD'08**. (**Top 6** cited papers among KDD 2008's papers)
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. **IJCAI'09**.



Thanks!

Demo: <http://arxivminer.org>

HP: <http://keg.cs.tsinghua.edu.cn/jietang/>