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 analystic
- 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 information need is not only about publication…"

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

Examples – Expertise search

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

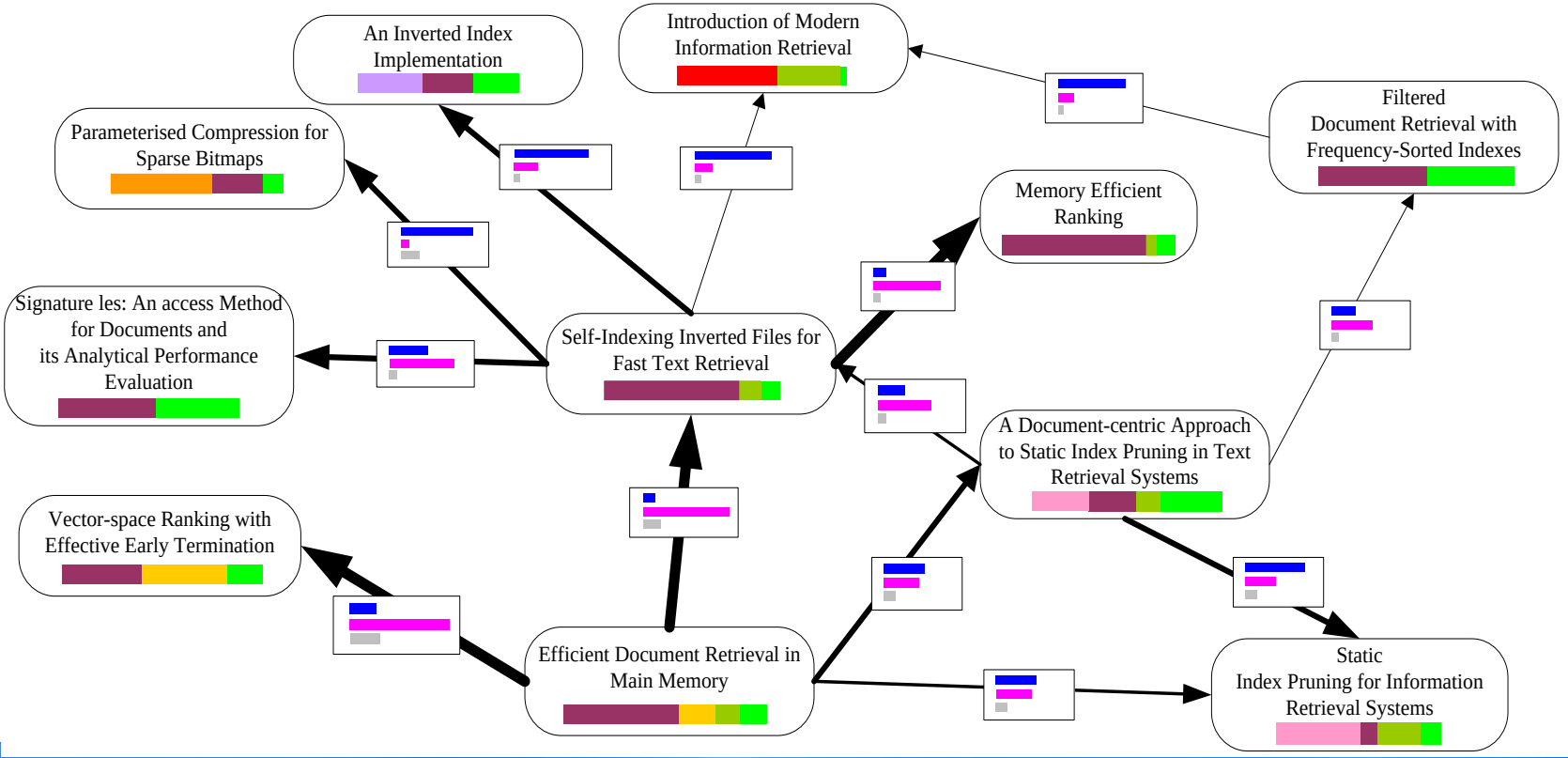
Researcher A

- **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

• an in-depth understanding of the research field?



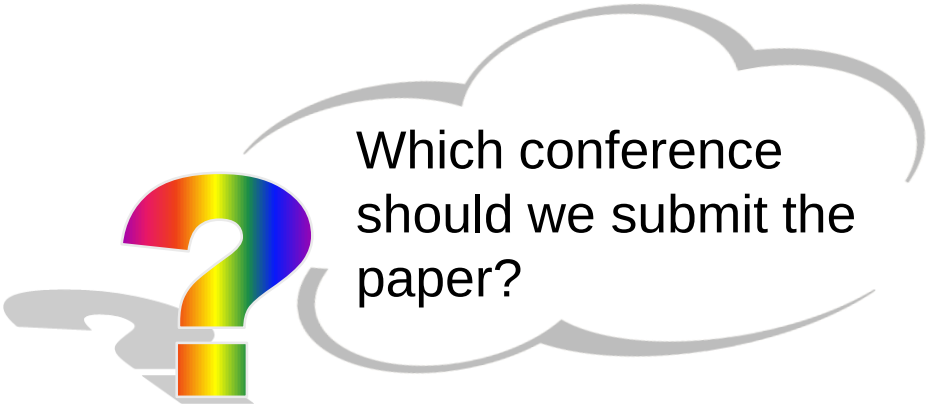
Topics

- Topic 31: Ranking and Inverted Index
- Topic 1 : Theory
- Topic 27: Information retrieval
- Topic 23: Index method
- Topic 21: Framework
- Topic 34: Parallel computing
- Topic 22: Compression
- Other

Citation Relationship Type

- Basic theory
- Comparable work
- Other

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 method have been proposed. Feature selection methods select a sub- set of words to reduce the dimensionality. Feature transfor- mation methods try to tackle not only the high dimensional- ity problem of "bag-of-words" representation, but indirectly consider synonymy and polysemy as well. Latent Semanti Indexing (LSI) [6] is one of these approaches which use singular value decomposition to identify a linear subspac in the original space of features. It is believed that the result- ing 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 in- terpretable components. Despite some work on interpreting the dimensions generated by these methods [5], these ap- proaches 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 proba- bilistic topics. Probabilistic topic modeling as a way of rep- resenting the content of words and documents has the dis- tinct advantage that each topic is individually interpretable providing a probability distribution over words that pick out a coherent cluster of correlated terms. The major dif- ference between cartesian space methods like LSI and sta- tistical 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 di- mensionality and ignoring polysemy and synonymy prop- erties, it still keeps the "bag-of-words" assumption intact. Recently, there has been increased research interest in mod- els sensitive to this kind of information [11].

content

Examples – Reviewer Suggestion

KDD Committee

Conference/Journal

Who are best matching reviewers for each paper?

Who will accept the invitation to review the paper

Paper content

Outline

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

AMiner.org

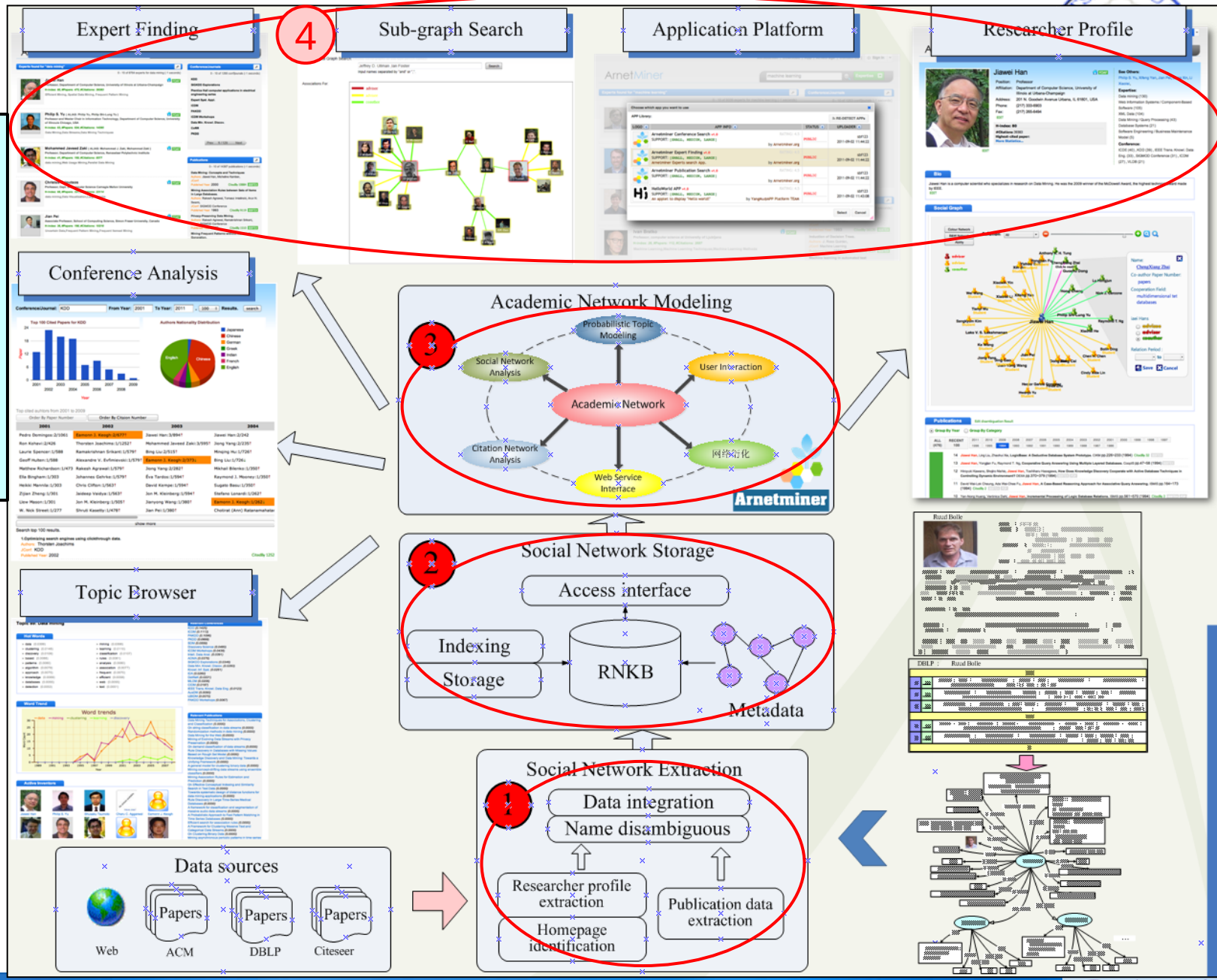
- Academic research social network analysis and mining system

提供全面的研究者网络分析与挖掘功能

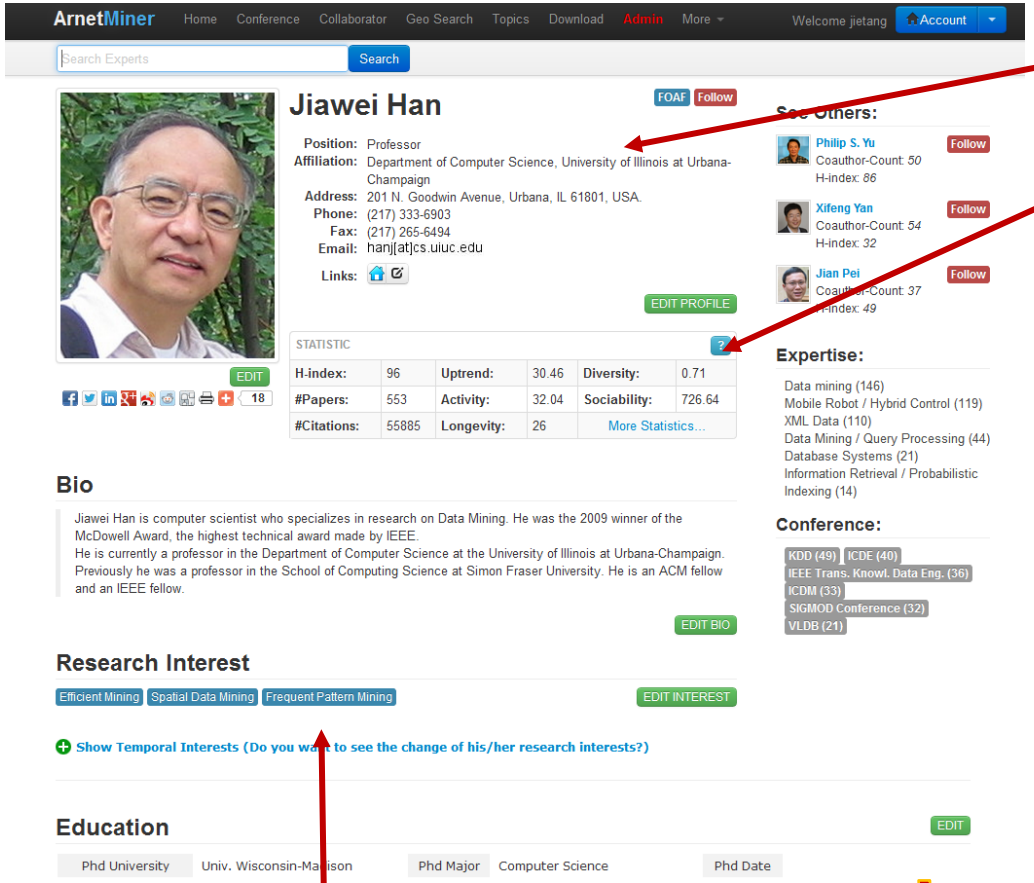
Papers published: ACM TKDD, IEEE TKDE, J. Informetrics, KDD'08-12, WWW'12, SIGMOD'09, IJCAI'09

http://aminer.org/

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



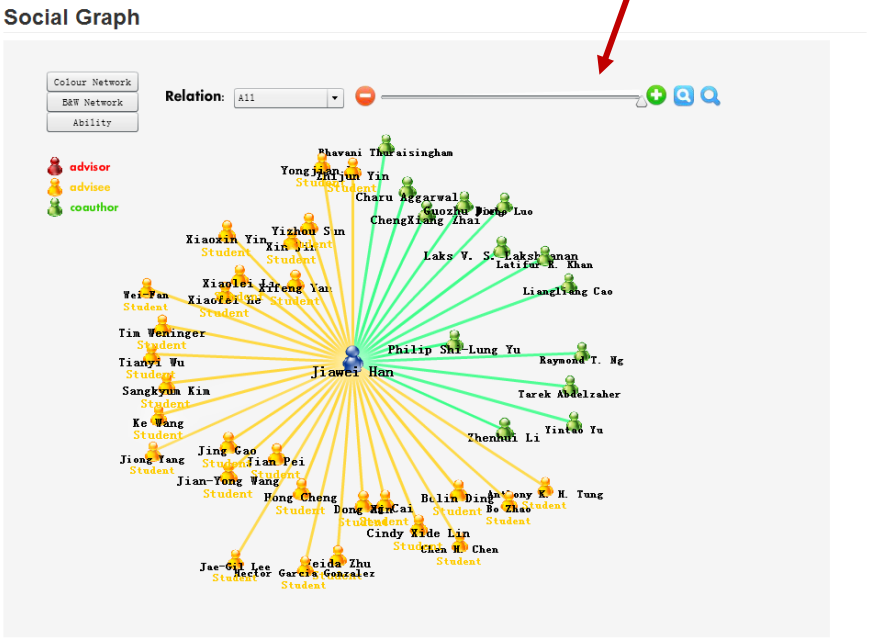
Person Search



Basic Info.

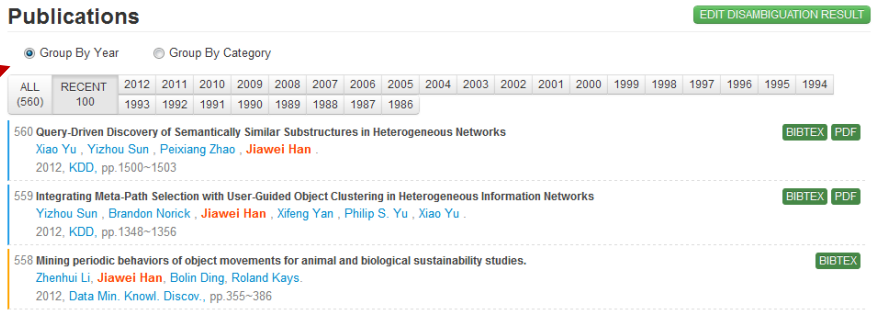
Citation statistics

Social Network



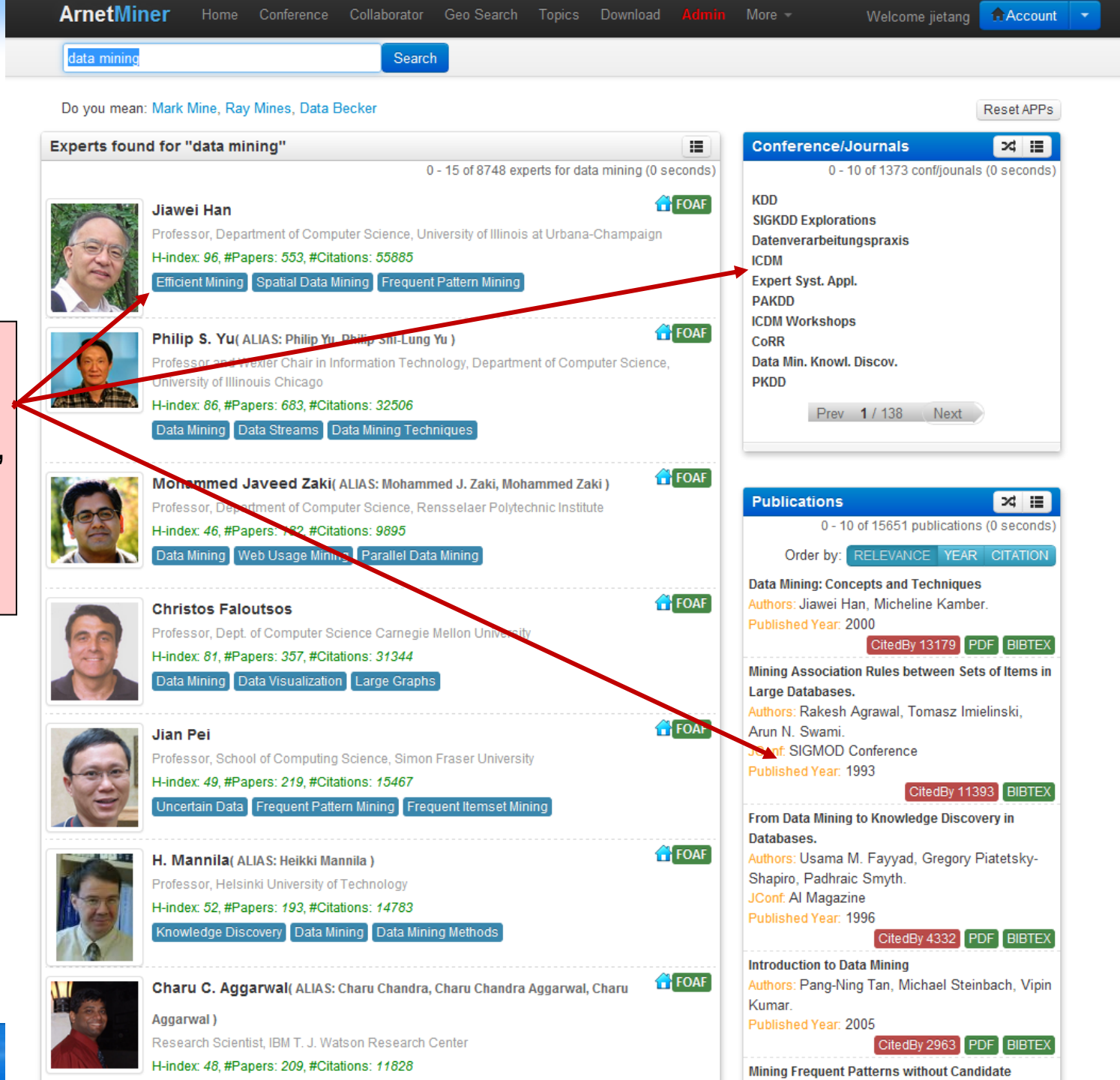
Research Interests

Publications



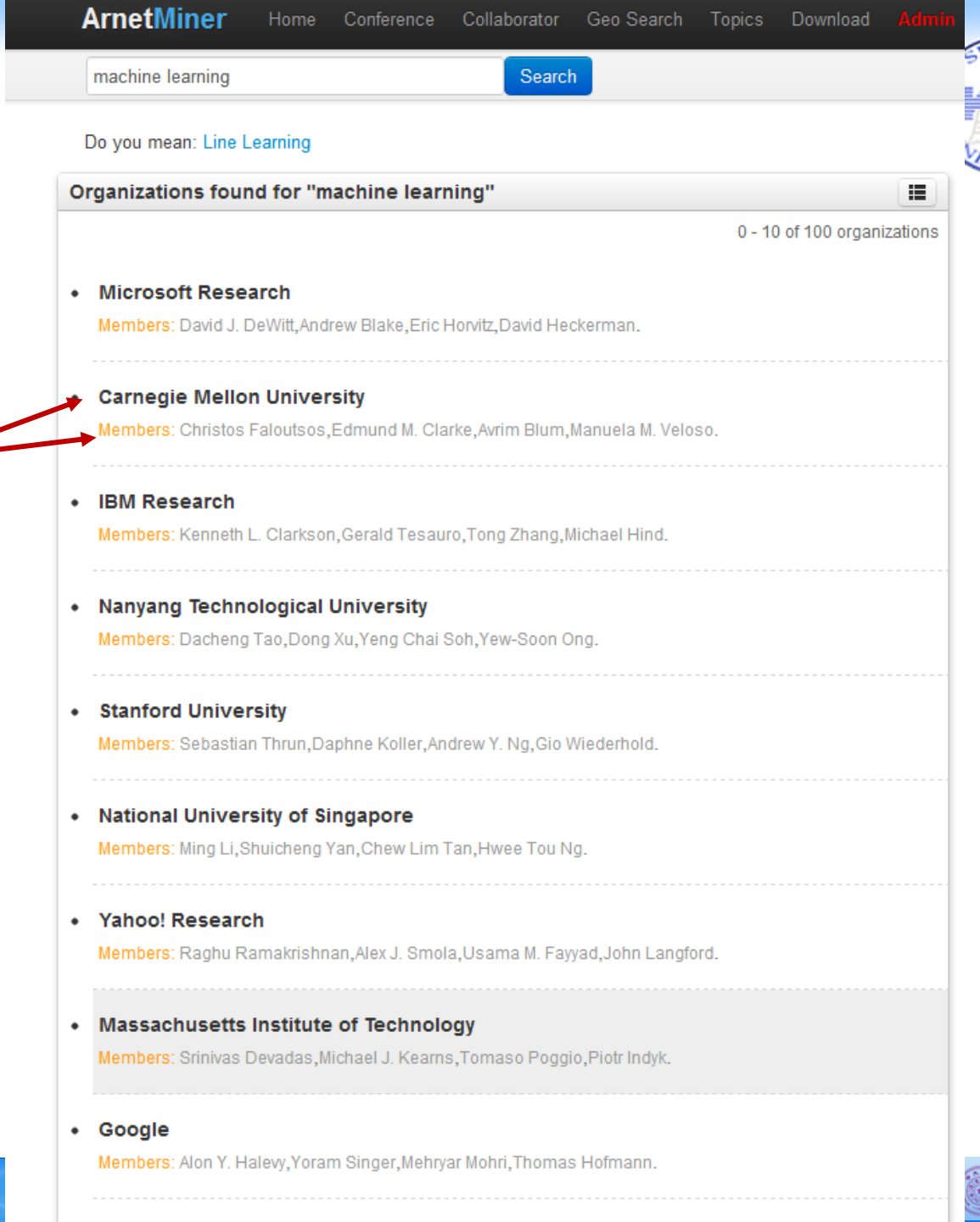
Expertise Search

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



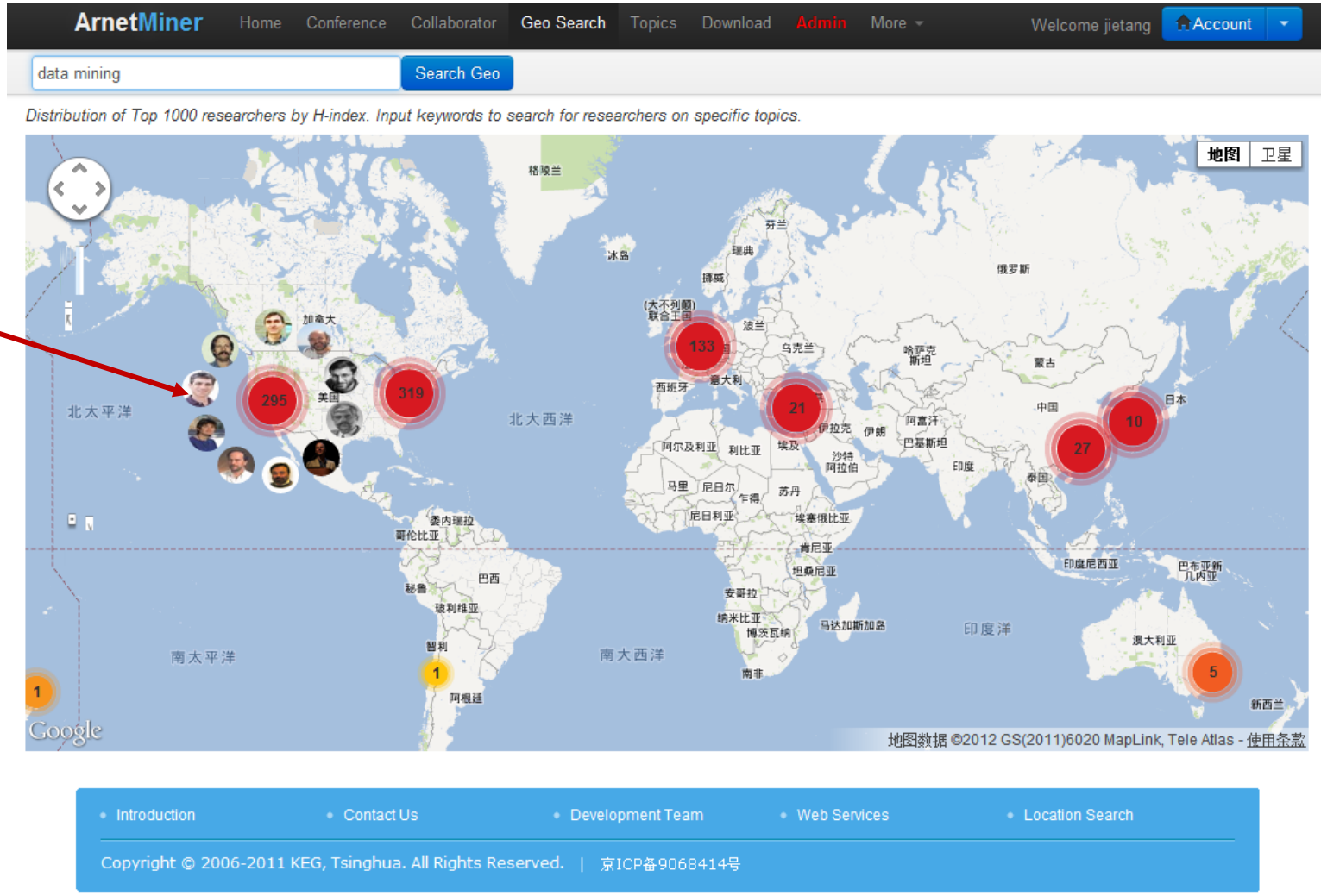
Organization Search

Ranking organizations on “machine learning”



Geographic search

Finding the most hot regions on “data mining”

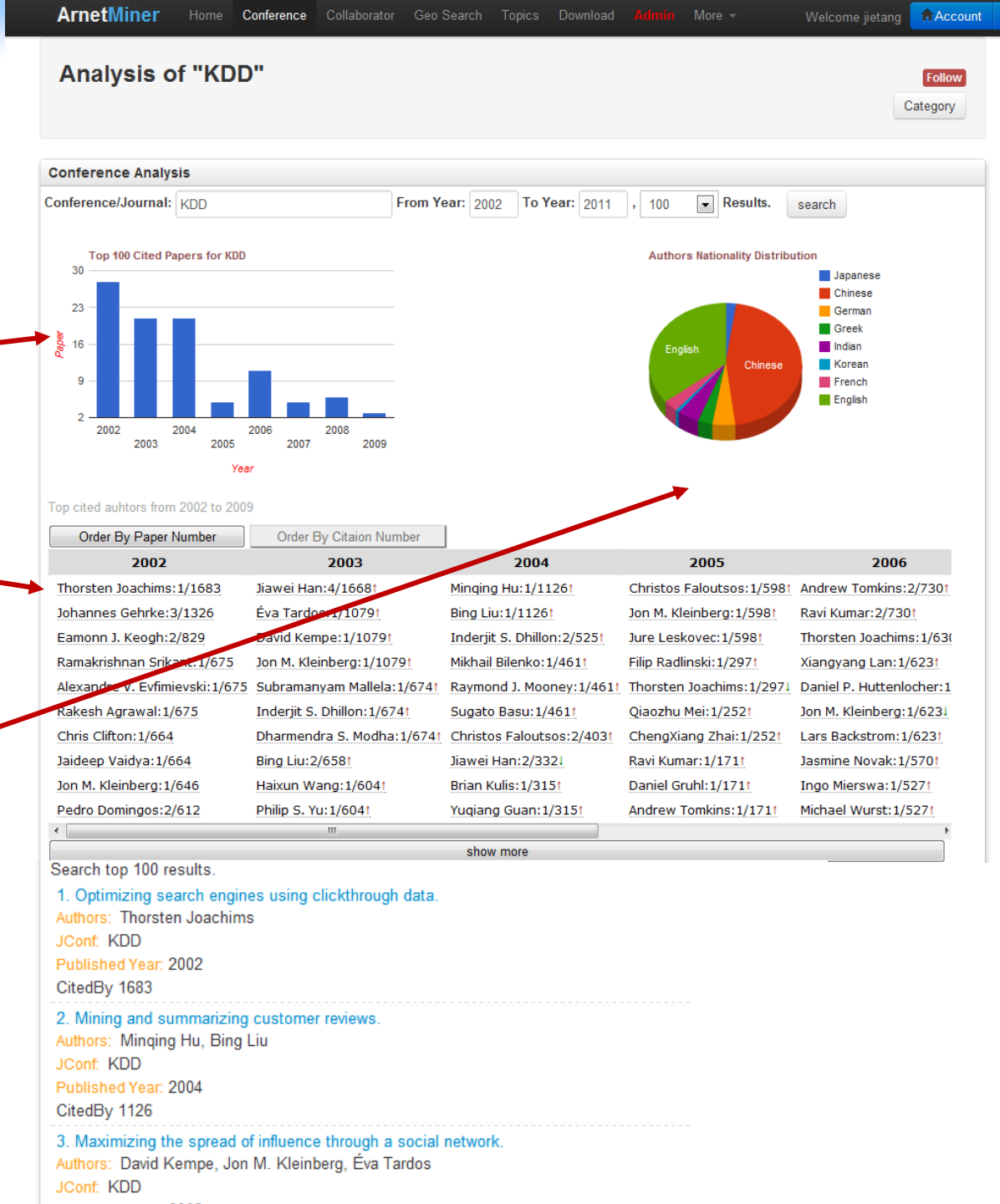


Conference Analysis

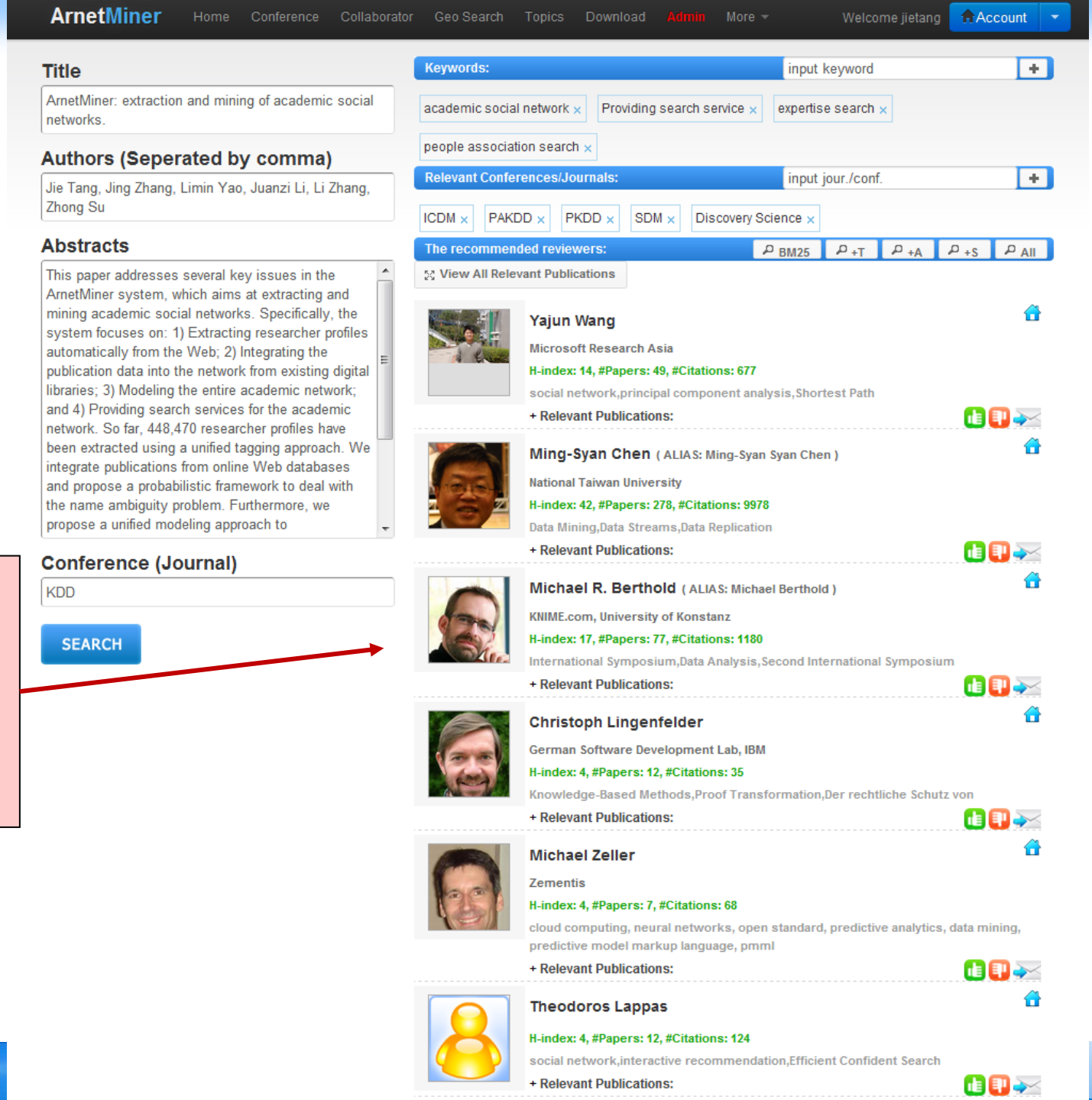
Which year is the most successful in the KDD's history?

Who are the most cited authors?

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



Reviewer Suggestion

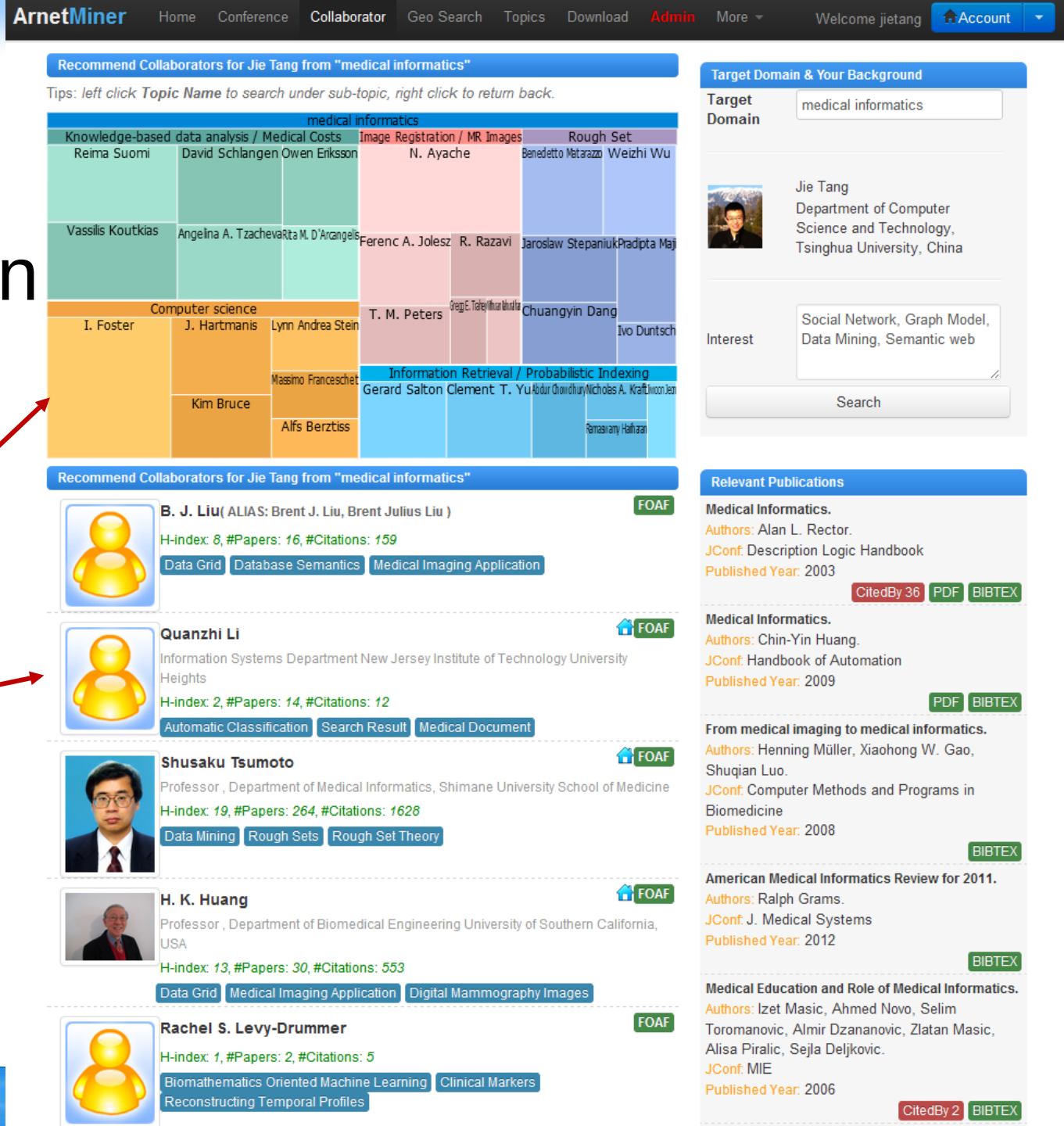


Inerest matching
COI avoiding
Load balancing
Forcast review quality

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

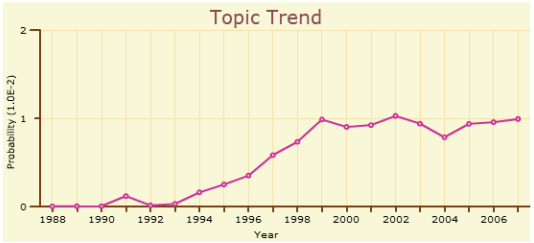
Active Inventors

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

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

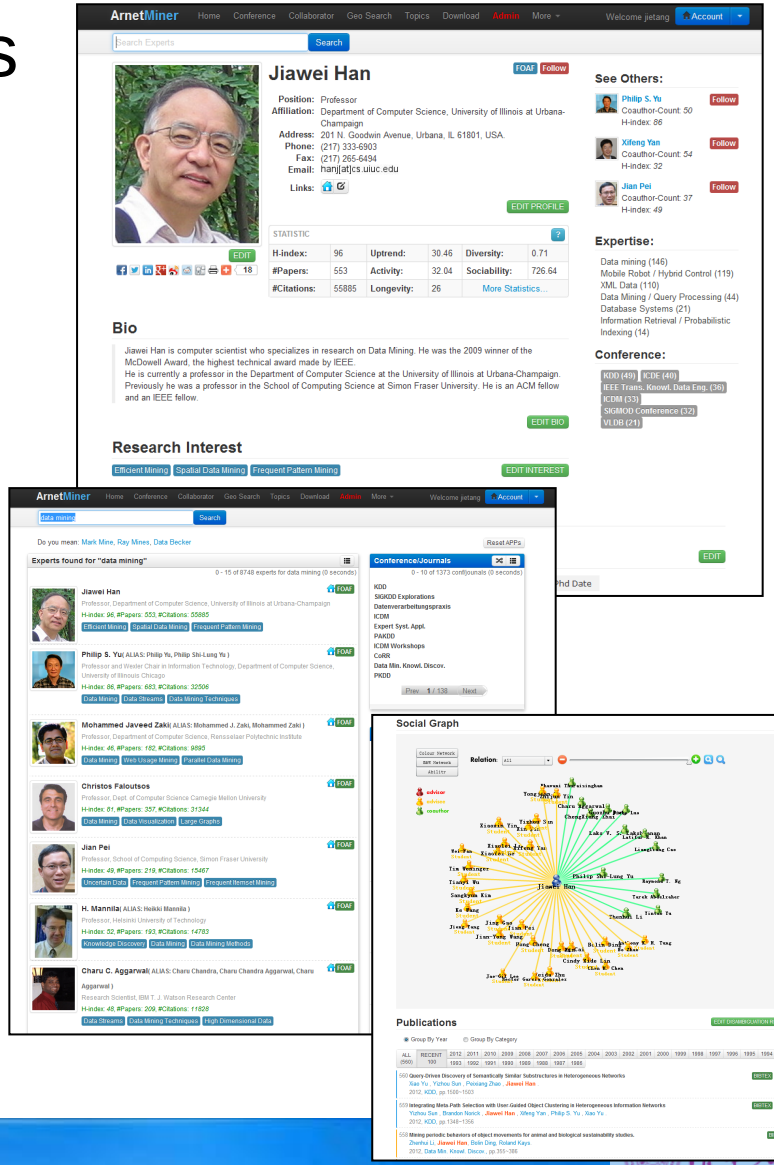
Expertise: Approximation Algorithms / Perfect Graphs Data mining

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/

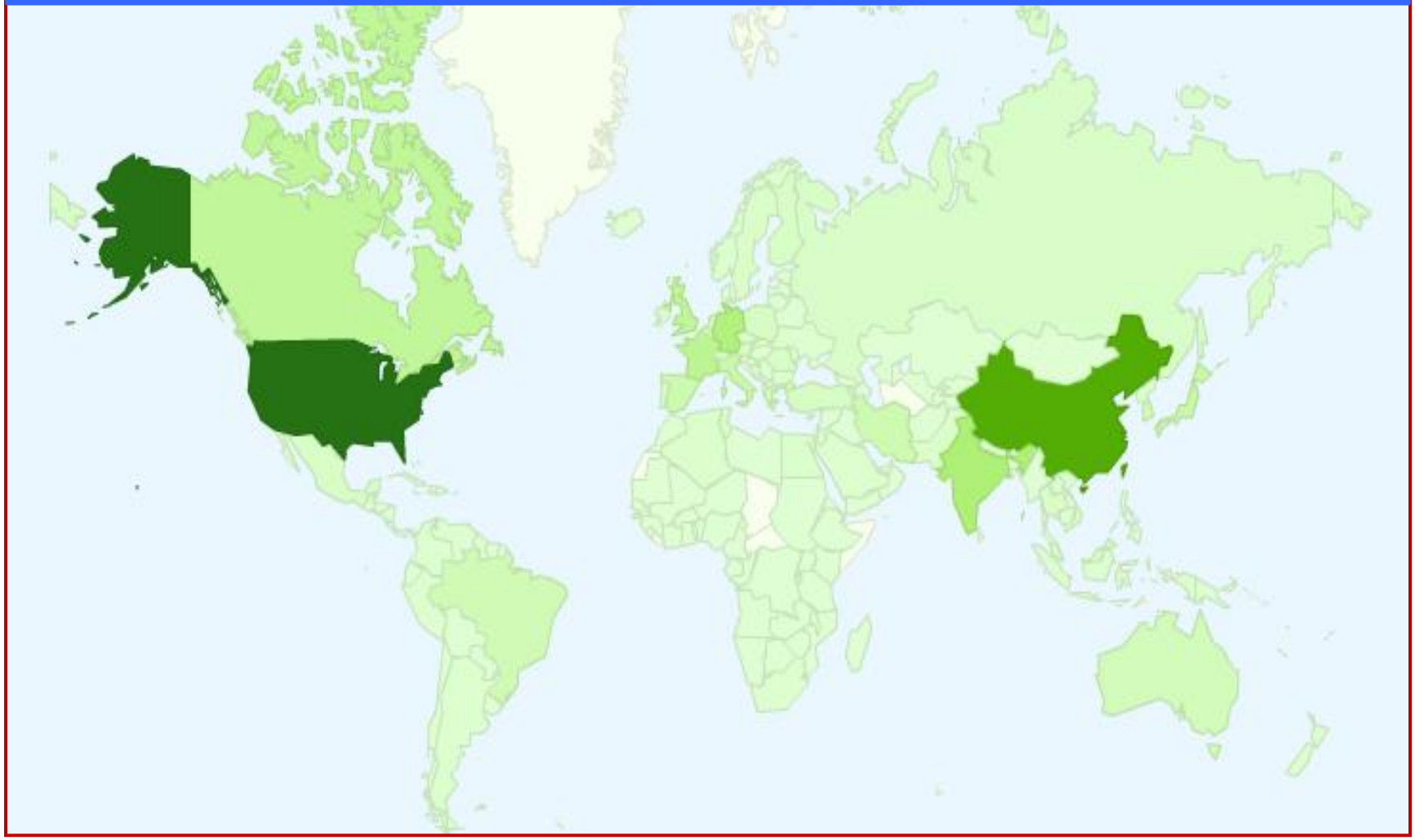
AMiner (ArnetMiner)

- 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 per month
  - 10% increase of visits per month
- Deep SN analysis, mining, and search



User Distribution

4.04 million IP from 220 countries/regions



User Distribution

4.04 million IP from 220 countries/regions

Top 10 countries

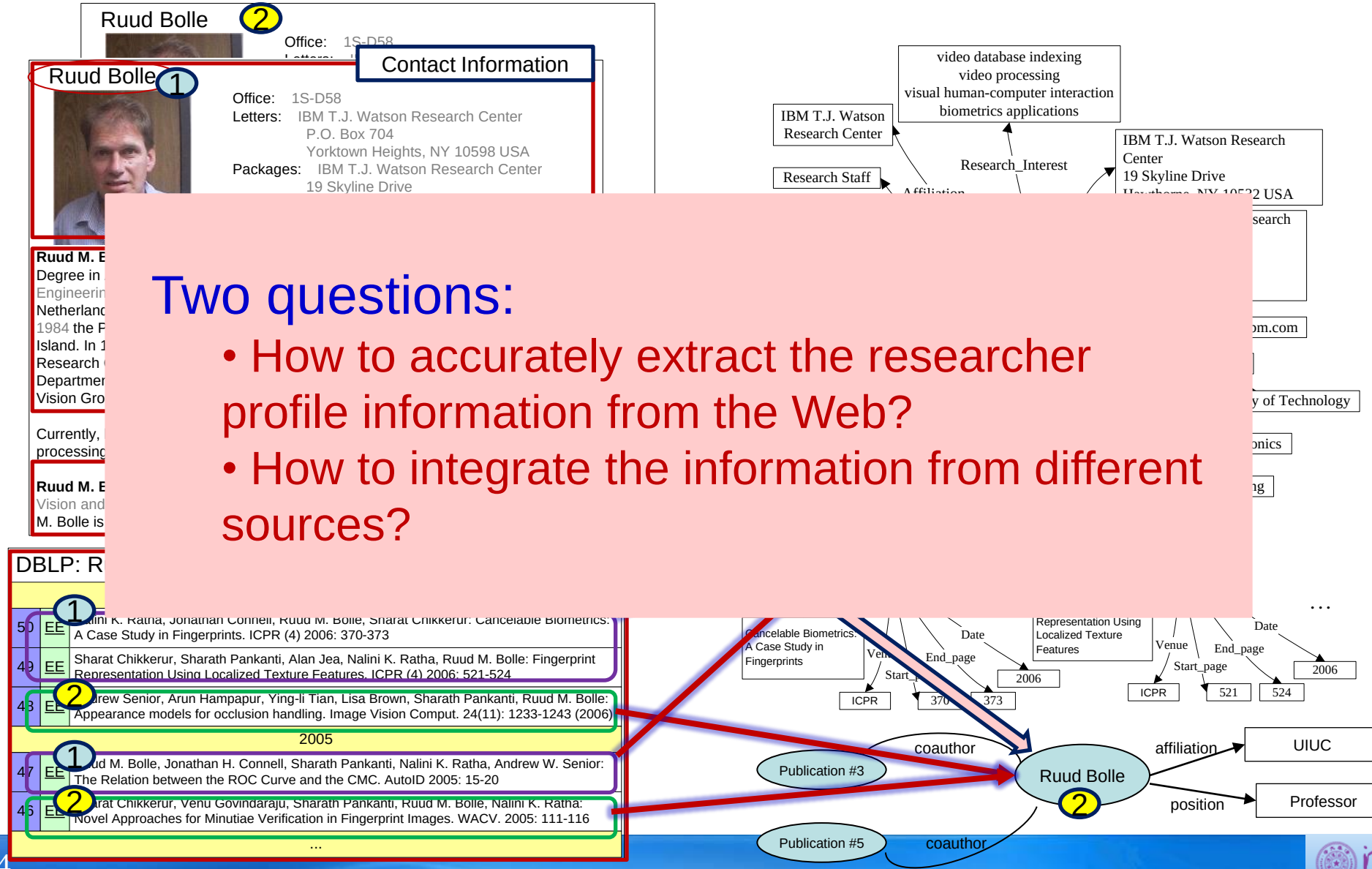
1. USA
2. China
3. Germany
4. India
5. UK
6. Canada
7. Japan
8. Spain
9. France
10. Italy

Outline

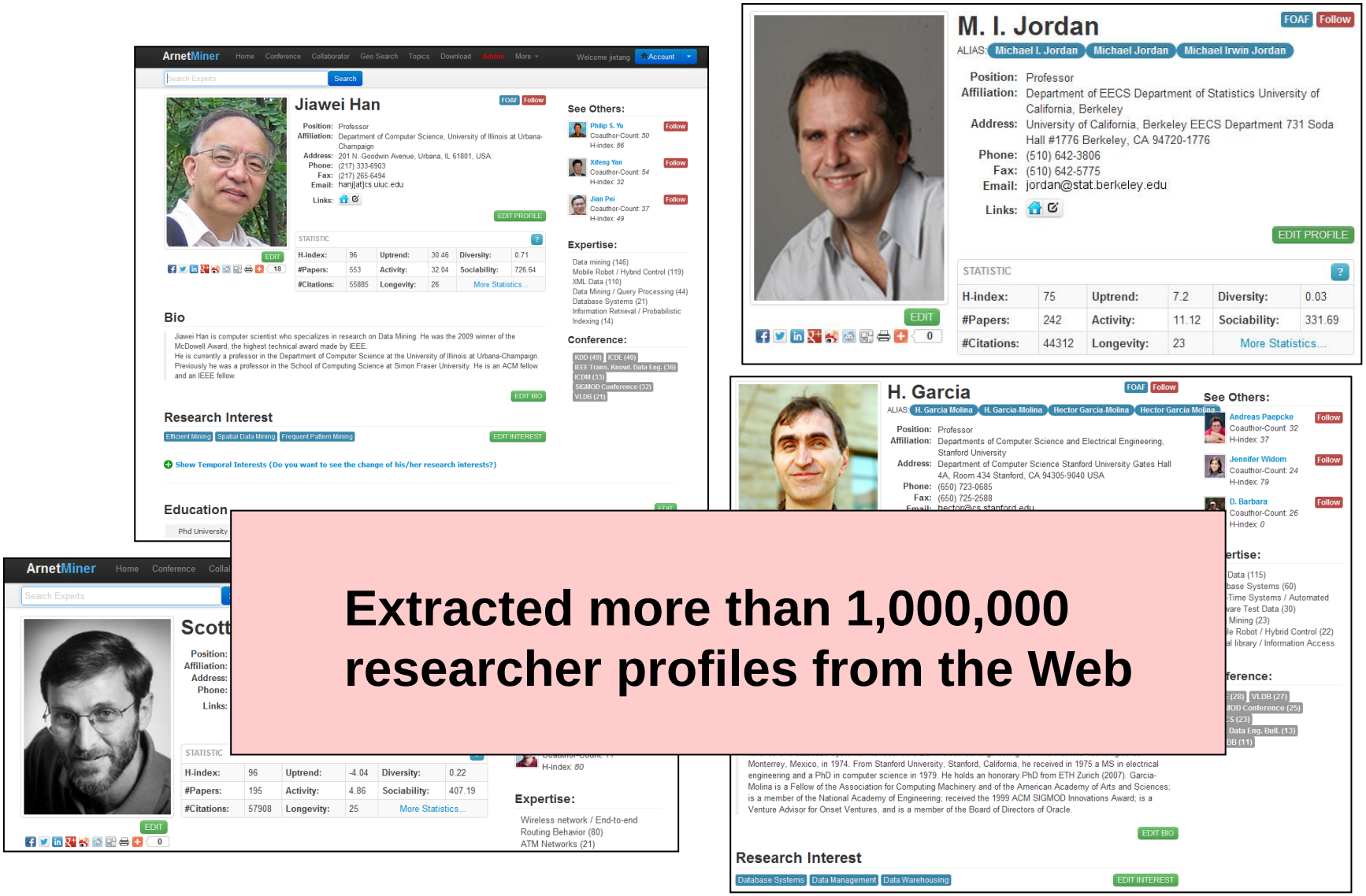
- ArnetMiner: Academic Social Network
- Core Techniques
  - Knowledge Acquisition
  - Modeling and Heterogeneous Ranking
  - Social Network Analysis

CT1: Knowledge Acquisition from Social Web

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



Researcher Profile Database[1]

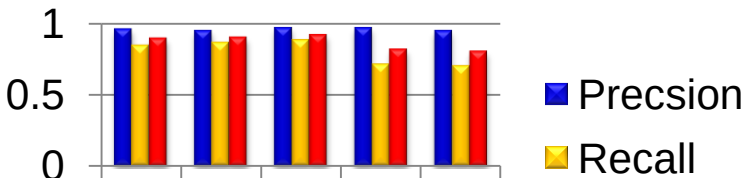


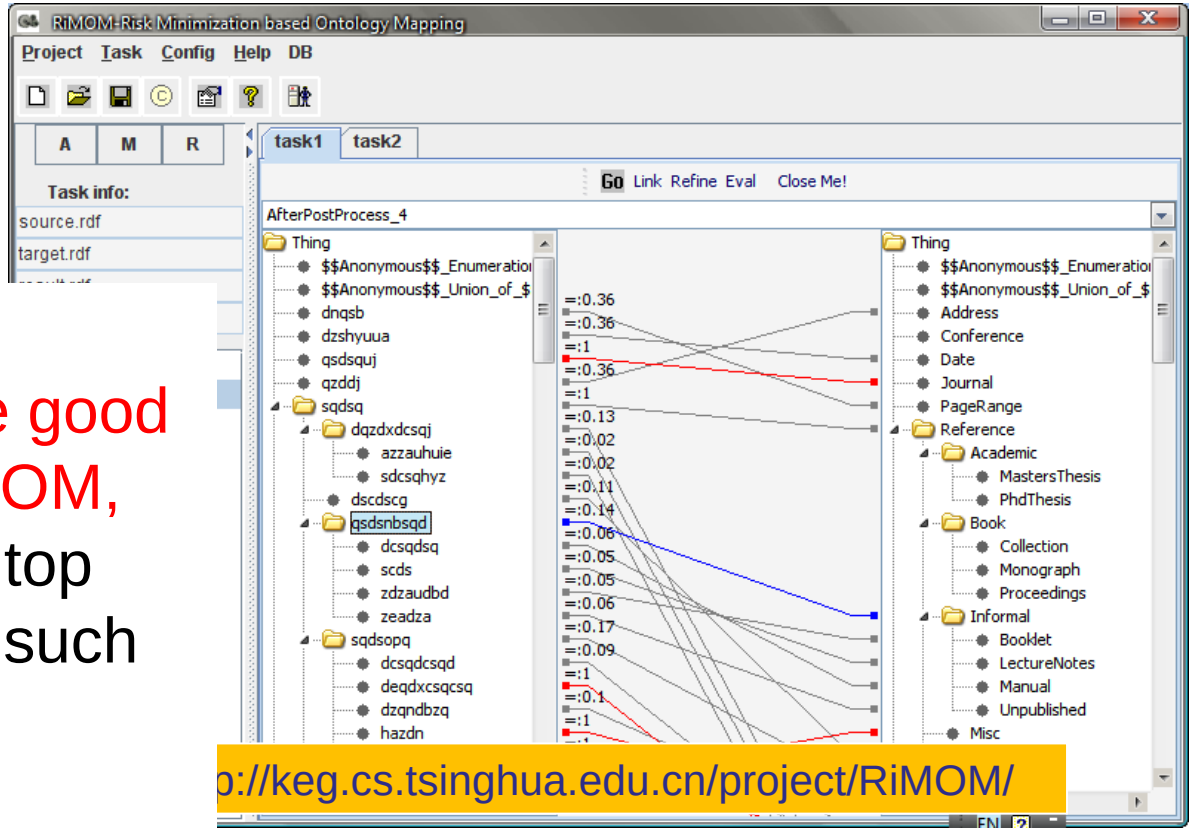
Extracted more than 1,000,000 researcher profiles from the Web

[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





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

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."

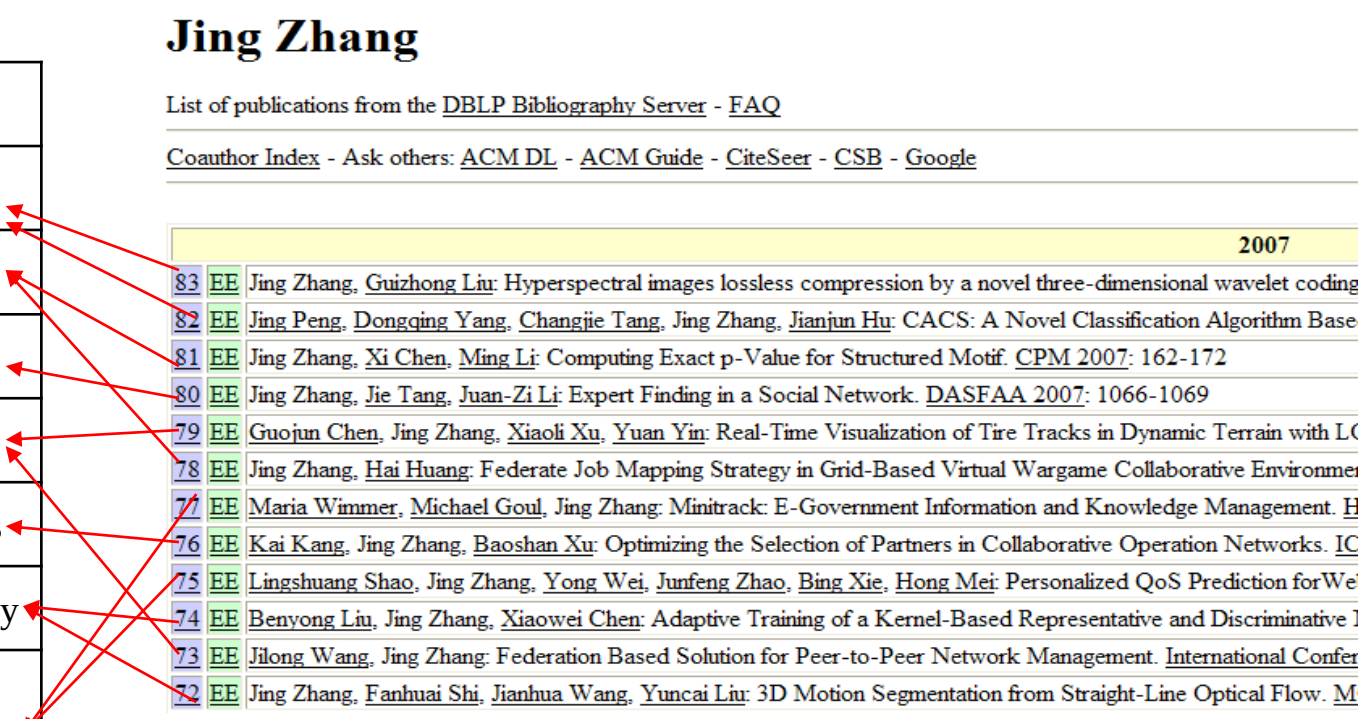
[1] J. Li, J. Tang, Y. Li, and Q. Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE Transaction on Knowledge and Data Engineering (TKDE). August 2009 (vol. 21 no. 8) pp. 1218-1232.

[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]

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



Jing Zhang

List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google

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. CPM 2007: 162-172

80 EE Jing Zhang, Jie Tang, Juan-Zi Li: Expert Finding in a Social Network. DASFAA 2007: 1066-1069

79 EE Guojun Chen, Jing Zhang, Xiaoli Xu, Yuan Yin: Real-Time Visualization of Tire Tracks in Dynamic Terrain with LO

78 EE Jing Zhang, Hai Huang: Federate Job Mapping Strategy in Grid-Based Virtual Wargame Collaborative Environmer

77 EE Maria Wimmer, Michael Goul, Jing Zhang: Minitrack: E-Government Information and Knowledge Management. H

76 EE Kai Kang, Jing Zhang, Baoshan Xu: Optimizing the Selection of Partners in Collaborative Operation Networks. IC

75 EE Lingshuang Shao, Jing Zhang, Yong Wei, Junfeng Zhao, Bing Xie, Hong Mei: Personalized QoS Prediction forWe

74 EE Benyong Liu, Jing Zhang, Xiaowei Chen: Adaptive Training of a Kernel-Based Representative and Discriminative I

73 EE Jilong Wang, Jing Zhang: Federation Based Solution for Peer-to-Peer Network Management. International Confer

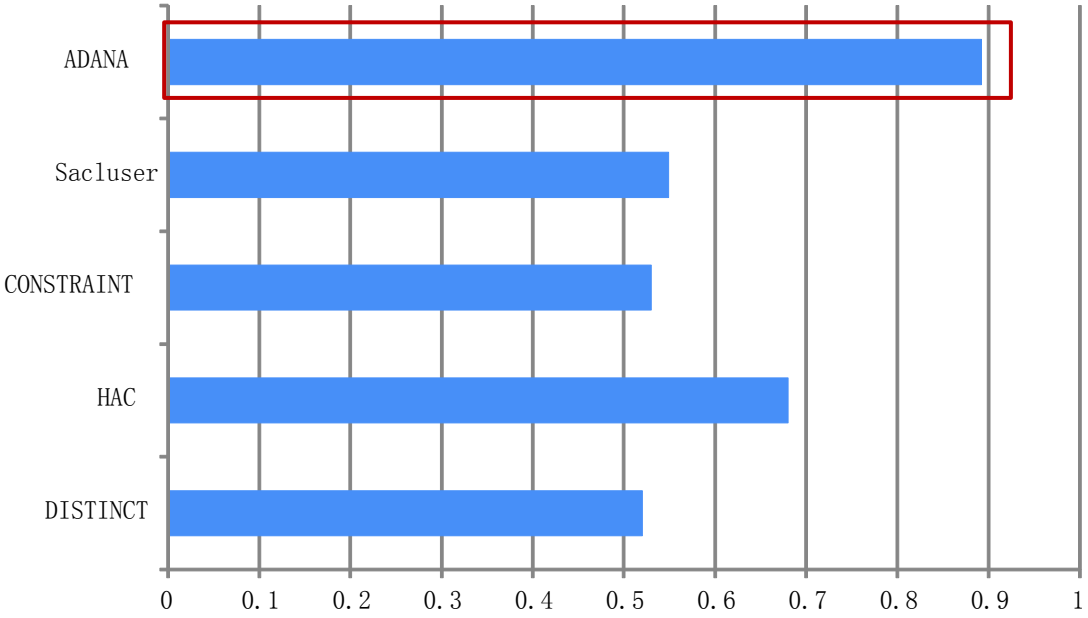
72 EE Jing Zhang, Fanhuai Shi, Jianhua Wang, Yuncai Liu: 3D Motion Segmentation from Straight-Line Optical Flow. M

- 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

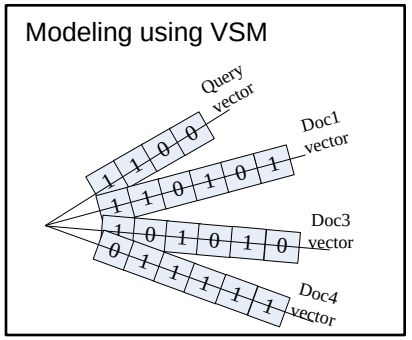


Outline

- 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)

Search with keyword



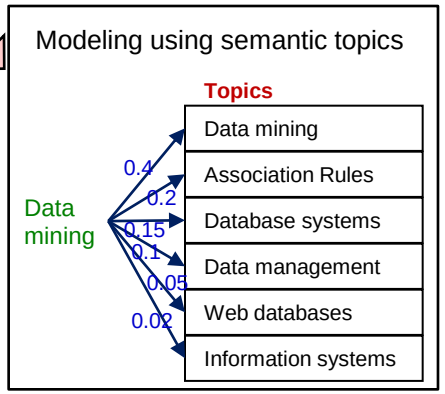
Return

[Principles of Data Mining.](#)
DJ Hand - Drug Safety, 2007 - drugsafety.adisonline.com

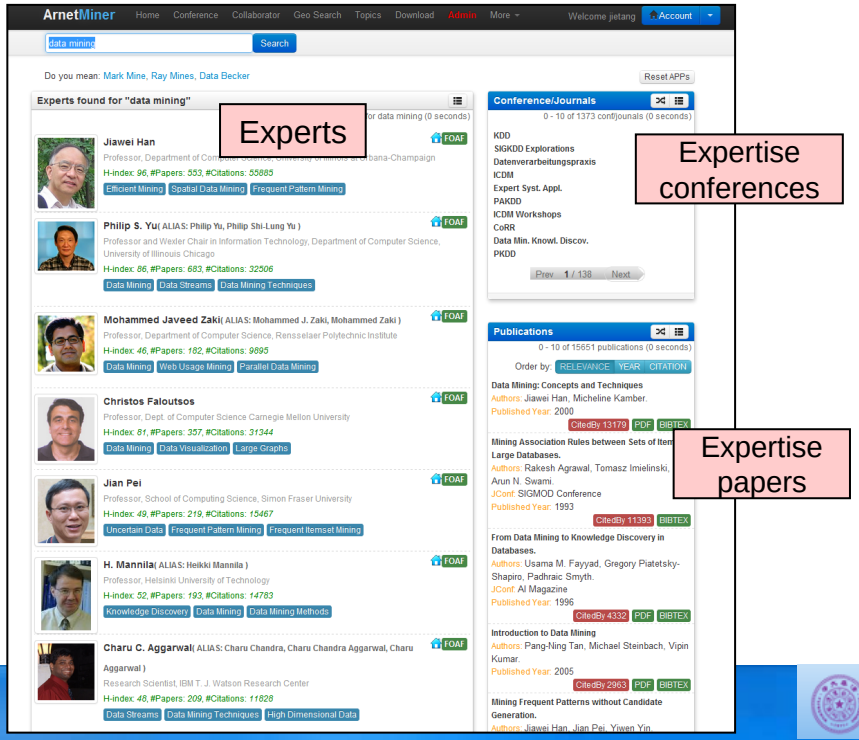
[Advances in Knowledge Discovery and Data Mining](#)
UM Fayyad, G Piatetsky-Shapiro, P Smyth, R…

[Data Mining: Concepts and Techniques](#)
J Han, M Kamber - 2001…

Search with semantic modeling

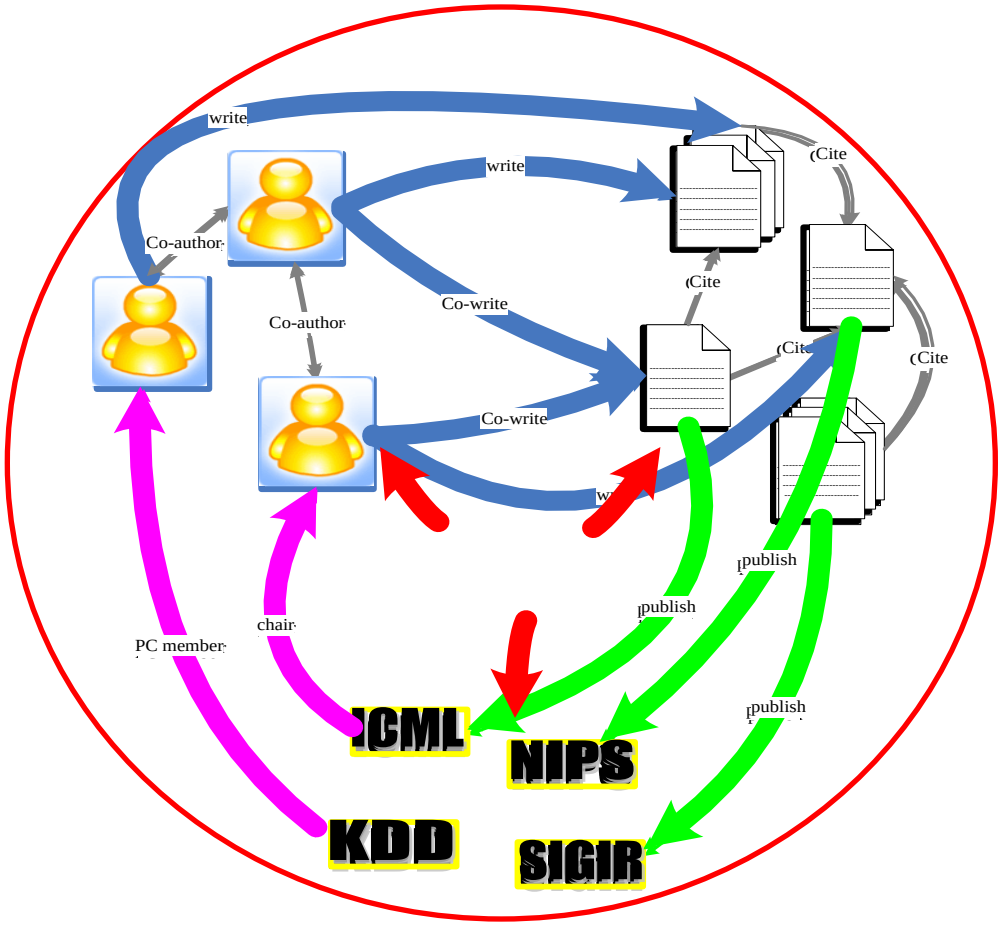


Return



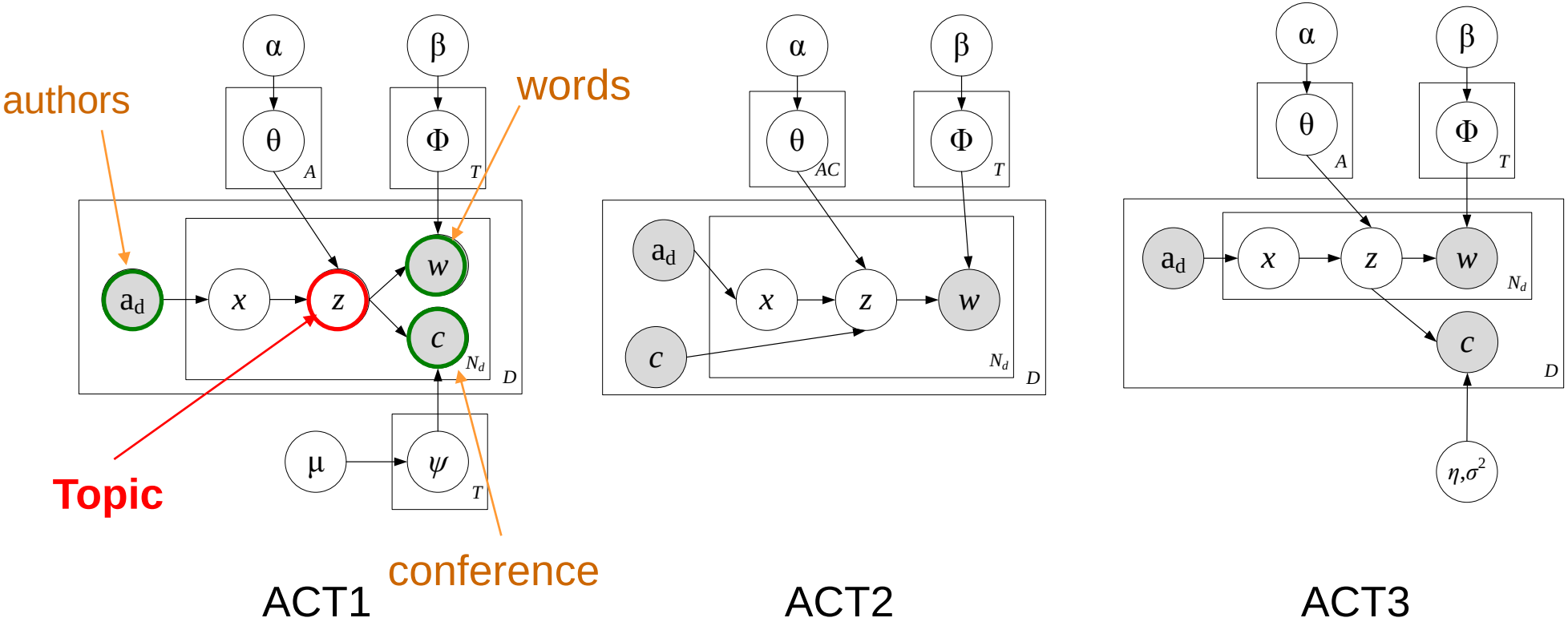
Challenges

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]

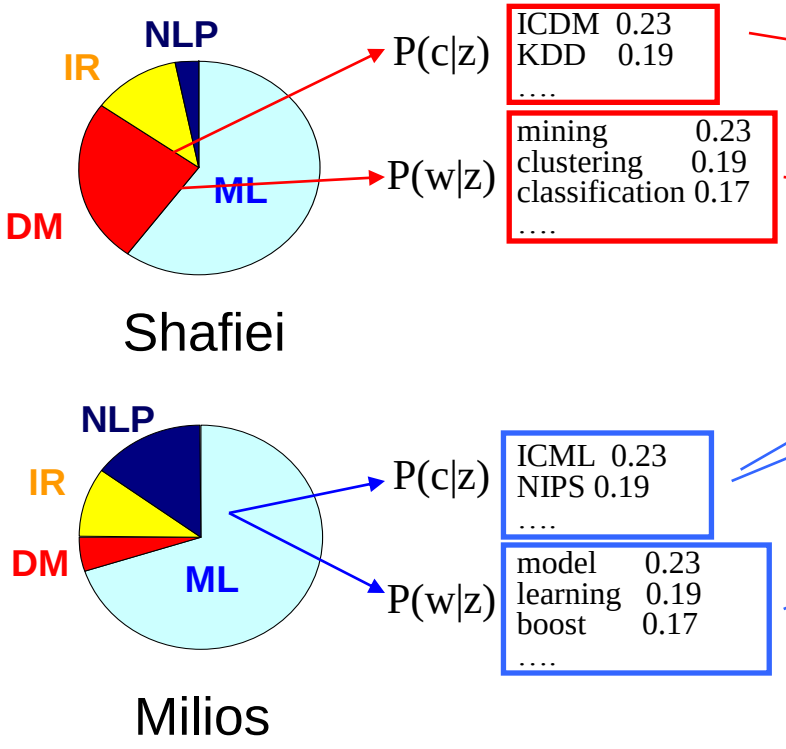


Author-Conference-Topic Model

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. SIGKDD'08. pages 990-998, 2008.

Generative Story of ACT1 Model

- Generative process



Paper

Latent Dirichlet Co-clustering

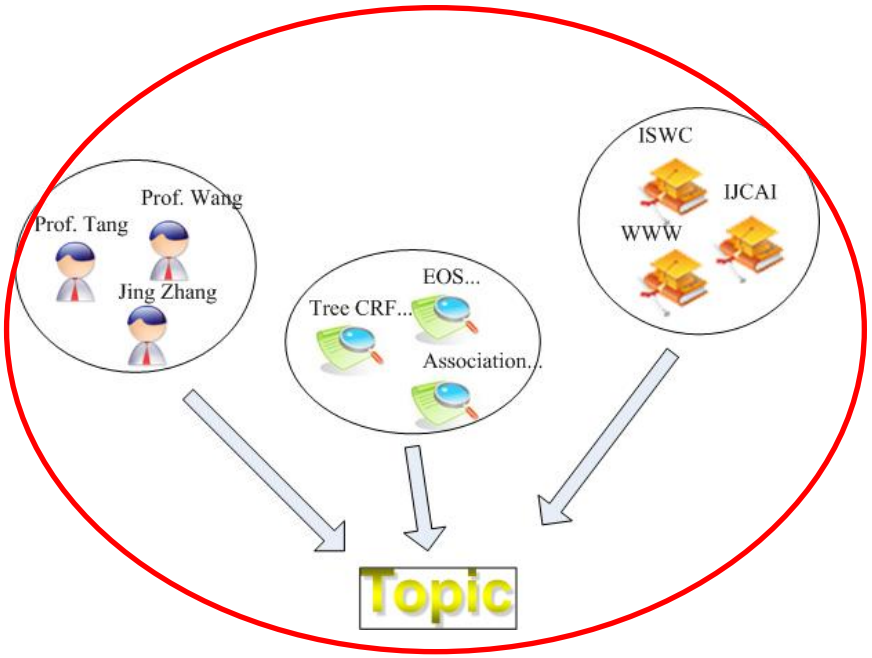
Shafiei and Milios

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 …

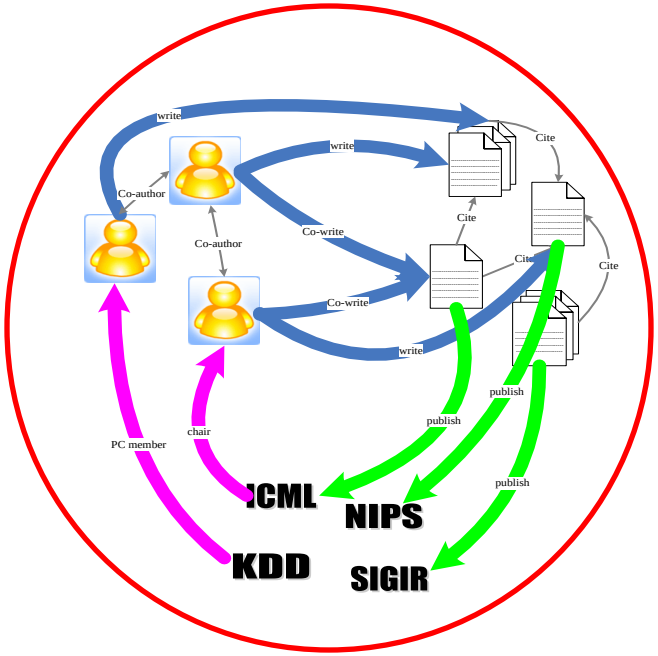
(2) Integrating with Random Walk[1]

Modeling academic network with topics



+

Random walk over the academic network



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.

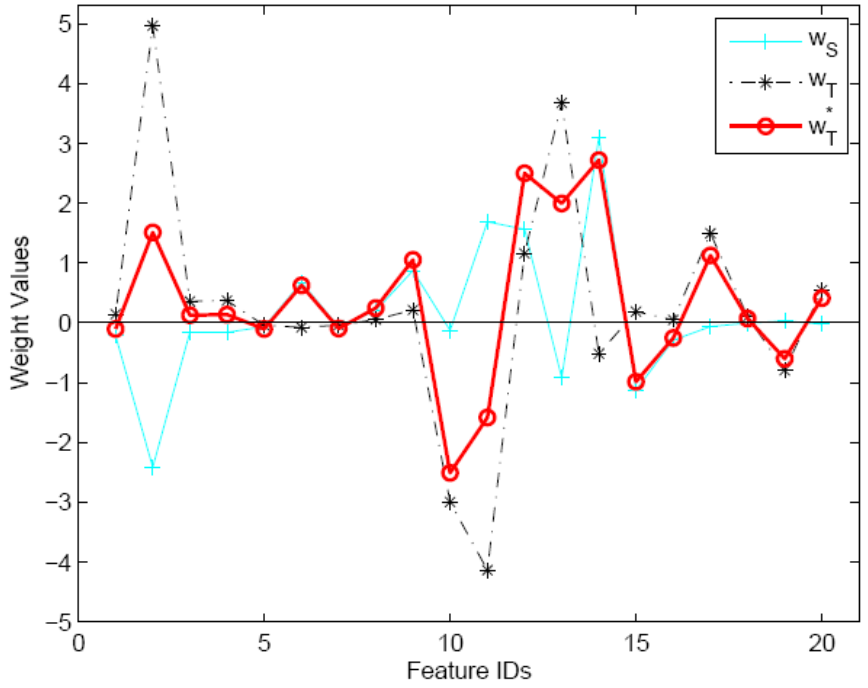
Experimental Results

- 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
RSVM	.8084	.6071	.5820	.5854	.6285



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.

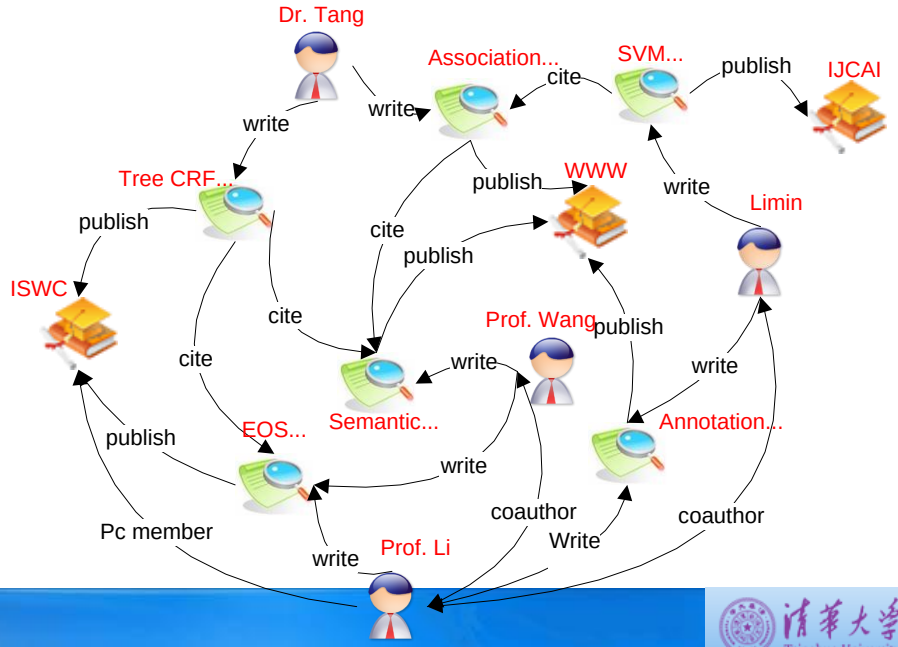
Approach	P@5	P@10	P@15	MAP	N@5	N@10
RSVM	.7714	**.8429**	.8285	.7756	.5545	.5947
RSVMt	.8000	.8286	.8476	.7837	.5923	.5999
MTRSVM	.8000	.8286	.8476	.7875	.6140	.6075
HCDRank	**.8285**	.7857	**.8571**	**.7971**	**.6189**	**.6112**
Language model	.6250	.6875	.6500	.6726	.3343	.3809

Outline

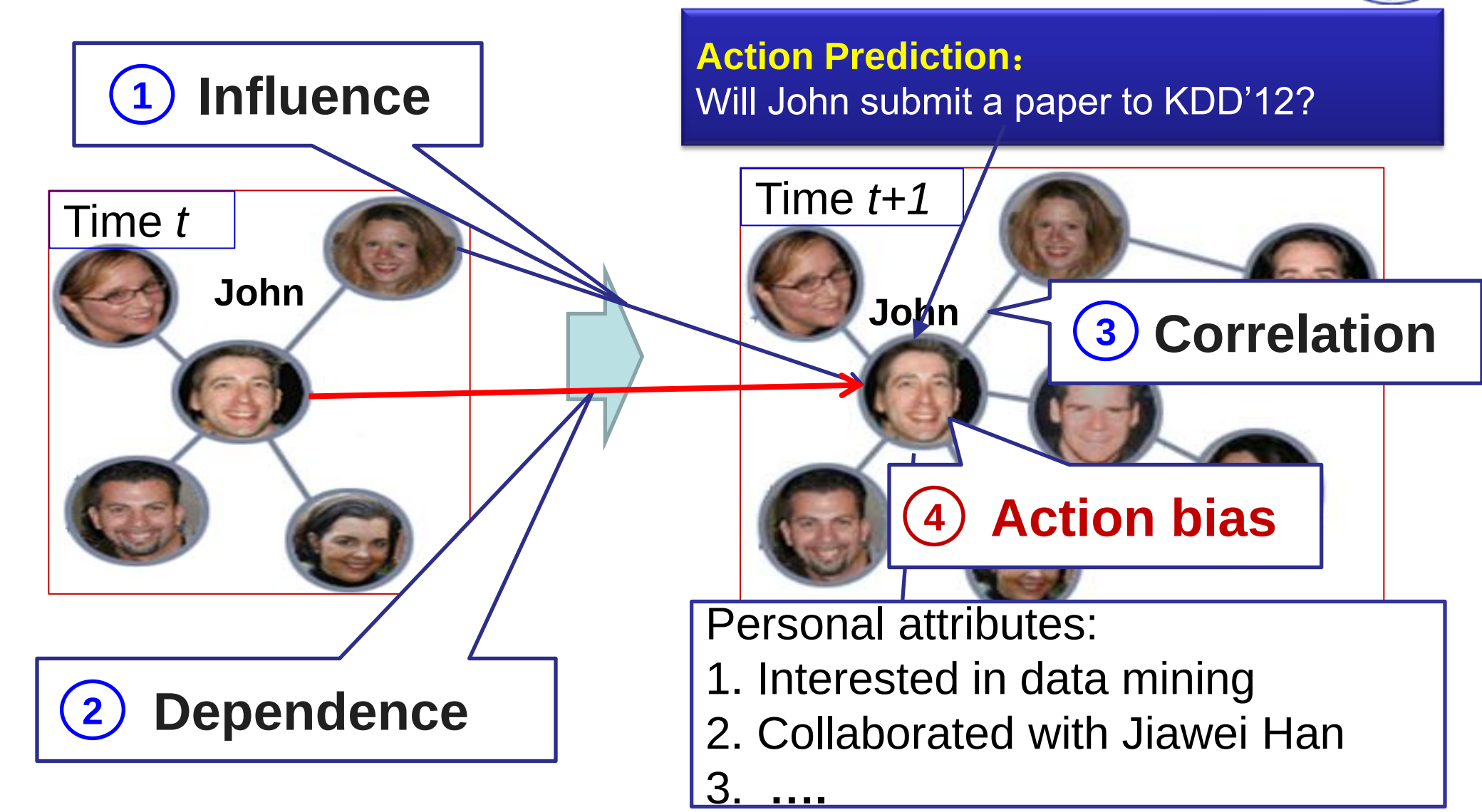
- ArnetMiner: Academic Social Network
- Core Techniques
  - Knowledge Acquisition
  - Modeling and Heterogeneous Ranking
  - **Social Network Analysis**

CT3: Social Network Analysis
(KDD'09, '10, '11, DMKD'12, JIS)

- **User modeling:** can we model and forcast 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 structrure?



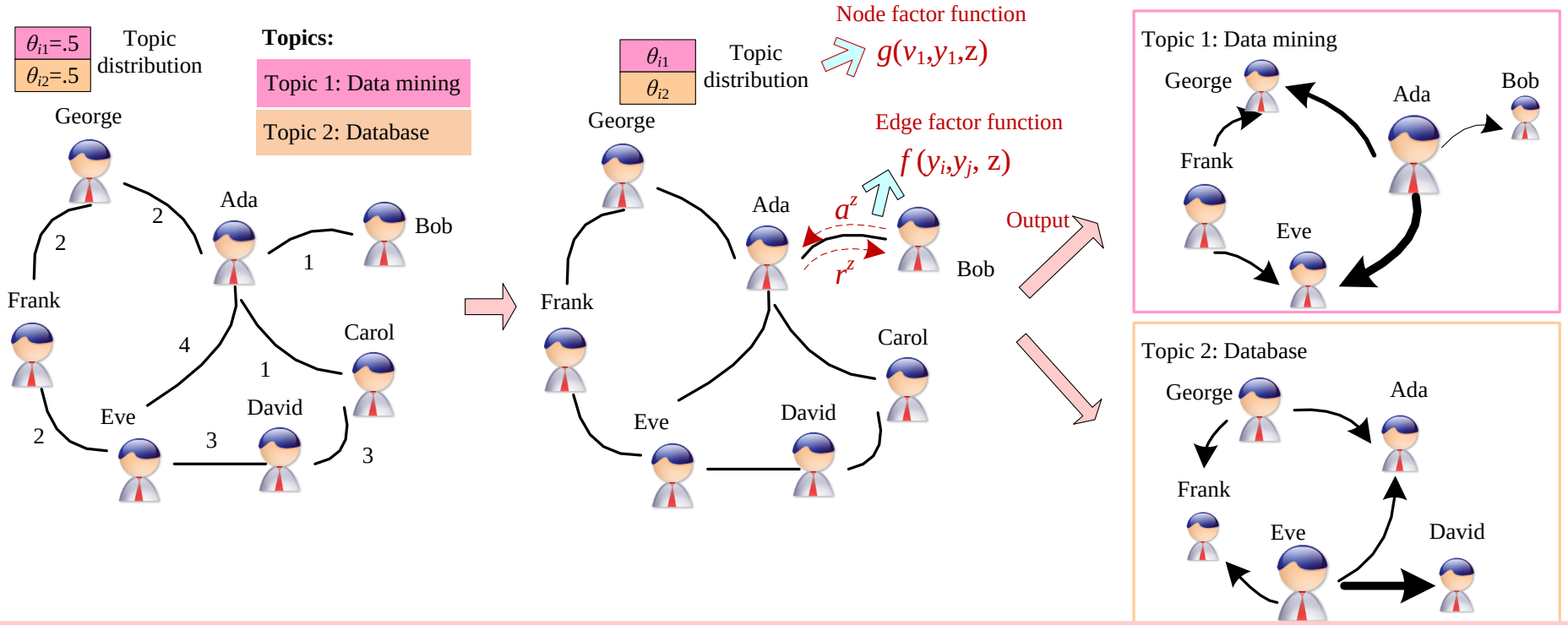
(1) Social Action Modeling and Prediction[1]



[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 anlysis** | **Output: topic-based social influences**



Several key challenges:

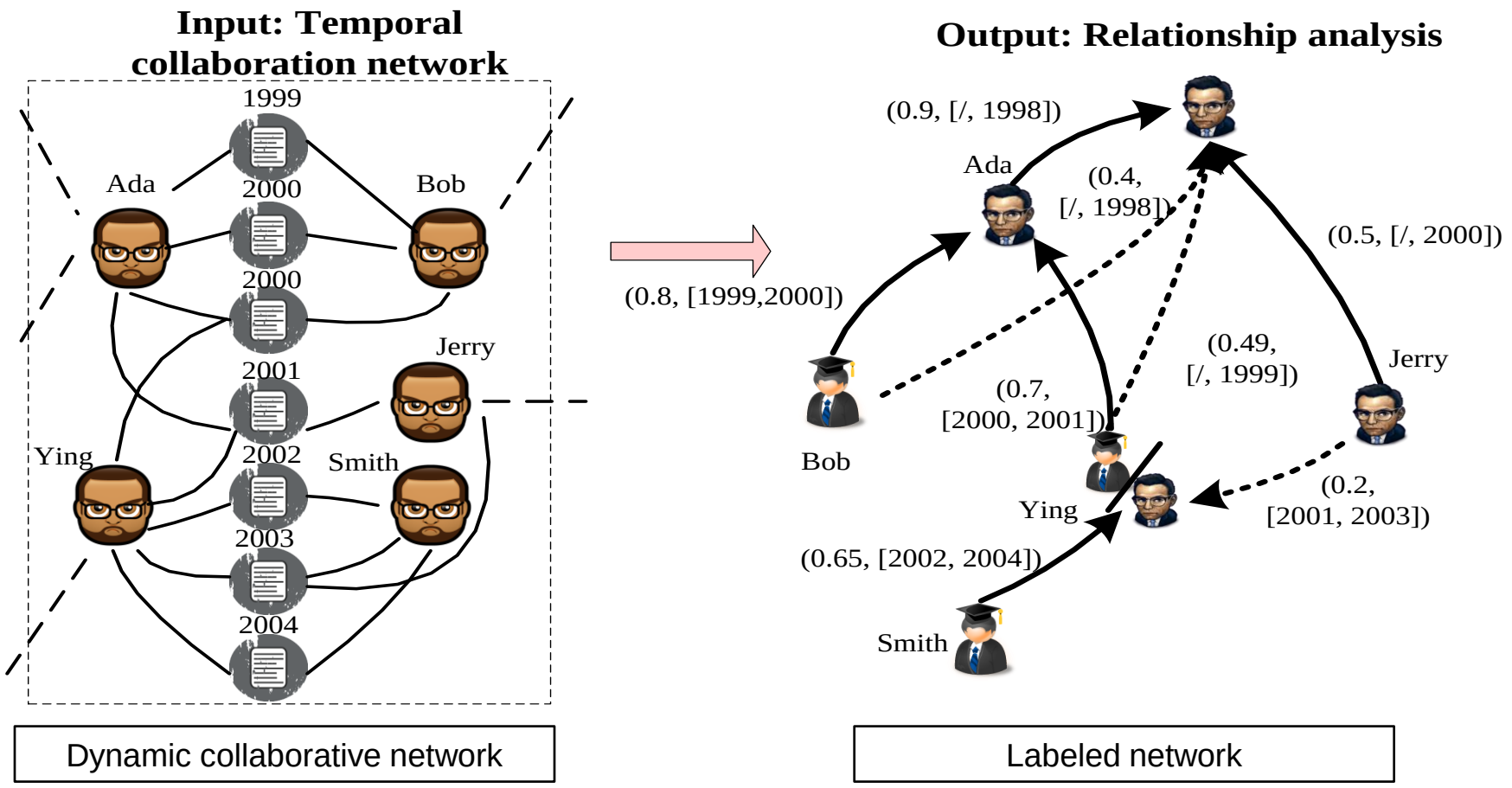
- **How to differentiate the social influences from different topics?**
- **How to incorporate both structure and content into a unified model?**
- **How to estimate the model on real-large networks?**

[1] J. Tang, J. Sun, C.Wang, and Z. Yang. Social influence analysis in largescale networks. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'09), pages 807–816, 2009.

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]



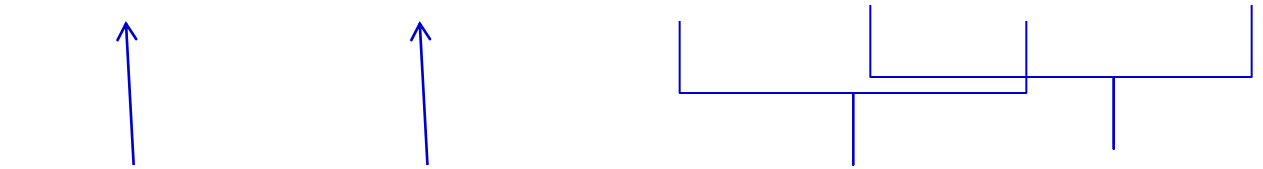
Output: potential types of relationships and their probabilities:
(type, prob, [s_time, e_time])

[1] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. SIGKDD'10. pages 203-212, 2010.

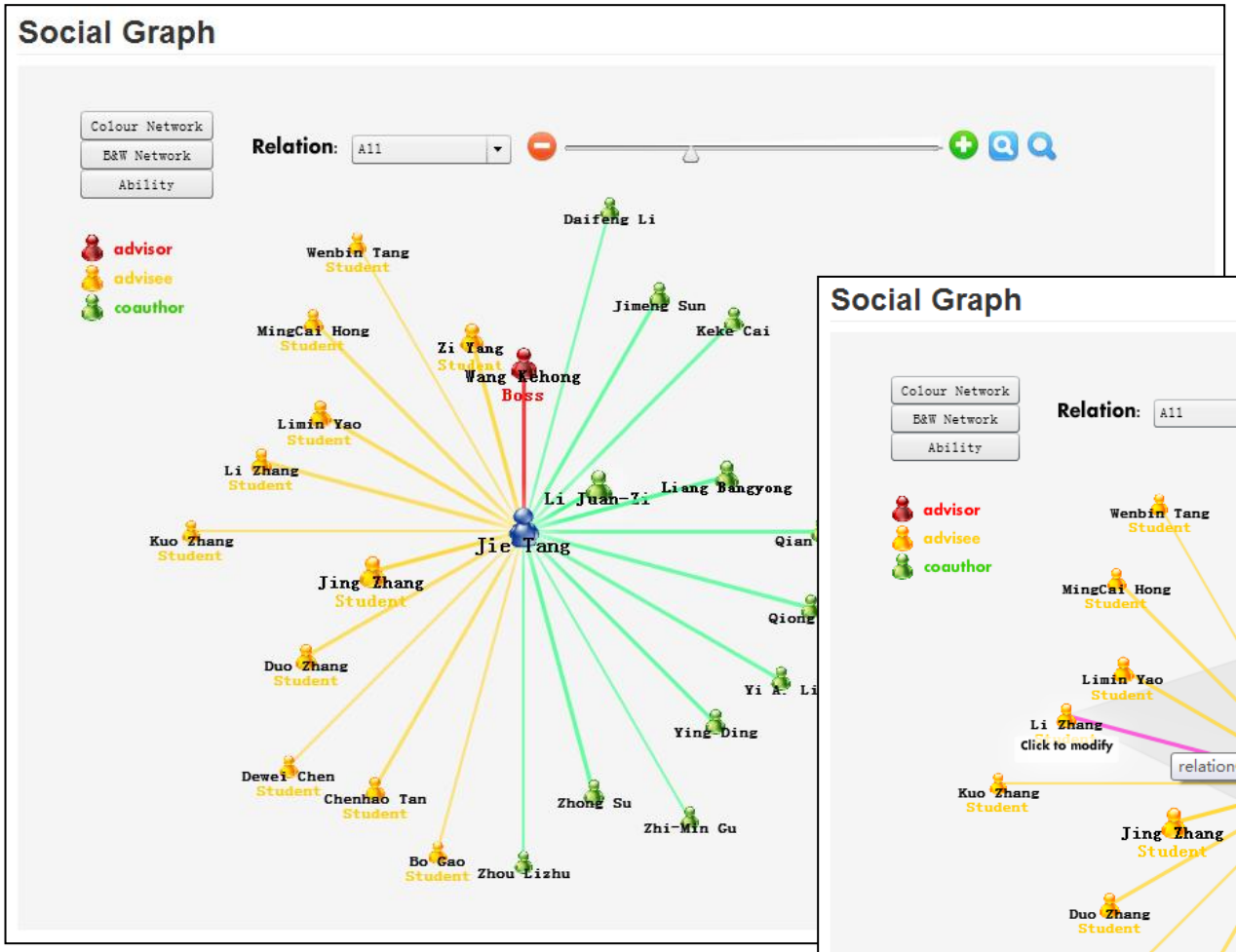
Results

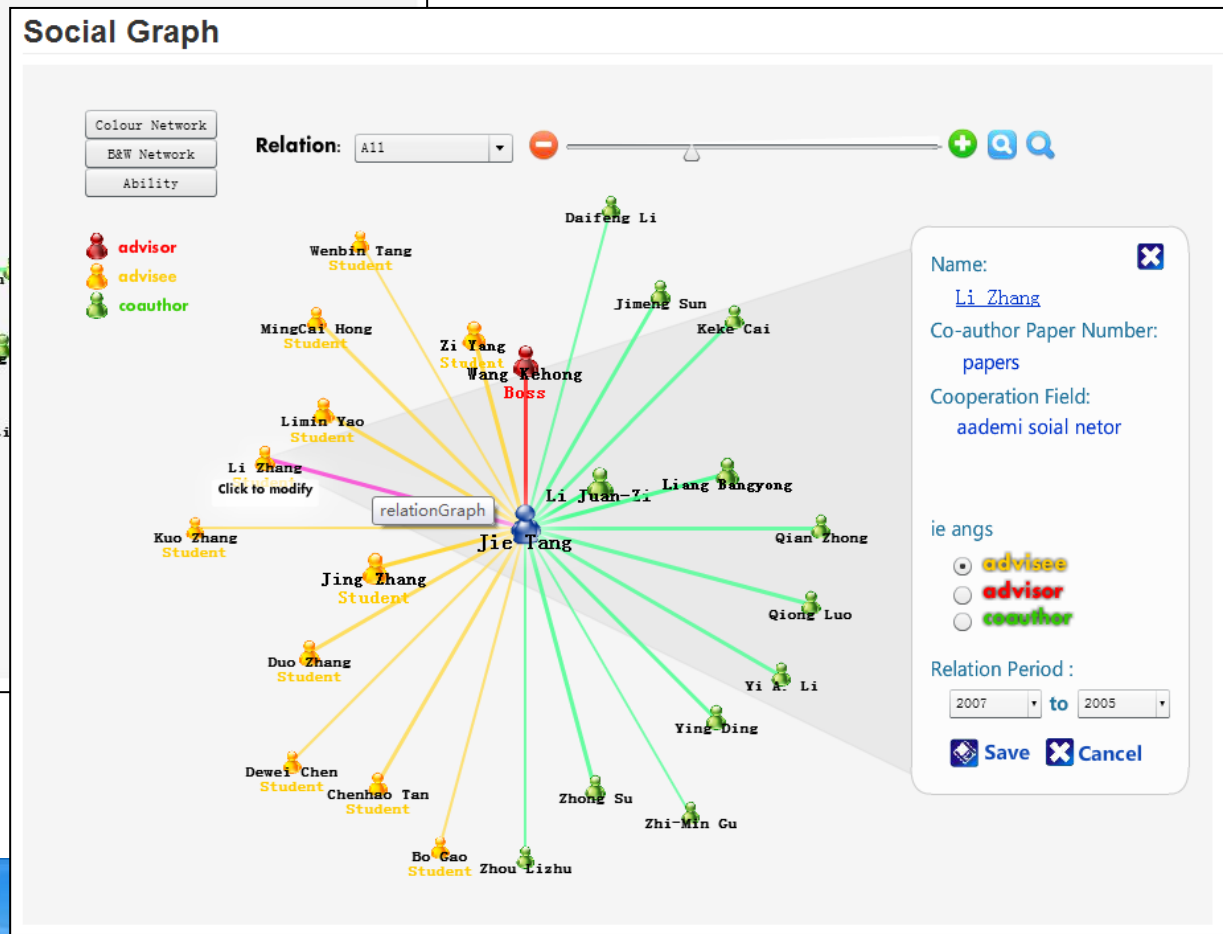
- DBLP data: 654, 628 authors, 1,076,946 publications, years provided.
- Ground truth: MathGenealogy Project; AI Genealogy Project; Faculty Homepage

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%**

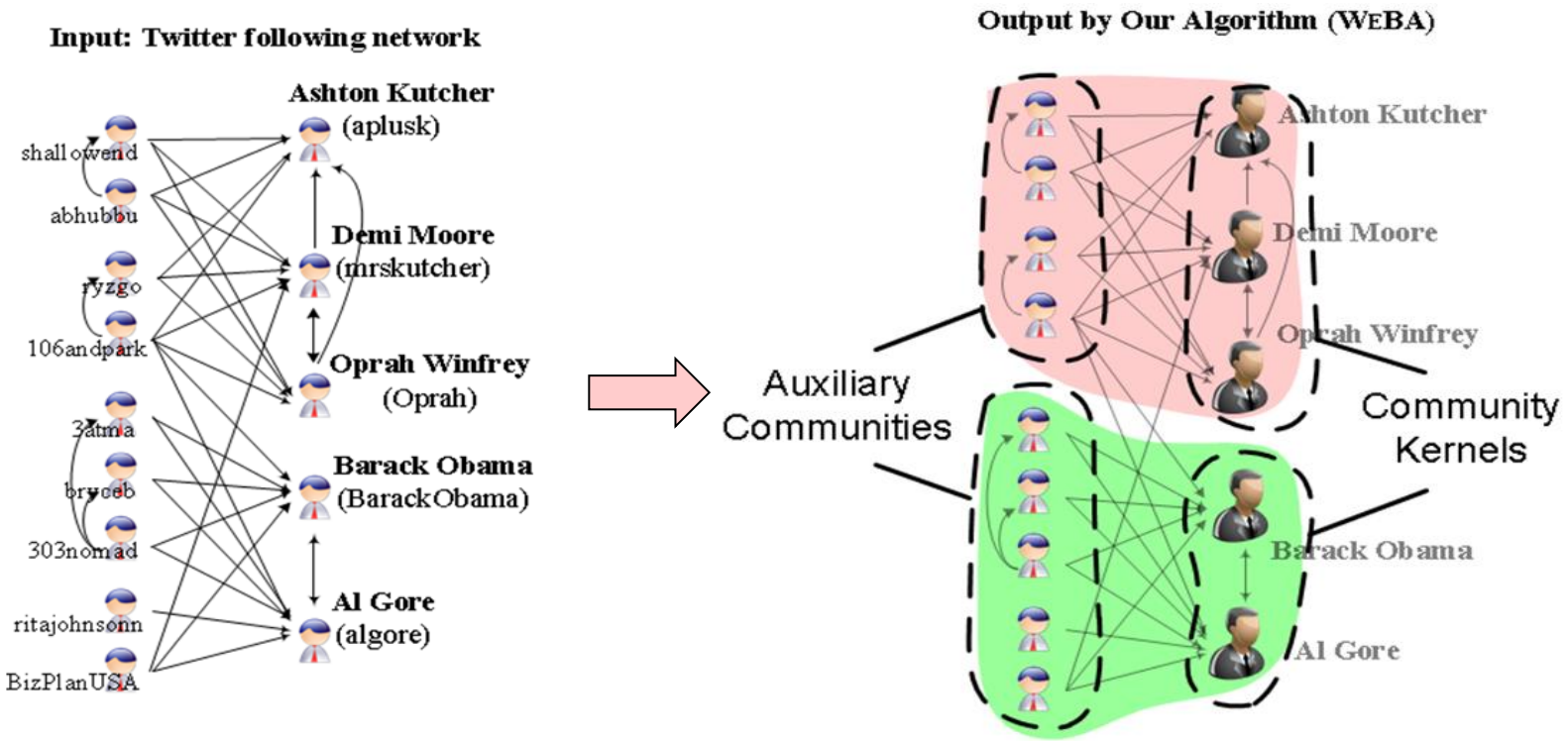


Results


Social Graph
Colour Network
B&W Network
Ability
Relation: All
advisor
advisee
coauthor
Daifeng Li
Wenbin Tang
Student
Jimeng Sun
Keke Cai
MingCai Hong
Student
Zi Yang
Student
Wang Kehong
Boss
Limin Yao
Student
Li Zhang
Student
Li Juan-Zi
Liang Bangyong
Kuo Zhang
Student
Jie Tang
Jing Zhang
Student
Qiong
Duo Zhang
Student
Yi A. Li
Ying-Ding
Dewei Chen
Student
Chenhao Tan
Student
Zhong Su
Zhi-Min Gu
Bo Gao
Student
Zhou Lizhu


Social Graph
Colour Network
B&W Network
Ability
Relation: All
advisor
advisee
coauthor
Daifeng Li
Wenbin Tang
Student
Jimeng Sun
Keke Cai
MingCai Hong
Student
Zi Yang
Student
Wang Kehong
Boss
Limin Yao
Student
Li Zhang
Click to modify
relationGraph
Li Juan-Zi
Liang Bangyong
Kuo Zhang
Student
Jie Tang
Qian Zhong
Jing Zhang
Student
Qiong Luo
Duo Zhang
Student
Yi A. Li
Ying-Ding
Dewei Chen
Student
Chenhao Tan
Student
Zhong Su
Zhi-Min Gu
Bo Gao
Student
Zhou Lizhu
Name:
Li Zhang
Co-author Paper Number:
papers
Cooperation Field:
aademi soial netor
ie angs
advisee
advisor
coauthor
Relation Period :
2007
to
2005
Save
Cancel

(4) Community Kernel[1]



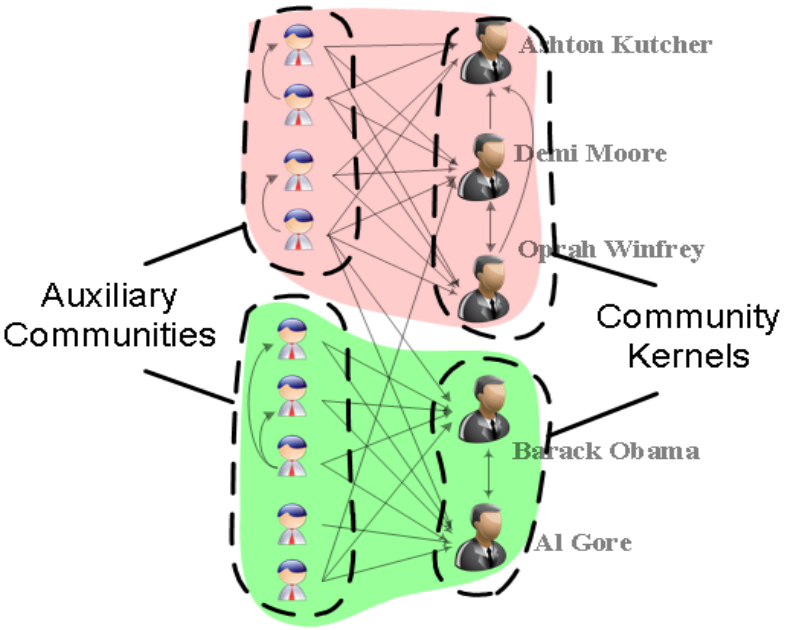
Pareto Principle: Less than 1% of the Twitter users (e.g. entertainers, politicians, writers) produce 50% of its content, while the others (e.g. fans, followers, readers) have much less influence and completely different social behavior.

Kernel users and ordinary users exhibit very different behaviors.

[1] L. Wang, T. Lou, J. Tang, and J. Hopcroft. Detecting Community Kernels in Large Social Networks. ICDM'11, pages 784-793, 2011.

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

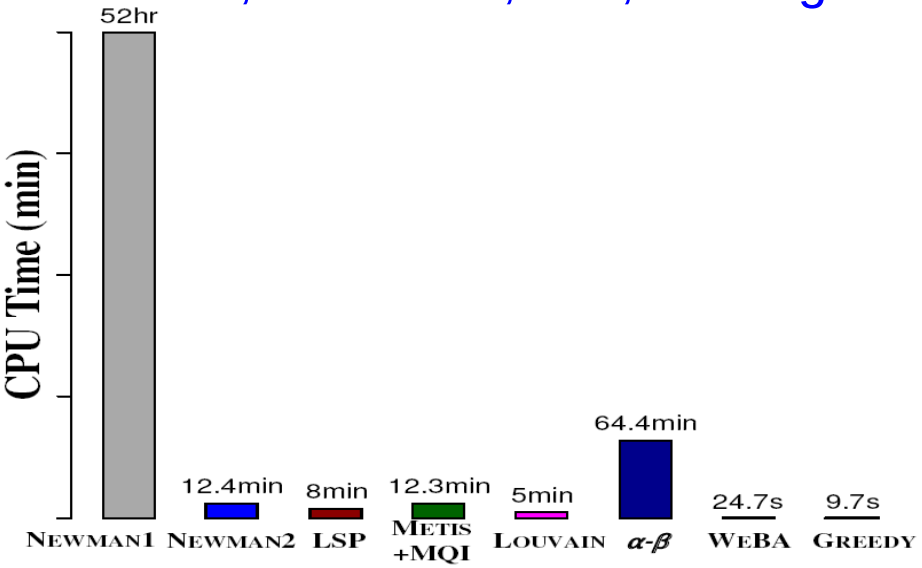
- **On average, WEBA improves Precision by 340% (wiki) and 70% (coauthor), and improves Recall by 130% (wiki) and 41% (coauthor).**

	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	0.069	0.007	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
α-β	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

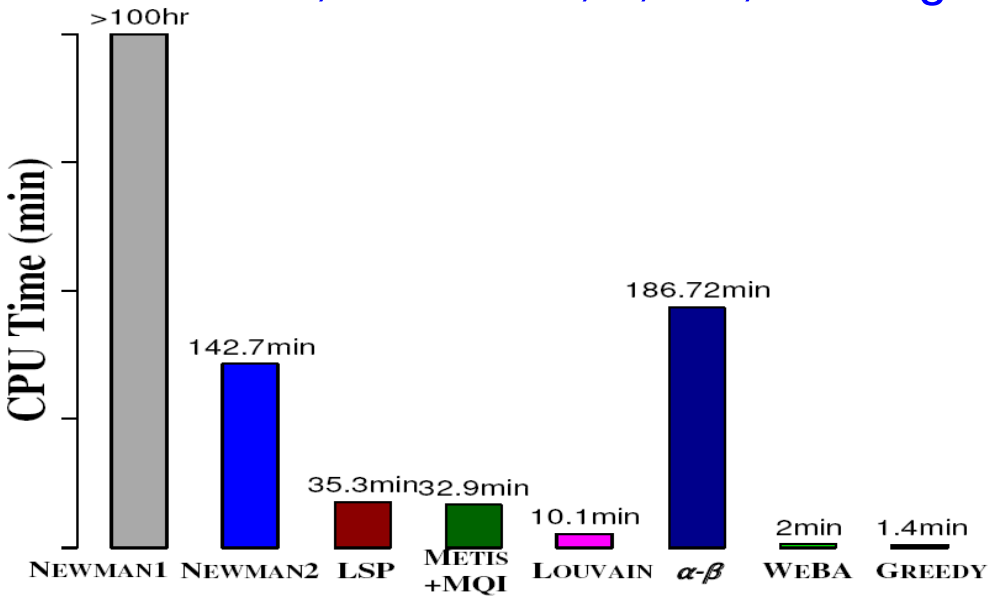
87%

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

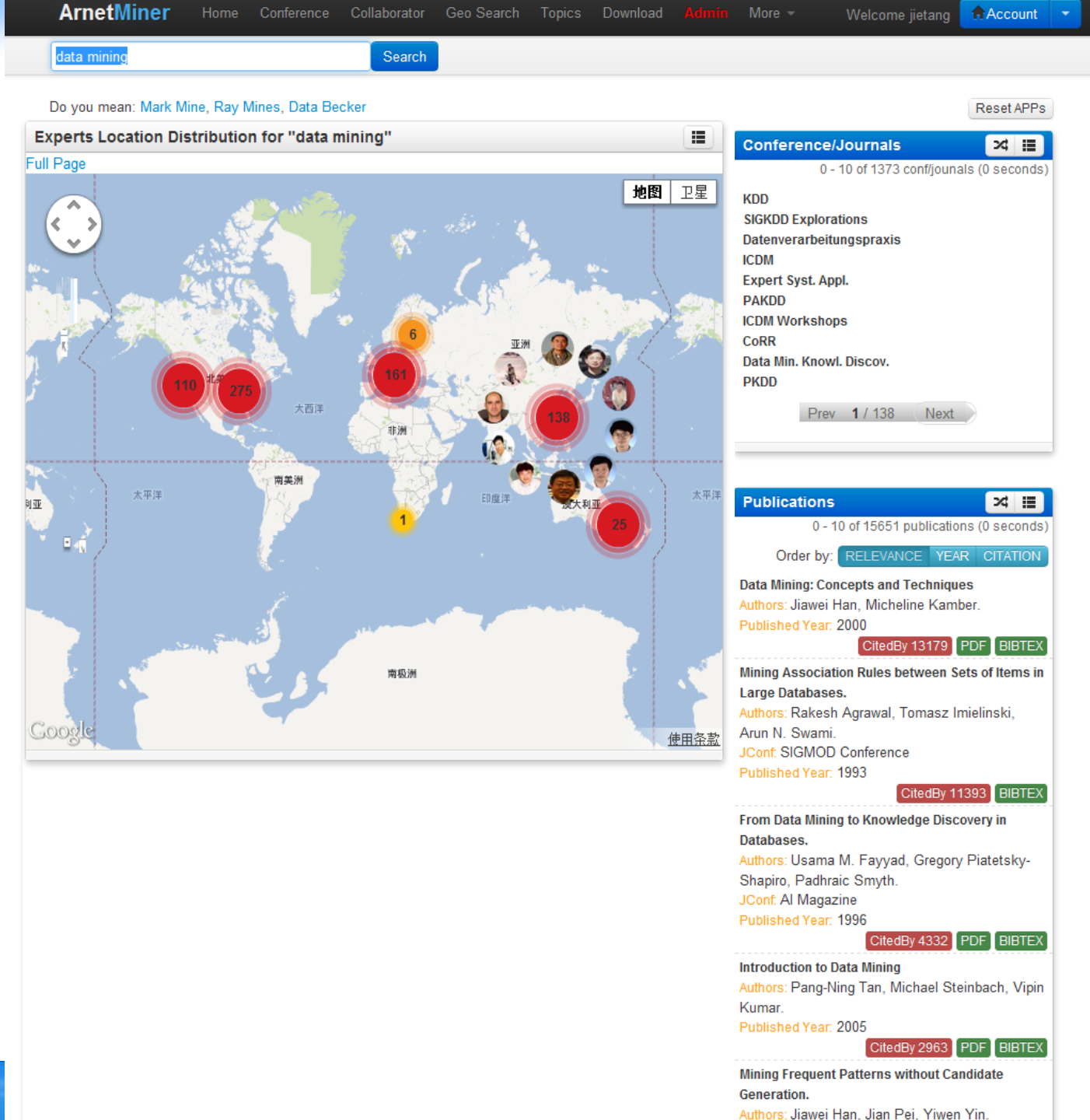
Widely used..

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

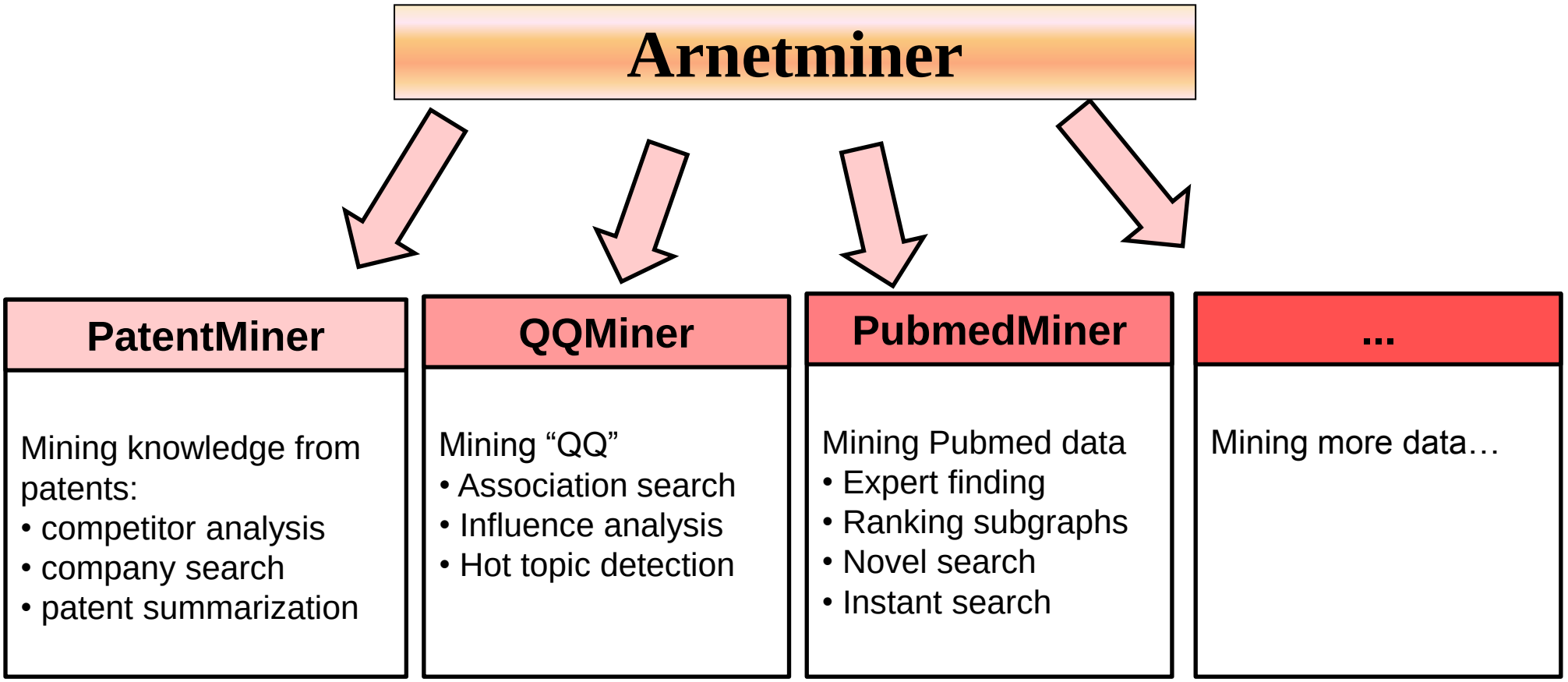
ArnetApp Platform

---to deploy your apps on Arnetminer.org

To customize the search on Arnetminer.

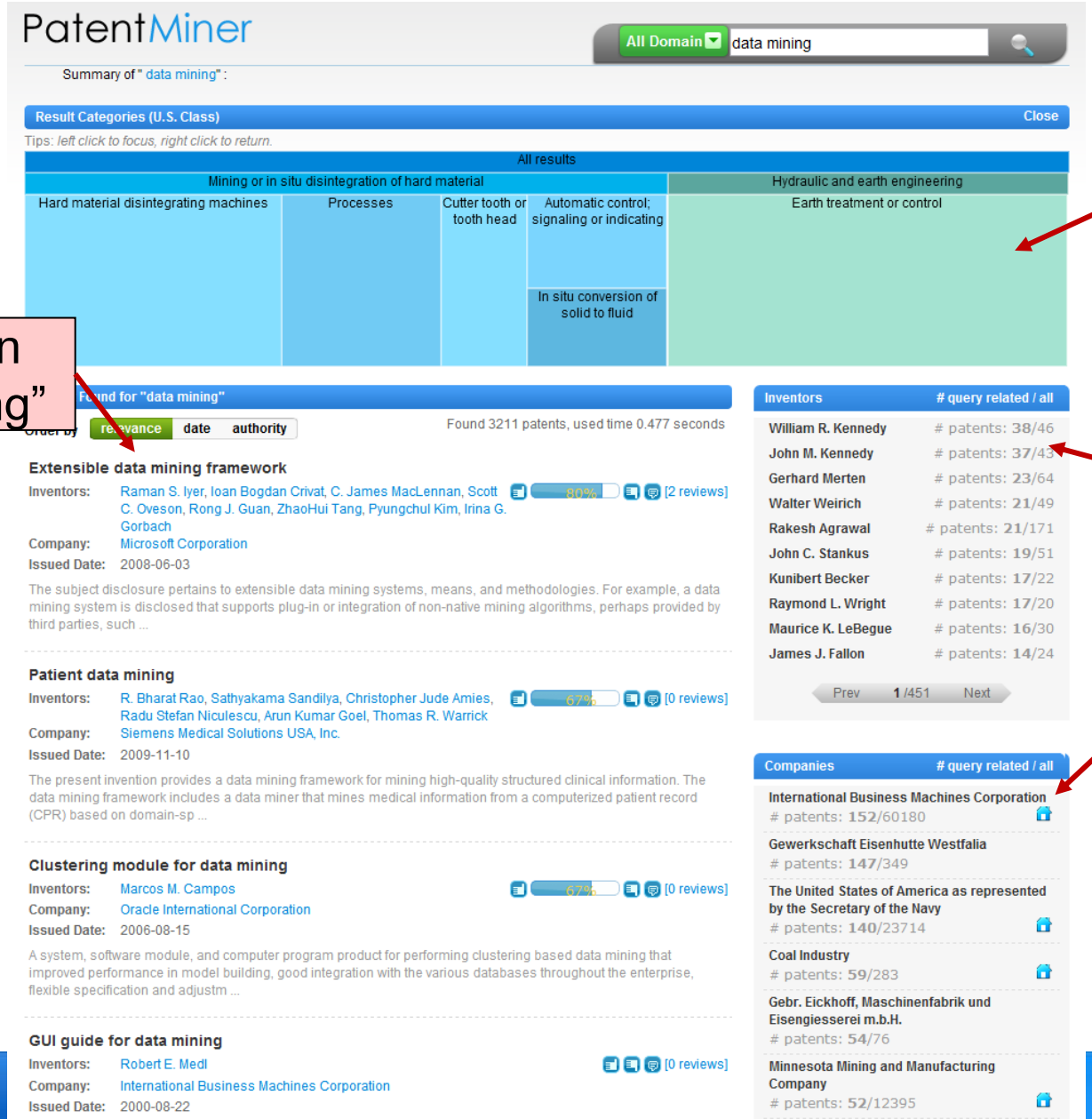


Arnetminer as a platform…



PatentMiner

Patent Search



Related sub topics

Patents on "data mining"

Inventor

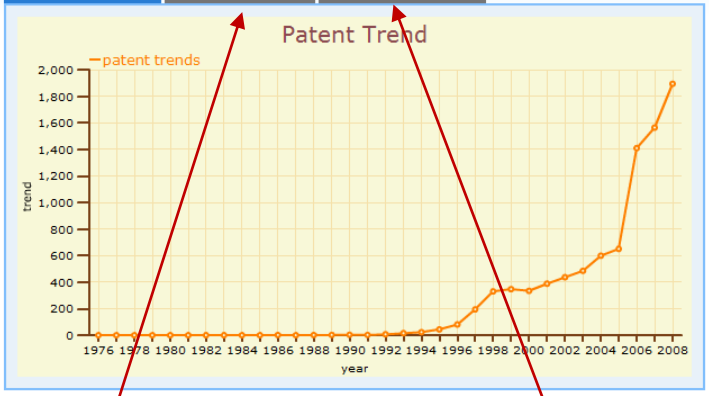
Company

PatentMiner

Microsoft Corporation

Trend Analysis:

Patent Trends | Topic Trends | Inventor Trends



Company Name: Microsoft Corporation

Patent Count: 13885

Website: microsoft.com

50%

[0 reviews]

Microsoft

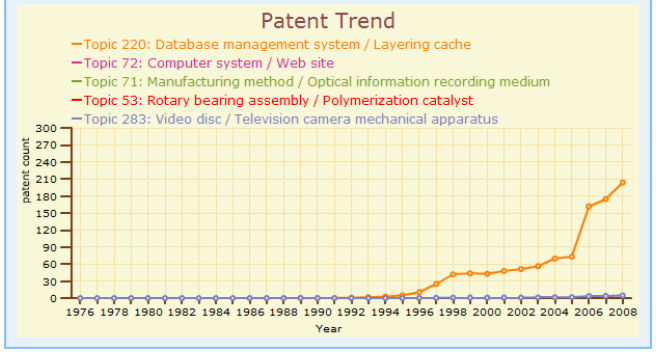
Competitors:

Global | Evolution | Topic | Topic Evolution

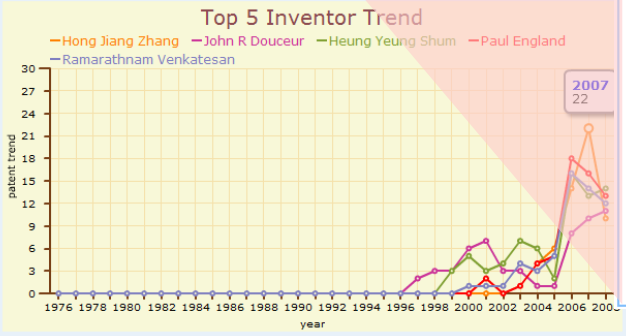
Heath Company
patents: 34

General Motors Corporation
patents: 10384

Patent Trends | Topic Trends | Inventor Trends



Patent Trends | Topic Trends | Inventor Trends



Competitors:

Global | Evolution | Topic | Topic Evolution

Find query-related competitor evolution for "Microsoft Corporation"

mobile phone search

1996--2000

Research In Motion Limited
patents: 887

Silverbrook Research Pty Ltd
patents: 2831

Nokia Corporation
patents: 3869

International Business Machines Corporation
patents: 56920

Hewlett-Packard Development Company, L.P.
patents: 11035

2001--2005

Research In Motion Limited
patents: 887

Silverbrook Research Pty Ltd
patents: 2831

Nokia Corporation
patents: 3869

International Business Machines Corporation
patents: 56920

Hewlett-Packard Development Company, L.P.
patents: 11035

2006--2010

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

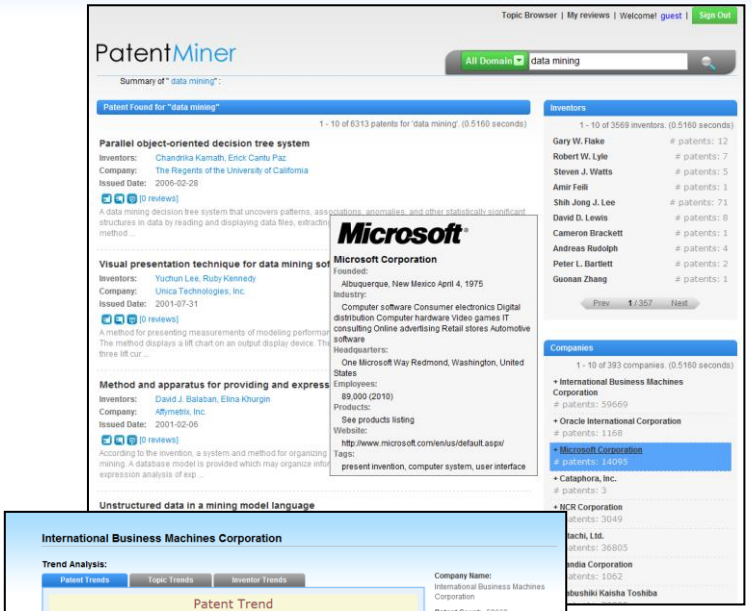
Tsinghua University

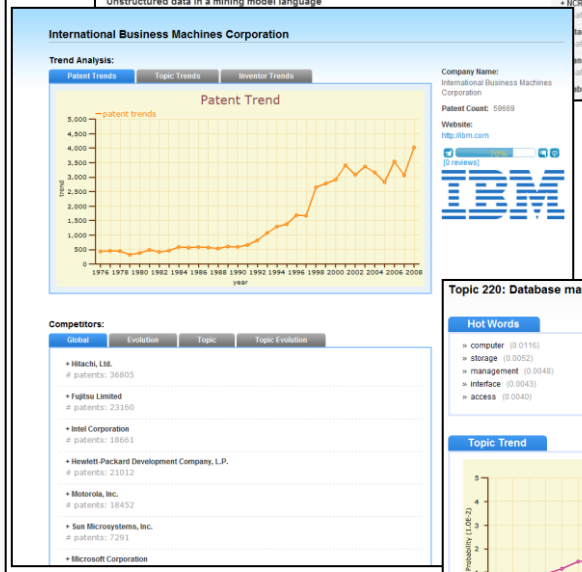
PatentMiner Today

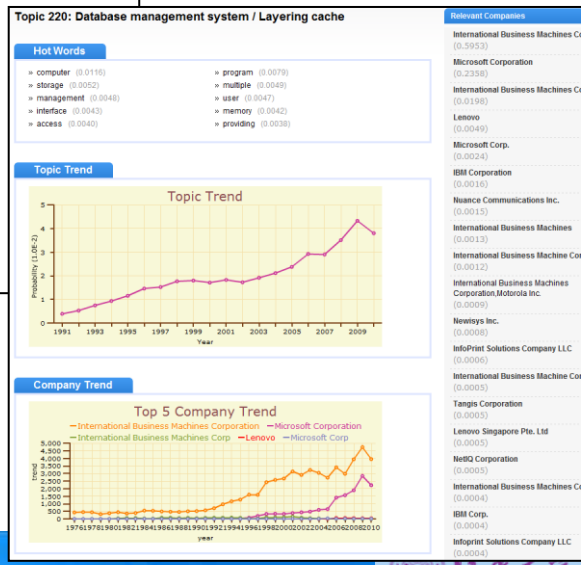
* Patent data:
  - > 3.8M patents
  - > 2.4M inventors
  - > 400K companies
  - > 10M citation relationships

* Journal data:
  - > 2k journal papers
  - > 3.7k authors

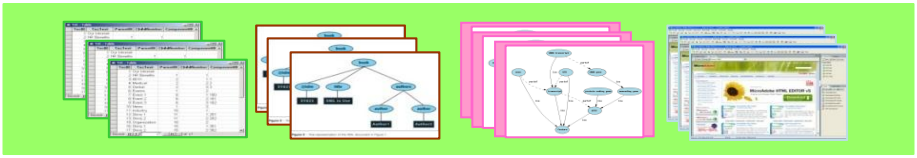
The crawled data is increasing to >300 Gigabytes.







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://arnetminer.org

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