Complex Social Network Mining
—Theory, Methodologies, and Applications

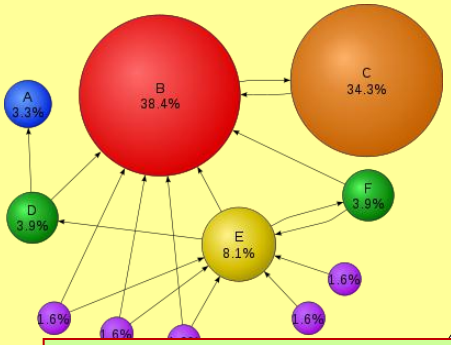
Jie Tang

Department of Computer Science and Technology
Tsinghua University
Email: jietang@tsinghua.edu.cn

Social Networks

Web 1.0 (1989)

Pages, hyperlinks
Relevance search



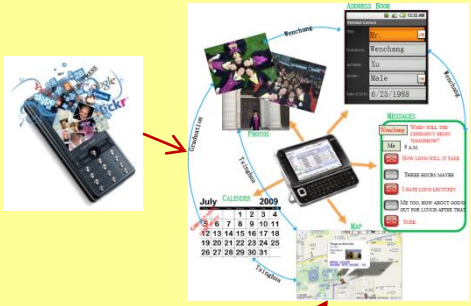
Web 2.0 (2004)

social networks

facebook

Mobile Web (2008-20)

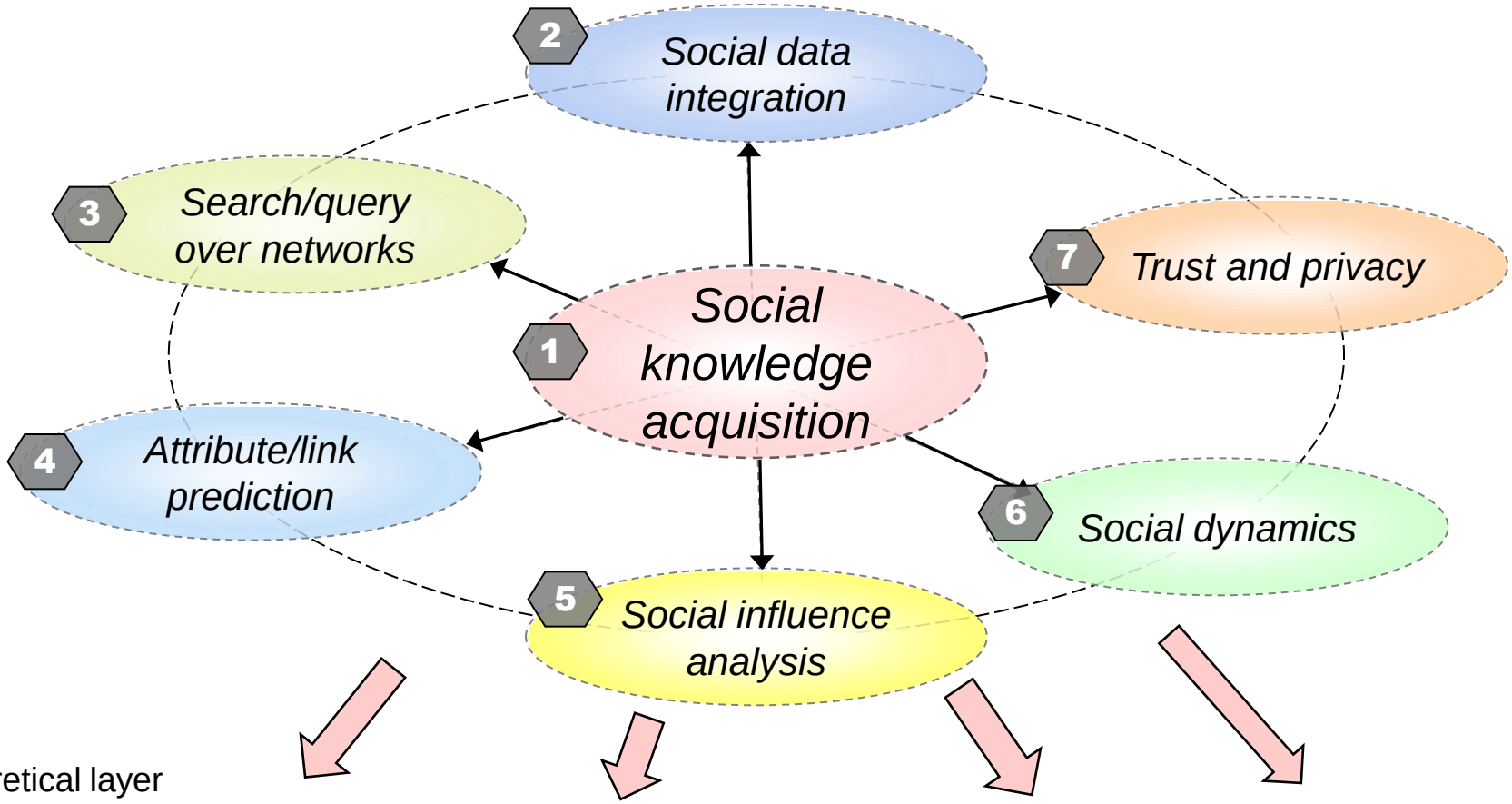
Connecting via
mobiles…



Web-based (or mobile-based) social networks already become a bridge to connect our real daily life and the virtual web space

Web-based Social Network Mining

—Theory, Methodologies, and Applications



Theoretical layer

Collective learning	Learning from users	Social theory	Graphical models

Web1.0: Web of Pages → Web 2.0: Web of People → Web 3.0: Web of Semantics

Outline

- ArnetMiner: Academic Social Network
- Core Techniques
  - Knowledge Acquisition
  - Semantic Integration
  - Heterogeneous Ranking
  - Social Influence Analysis
- Demo

ArnetMiner.org

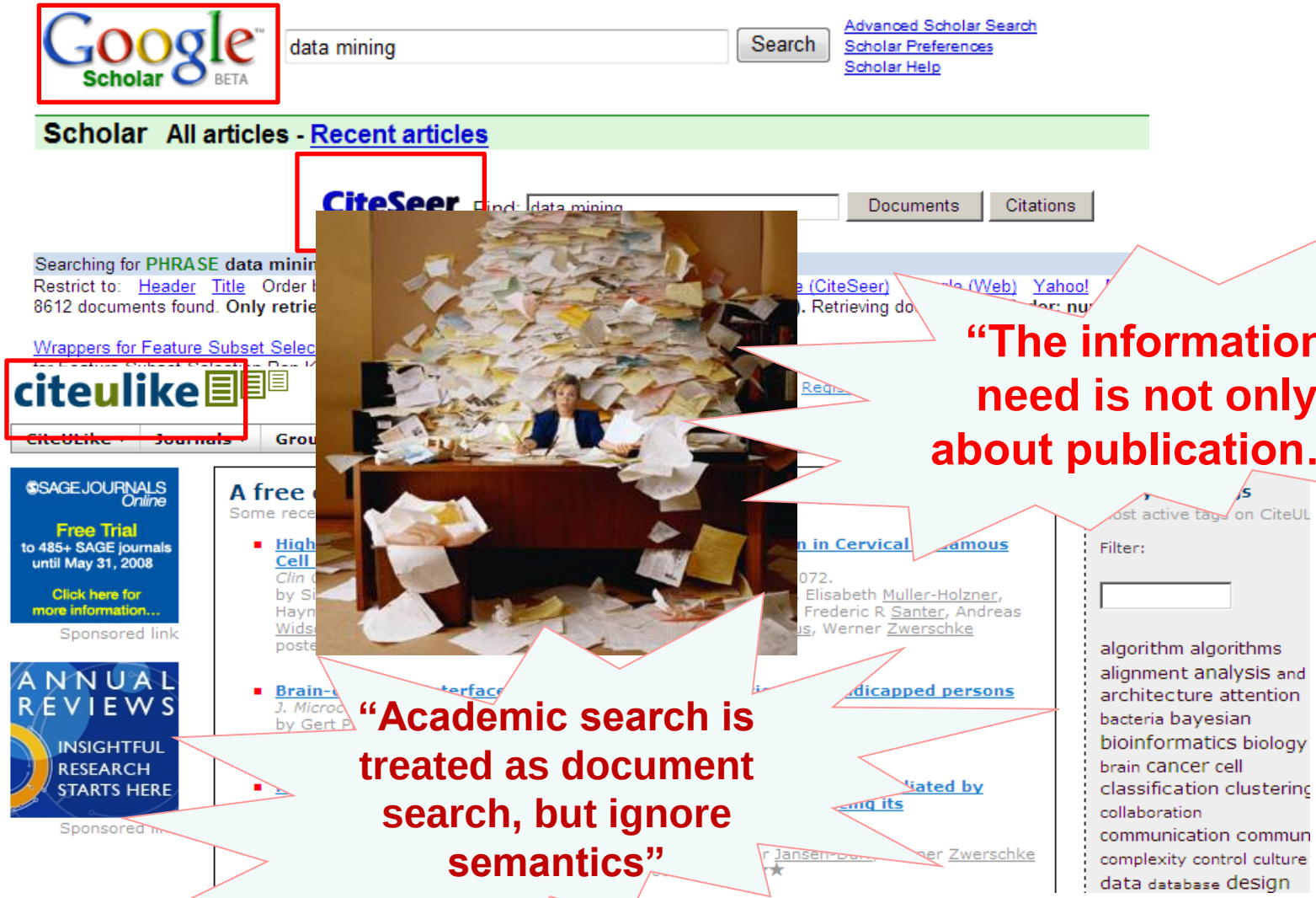
- Academic research social network analysis and mining system

提供全面的研究者网络分析与挖掘功能

Papers published: ACM TKDD, KDD’08-10, SDM’09, ICDM’07-09, CIKM’07-09, DKE, JIS

http://arnetminer.org/

Why Arnetminer.org?



"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.

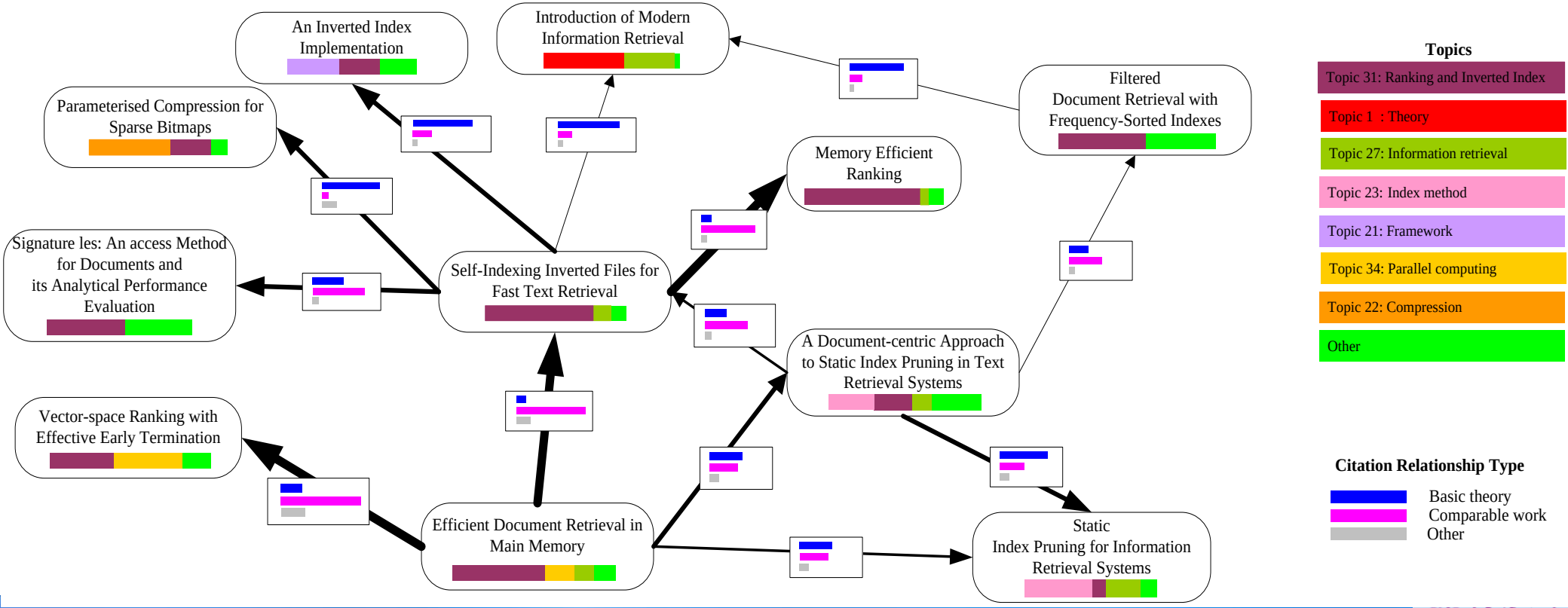
Researcher A

- **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?



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

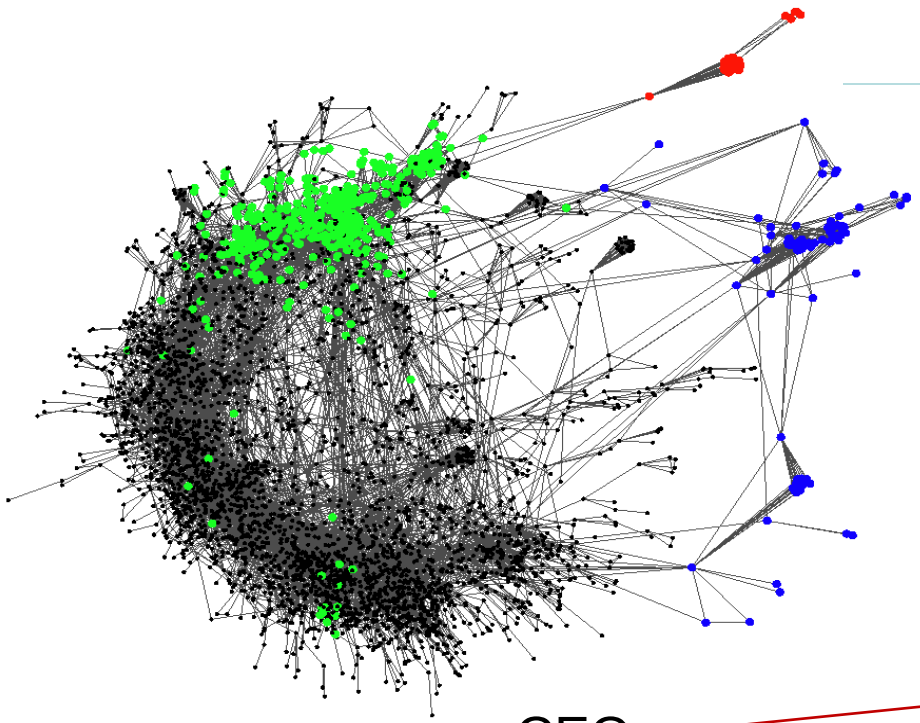
KDD Committee

conference

Who are best matching reviewers for each paper?

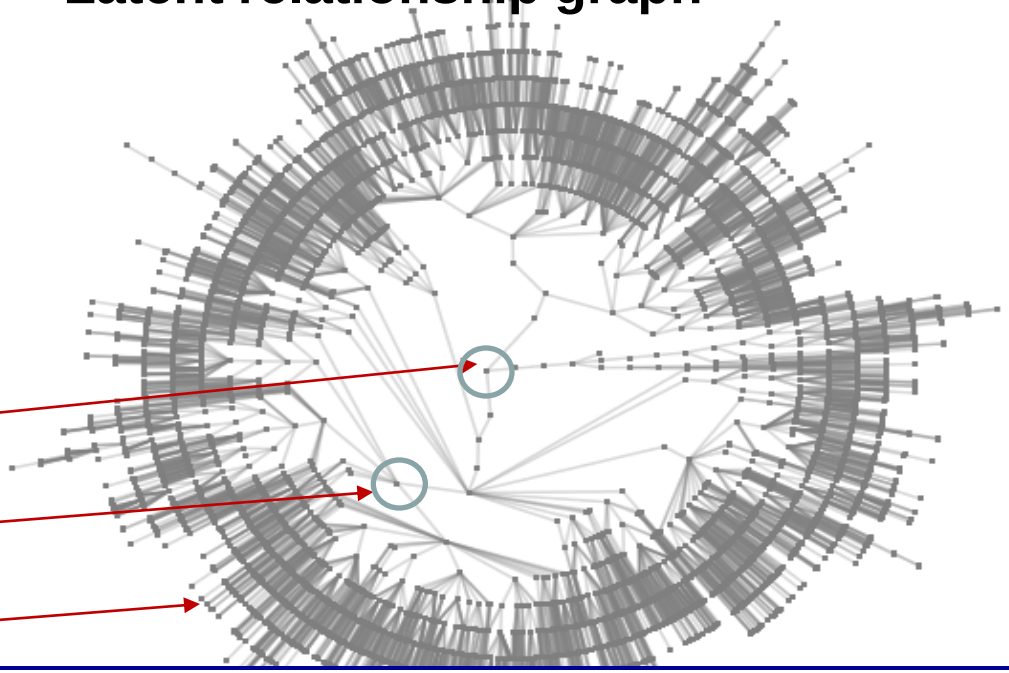
Paper content

Our Social Network is Black White



Social network without role/relationship info, e.g. a company's email network

Latent relationship graph

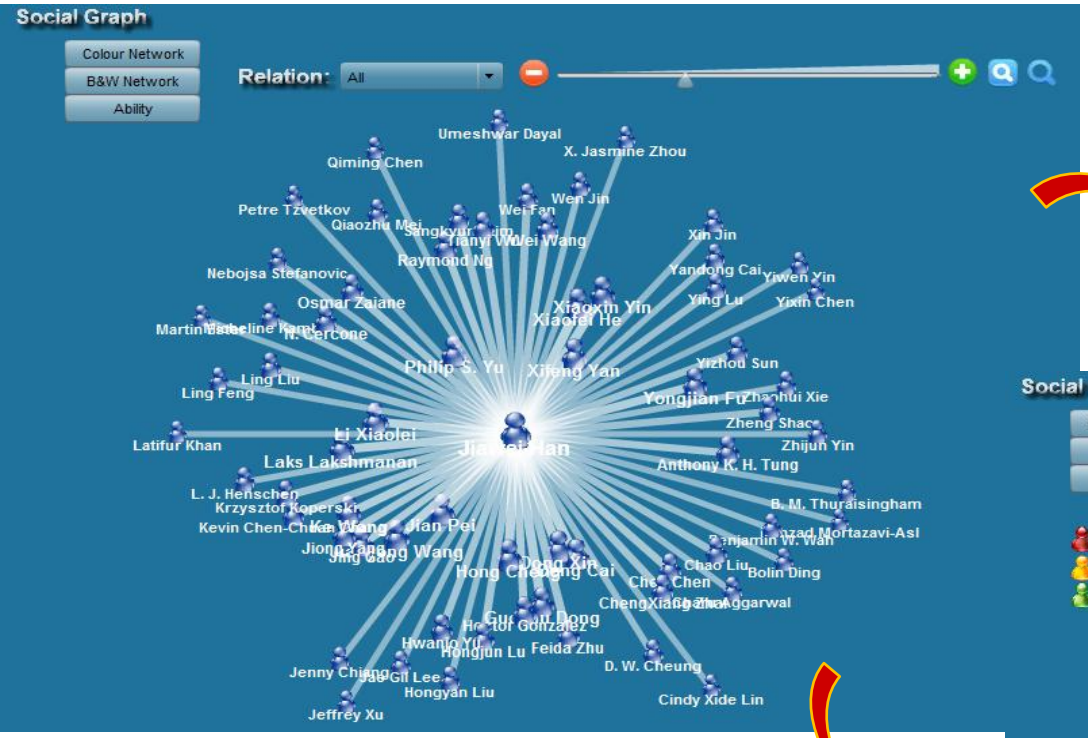


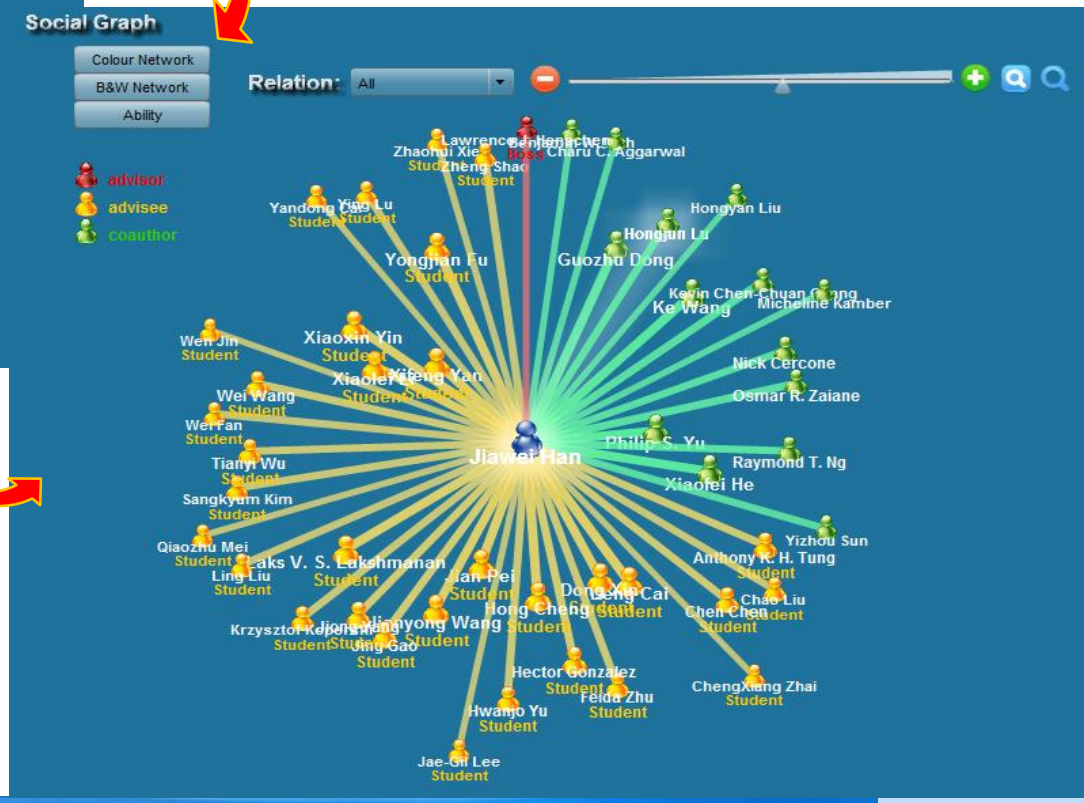
How to infer

- CEO
- Manager
- Employee

Fortunately, user interactions form *implicit groups*

From BW to Color

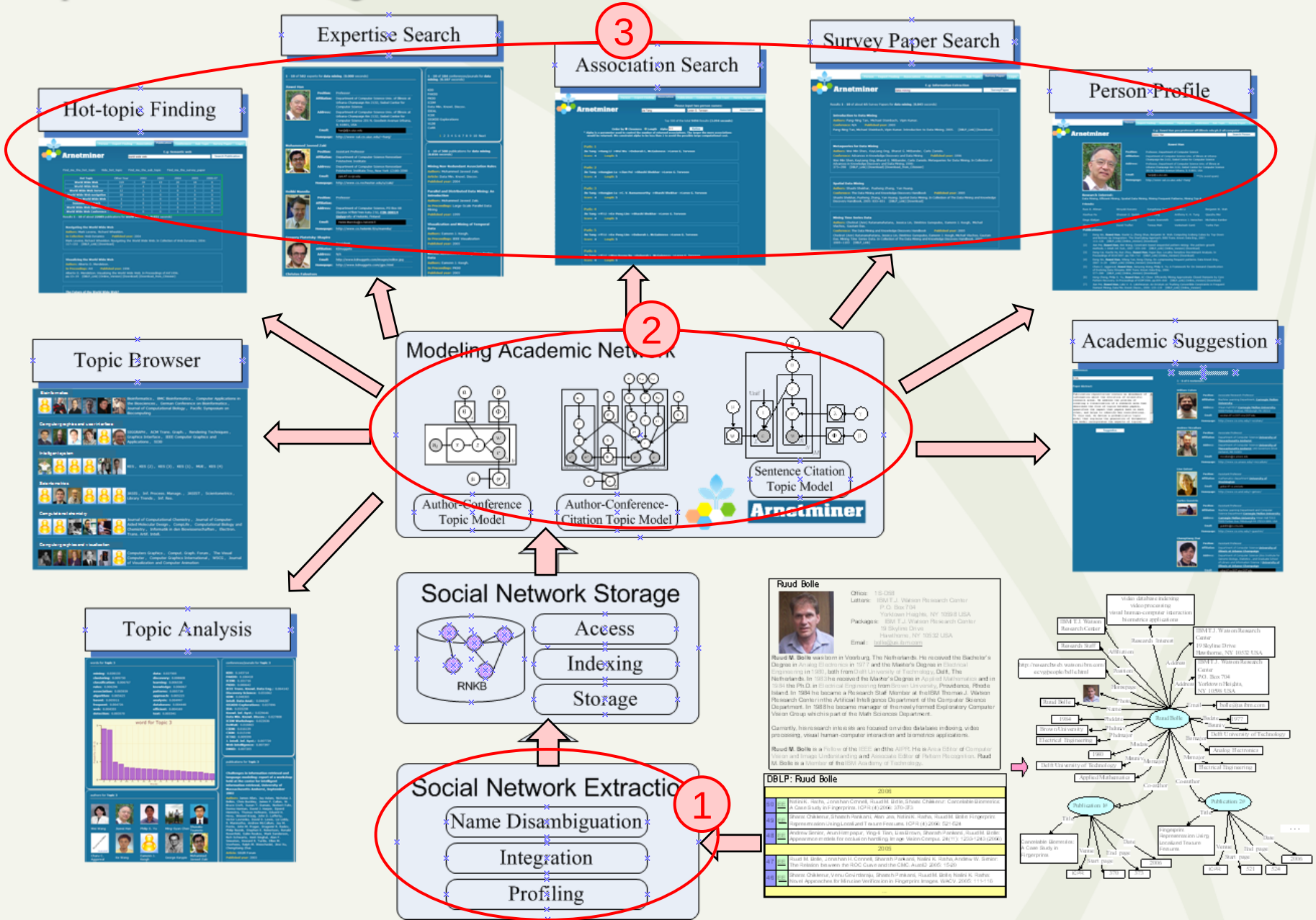




Jie Tang, Jing Zhang, Feng Wang, Limin Yao, Duo Zhang, Liu Liu, and Juanzi Li
Knowledge Engineering Group, DCST, Tsinghua University
{tangjie, zhangjing, wangfeng, ylm, zhangduo, ljz}@keg.cs.tsinghua.edu.cn

Arnetminer

Http://www.arnetminer.org



Person Search

Basic Info.

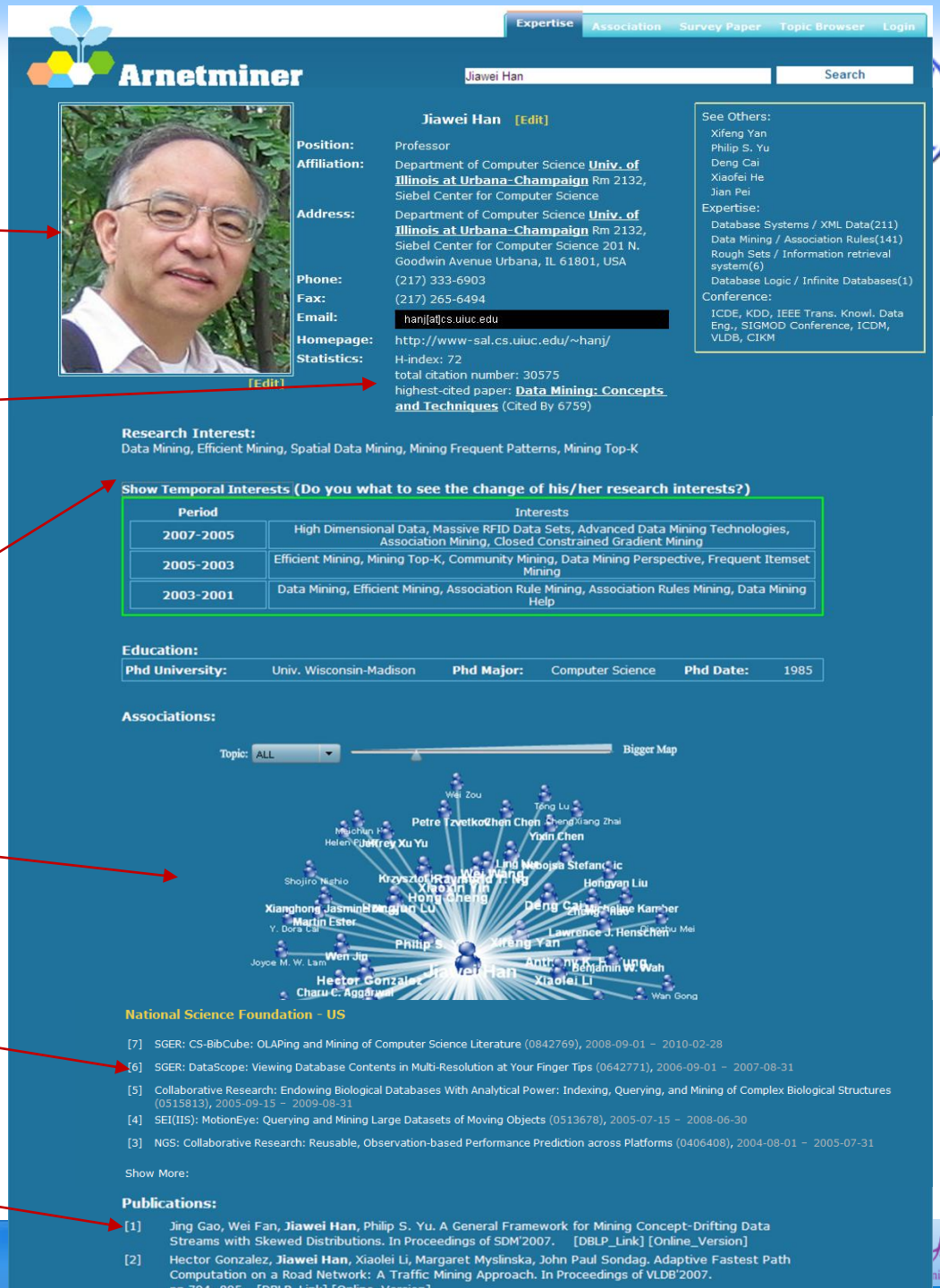
Citation statistics

Research Interests

Social Network

Fundings

Publications



Expertise Search

Finding experts, expertise conferences, and expertise papers for “information retrieval”

Did you mean author: *Zanthe Information*

Rank Algorithm: By VSM, **By ACT**

1 - 15 of **500** experts for **information retrieval**. (**0.297** seconds)

Gerard Salton (Professor, Harvard University Engineering College Library Committee)

H-index: 32; Papers: 138; Citation: 5579 [FOAF]

highest-cited paper: Document Length Normalization (1996) at Inf. Process. Manage.(Cited By 726)

Homepage: http://www.cs.cornell.edu/Info/Department/Annual95/Faculty/Salton.html

Expertise: Information Retrieval / Probabilistic Indexing; Computer / Systems;

W. Bruce Croft (Distinguished Professor, University of Massachusetts Amherst)

H-index: 56; Papers: 203; Citation: 14073 [FOAF]

highest-cited paper: A Language Modeling Approach to Information Retrieval (1998) at SIGIR(Cited By 1129)

Homepage: http://ciir.cs.umass.edu/personnel/croft.html

Expertise: Information Retrieval / Probabilistic Indexing; Database Systems;

C. J. van Rijsbergen (Professor, Information Retrieval Group Department of Computing Science University of Glasgow)

H-index: 23; Papers: 74; Citation: 2243 [FOAF]

highest-cited paper: A Non-Classical Logic for Information Retrieval (1986) at Comput. J.(Cited By 363)

Homepage: http://www.dcs.qla.ac.uk/~keith/

Expertise: Information Retrieval / Probabilistic Indexing;

Jian-Yun Nie (Full professor, RALI Lab Department IRO University of Montreal)

H-index: 22; Papers: 83; Citation: 2144 [FOAF]

highest-cited paper: Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web (1999) at SIGIR(Cited By 196)

Homepage: http://www.iro.umontreal.ca/~nie/index_Eng.html

Expertise: Information Retrieval / Probabilistic Indexing; Reutilisation de l / Une application; Natural Language System / Statistical Machine Translation;

James Allan (associate professor , Department of Computer Science University of Massachusetts)

H-index: 36; Papers: 113; Citation: 4186 [FOAF]

highest-cited paper: Automatic Query Expansion Using SMART: TREC 3 (1994) at TREC(Cited By 345)

Homepage: http://ciir.cs.umass.edu/~allan/

Expertise: Information Retrieval / Probabilistic Indexing; Natural Language System / Statistical Machine Translation; Digital library / Information Access;

1 - 10 of **500** conferences/jounals for **information retrieval**. (**0.281** seconds)

SIGIR
Inf. Process. Manage.
TREC
JASIST
Commun. ACM
JASIS
CIKM
RIAO
CLEF
Journal of Chemical Information and Computer Sciences

1 2 3 4 5 6 7 8 9 10 Next

1 - 10 of **500** publications for **information retrieval**. (**0.281** seconds)

Modern Information Retrieval [Bibtex] CitedBy 6720

Authors: Ricardo A. Baeza-Yates , Berthier A. Ribeiro-Neto .

Book:

Published year: 1999

A Language Modeling Approach to Information Retrieval [Bibtex] CitedBy 1129

Authors: Jay M. Ponte , W. Bruce Croft .

In Proceedings: SIGIR

Published year: 1998

Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval [Bibtex] CitedBy 706

Course Search

Finding courses for "data mining"

Arnetminer Course Search Result:
returns *1468* results for *`data mining`*

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

- **Data Mining and Text Mining**

Course Course Spring 2008: CS 583 - Data Mining and Text Mining Fall 2007: CS 583 - Data Mining and Text Mining Spring 2007: CS 583 - Data Mining and Text Mining Past teaching

by: Bing Liu (Professor)
Affiliation: Department of Computer Science University of Illinois at Chicago (UIC)
Homepage: http://www.cs.uic.edu/~liub/

- **Data Mining**

Course CISC-304 (Logic and Programming) CISC-483-683 (Data Mining)

by: Sandra Carberry (Professor)
Affiliation: Department of Computer and Information Sciences Joint Appointment in Department of Lingustics University of Delaware
Homepage: http://www.cis.udel.edu/~carberry/

- **CS701: Database Systems and Design CS705: Introduction to Data Mining**

Course CS701: Database Systems and Design CS705: Introduction to Data Mining (this is a derivative of CS499/699: Data Mining (course website)) CS400/600: Data Structures and Software Design CS405/605: Introduction to Database Management Systems CS466/666: Introduction to Formal Languages

by: Guozhu Dong (Professor)
Affiliation: Data Mining Research Lab Dept of Computer Sci & Engr Wright State University
Homepage: http://www.cs.wright.edu/~gdong/

- **Data Mining: Principles and Algorithms**

Course Fall, 2009 CS290D - Introduction to Data Mining Spring, 2009 CS290D - Data Mining: Principles and Algorithms Winter, 2009 CS595D - Graph Mining, Weekly Seminar

by: Xifeng Yan (Research Staff Member)
Affiliation: IBM T.J.Watson Research Center
Homepage: http://www.xifengyan.net/

- **UIUC CS591Han Advanced Topics in Data Mining**

Course UIUC CS512: Data Mining: Principles and Algorithms (Spring 2009) 9:30-10:45am Tues/Thurs. 1131 SC, CRN: 31604 UIUC CS591Han Advanced Topics in Data Mining (Spring 2008) 4-5pm Thurs., 3403 SC, CRN: 35953/ Spring'09: 40247 UIUC CS591 Yahoo!-DAIS (Data and Information Systems

by: Jiawei Han (Professor)
Affiliation: Department of Computer Science Univ. of Illinois at Urbana-Champaign Rm 2132, Siebel Center for Computer Science
Homepage: http://www-sal.cs.uiuc.edu/~hanj/

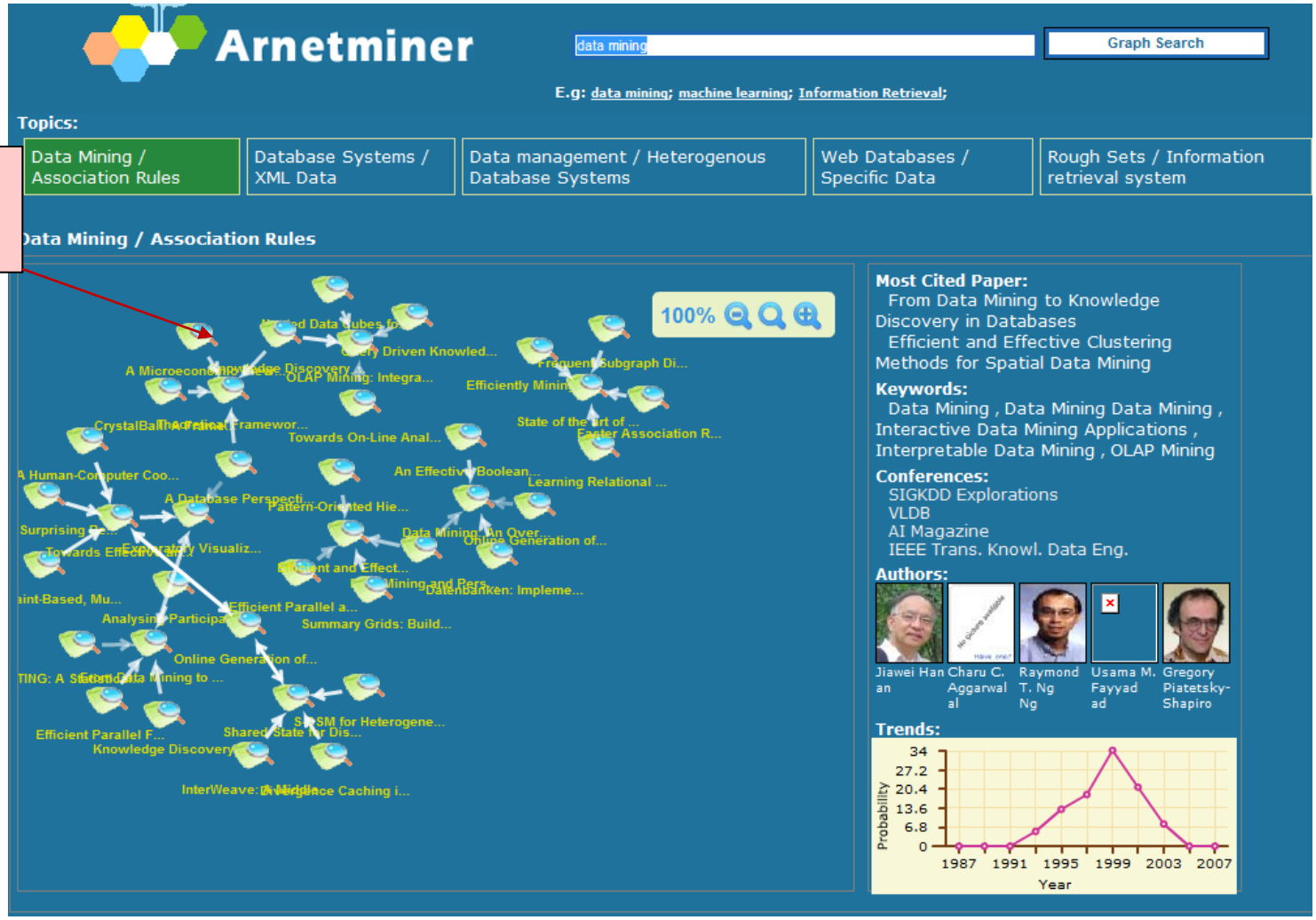
Association Search

Finding associations between persons
- high efficiency
- Top-K associations

Usage:
- to find a partner
- to find a person with same interests

Sub-Graph Search

Sub graphs



Topic Browser

Topic 13: Computing programming

Commun. ACM , IFIP Congress , IFIP Congress (2) ,
Comput. Surv. , Advances in Computers , IFIP Congre

Topic 14: Computer networks and communication

INFOCOM , IEEE/ACM Trans. Netw. , IEEE Journal on
Areas in Communications , SIGCOMM , Computer Netw
ICNP

Topic 15: Automatic reasoning

CADE , J. Autom. Reasoning , J. Log. Comput. , JELIA
Math. Artif. Intell. , KR

Topic 16: Knowledge discovery and data mining

KDD , PAKDD , ICDM , PKDD , IEEE Trans. Knowl. Da
, SDM

Topic 17: Human computer interaction

CHI Extended Abstracts , CHI , Interactions , INTERA
Interacting with Computers , CHI Conference Compani

Topic 18: Evolvable Hardware

Evolvable Hardware , ICES , AHS , BioADIT , Conf.
Computing Frontiers , UC

200 topics have been discovered automatically from the academic network

Hot words in **Topic 16: Knowledge discovery and data mining**

mining:	0.041622	**data:**	0.041032
clustering:	0.014825	**discovery:**	0.010956
learning:	0.010630	**classification:**	0.010569
rules:	0.008899	**association:**	0.008655
patterns:	0.007962	**knowledge:**	0.007820
approach:	0.007636	**algorithm:**	0.007596
frequent:	0.007473	**based:**	0.007311
analysis:	0.007209	**databases:**	0.006211
efficient:	0.006129	**web:**	0.005417
text:	0.004908	**detection:**	0.004887

Hot conferences/jounals in **Topic 16: Knowledge discovery and data mining**

KDD:	0.143204
PAKDD:	0.106370
ICDM:	0.100902
PKDD:	0.082549
IEEE Trans. Knowl. Data Eng.:	0.064238
SDM:	0.047920
Intell. Data Anal.:	0.044338
SIGKDD Explorations:	0.037705
IDA:	0.033127
Knowl. Inf. Syst.:	0.030880
Data Min. Knowl. Discov.:	0.027447
DaWaK:	0.024374
CIKM:	0.022848
ICDM Workshops:	0.022806
Discovery Science:	0.020771
MLDM:	0.017677
CIDM:	0.016194
J. Intell. Inf. Syst.:	0.008055
KDID:	0.007081
DMKD:	0.007059

Active authors in **Topic 16: Knowledge discovery and data mining**

Jiawei Han, Eamonn J. Keogh, Philip S. Yu, Mohammed Javeed Zaki, Ming-Syan Chen, Wei Wang, Srinivasan Parthasarathy, Charu C. Aggarwal, Qiang Yang, Shusaku Tsumoto

Data Mining	Association Rules
data stream	Association Mining
Data Mining Techniques	Data Streams
Evolving Data Streams	Mining Association Rules
Mining Time Series Data	data mining application

Topic Trend in **Topic 16: Knowledge discovery and data mining**

Publications in **Topic 16: Knowledge discovery and data mining**

Data Mining Techniques for Associations, Clustering and Classification
Authors: Charu C. Aggarwal, Philip S. Yu.
In Proceedings: PAKDD
Published year: 1999

Data Mining for the Web
Authors: Myra Spiliopoulou.
In Proceedings: PKDD
Published year: 1999

On string classification in data streams
Authors: Charu C. Aggarwal, Philip S. Yu.
In Proceedings: KDD
Published year: 2007

Matching in Time Series Databases
Authors: Eamonn J. Keogh, Padhraic Smyth.
In Proceedings: KDD
Published year: 1997

Knowledge Discovery and Data Mining: Towards a Unifying Framework
Authors: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth.
In Proceedings: KDD
Published year: 1996

Efficiently mining frequent trees in a forest
Authors: Mohammed Javeed Zaki.
In Proceedings: KDD
Published year: 2002

Mining of Evolving Data Streams with Privacy Preservation
Authors: Philip S. Yu.
In Proceedings: PAKDD

Academic Performance Measurement

Academic Statistics

g H-index, Total citation number, Uptrend, Activity, Longevity, Diversity, Sociability and New Star. We also list expert ranks for 200 different domains (topics), which are discovered using an author-conference-topic model, and is available here. More information about How we calculate academic statistics for an expert please click here.

If you find a bug, please Email to

Home | H-index | Citation

Top 3 by H-index: [?]

1 Hector Garcia-Molina (Prof

H-index: 95; Papers: 37
highest-cited paper: Th
By 1332)
Homepage: http://www

Expertise: XML Data / Relational D
Systems; Digital library; Extende

2 Christos H. Papadimitriou

H-index: 90; Papers:
highest-cited paper:
Homepage: http://w

Expertise: Optimal Parallel Algorit
Complexity Classes; Graph Transf

3 Anil K. Jain (Distinguished Pro

H-index: 83; Papers:
highest-cited paper:
Homepage: http://w

Expertise: Object Recognition / Mu
Image registration; Character Rec

Top 3 by Citation: [?]

Personal Statistics

Academic Statistic for Wendy Hall:

[Edit Basic Information], [Follow]
Head Professor , Computer Science at the University of Southampton School of Electronics and Computer Science (ECS) UK Prime Minister's Council for Science and Technology British Computer Society
Homepage: http://www.ecs.soton.ac.uk/~wh/

Expertise: Semantic web / Hypermedia Systems; Multimedia Applications / Content-Based Video Retrieval; Intelligent User Interfaces / Information Agents;

The result is not necessarily accurate. If you find a bug, please Email to us.

- In Computer Science:

1. H-index [?]	Top 2185		23	vs. Rank1: 95.0 (Hector Garcia-Molina)
2. Citation [?]	Top 2739		2315	vs. Rank1: 86803.0 (David E. Goldberg)
3. Uptrend [?]	Top 713264	96.9948%	-0.40	vs. Rank1: 28.5705 (Sumit Gulwani)
4. Activity [?]	Top 4865		4.60	vs. Rank1: 72.7481 (Philip S. Yu)
5. Longevity [?]	Top 12564		21	vs. Rank1: 67.0 (J. Barkley Rosser)
6. Diversity [?]	Top 135829		0.07	vs. Rank1: 1.78526 (R. Singh)
7. Sociability [?]	Top 514		218.69	vs. Rank1: 1057.92 (Wei Wang)

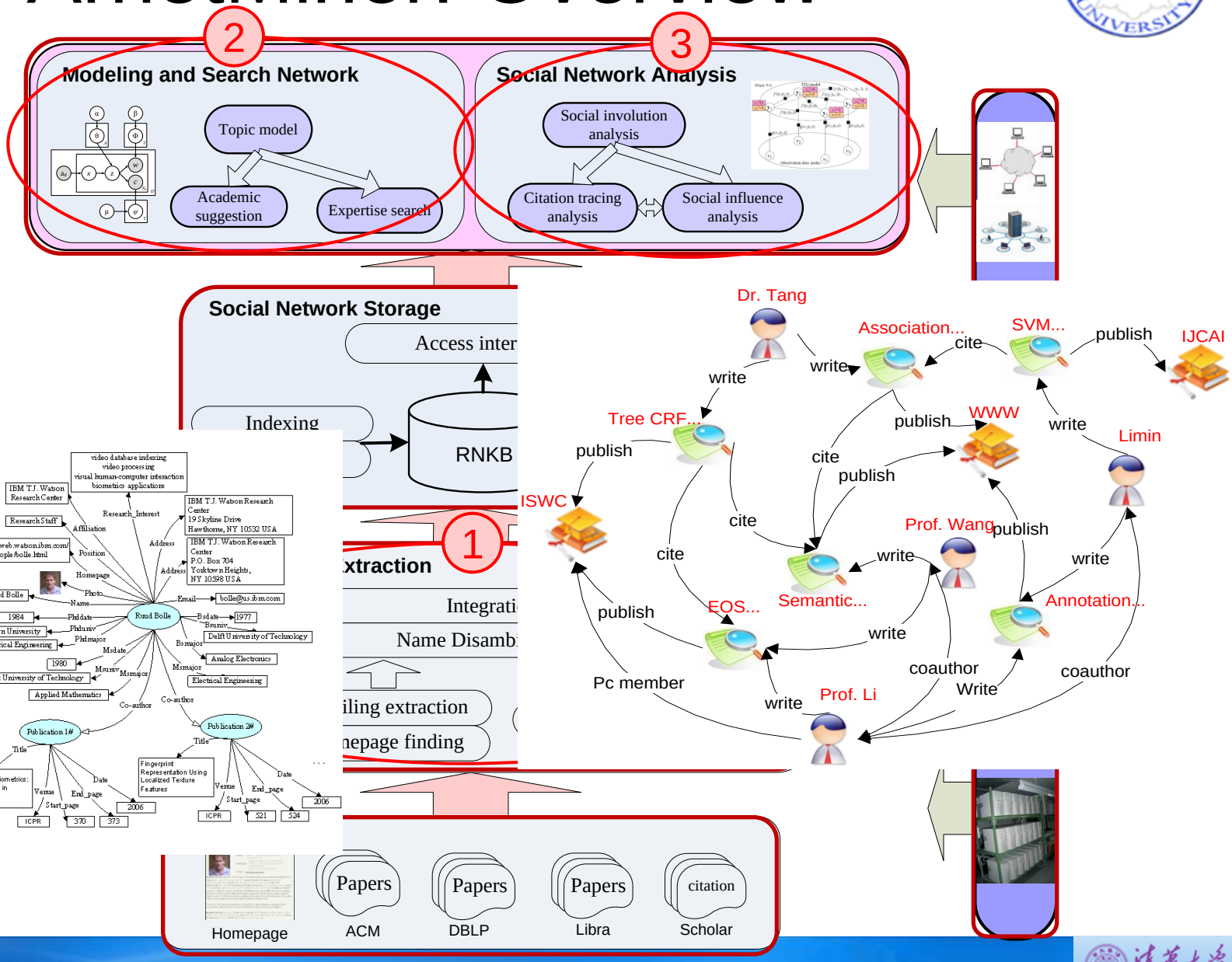
- AREA: Semantic web / Hypermedia Systems

1. H-index [?]	Top 54		23	vs. Rank1: 95.0 (Hector Garcia-Molina)
2. Citation [?]	Top 76		2315	vs. Rank1: 34981.0 (Hector Garcia-Molina)
3. Uptrend [?]	Top 3579	89.4974%	-0.40	vs. Rank1: 23.9716 (Ravi Kumar)
4. Activity [?]	Top 177		4.60	vs. Rank1: 41.6912 (Ravi Kumar)
5. Longevity [?]	Top 114		21	vs. Rank1: 41.0 (Gene H. Golub)
6. Diversity [?]	Top 2019	76%	0.07	vs. Rank1: 1.6156 (Gail E. Kaiser)
7. Sociability	Top 20		218.69	vs. Rank1: 461.368 (Hector Garcia-Molina)

Outline

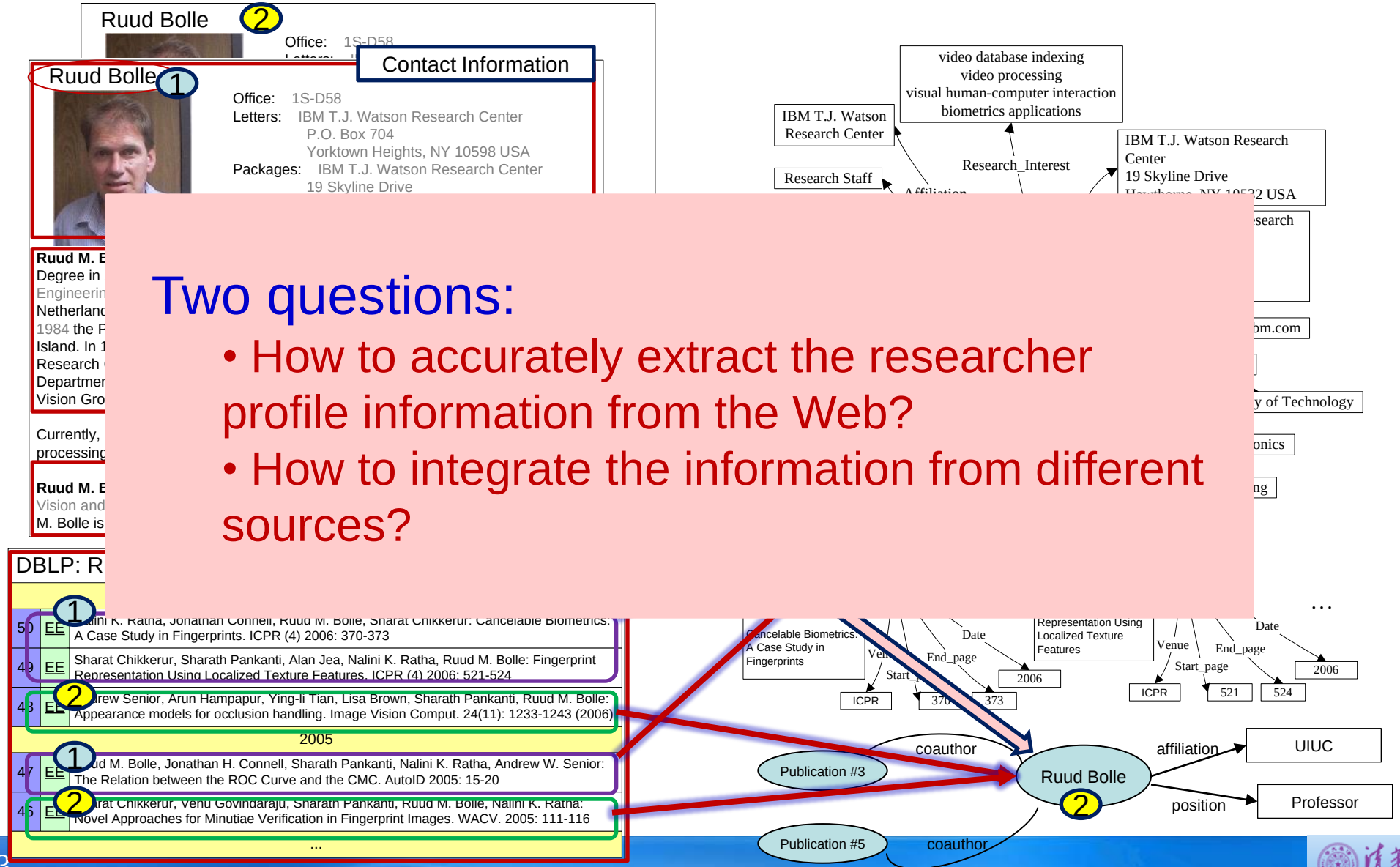
- ArnetMiner: Academic Social Network
- Core Techniques
  - Knowledge Acquisition
  - Semantic Integration
  - Heterogeneous Ranking
  - Social Influence Analysis
- Demo

ArnetMiner: Overview



CT1: Knowledge Acquisition from Social Web

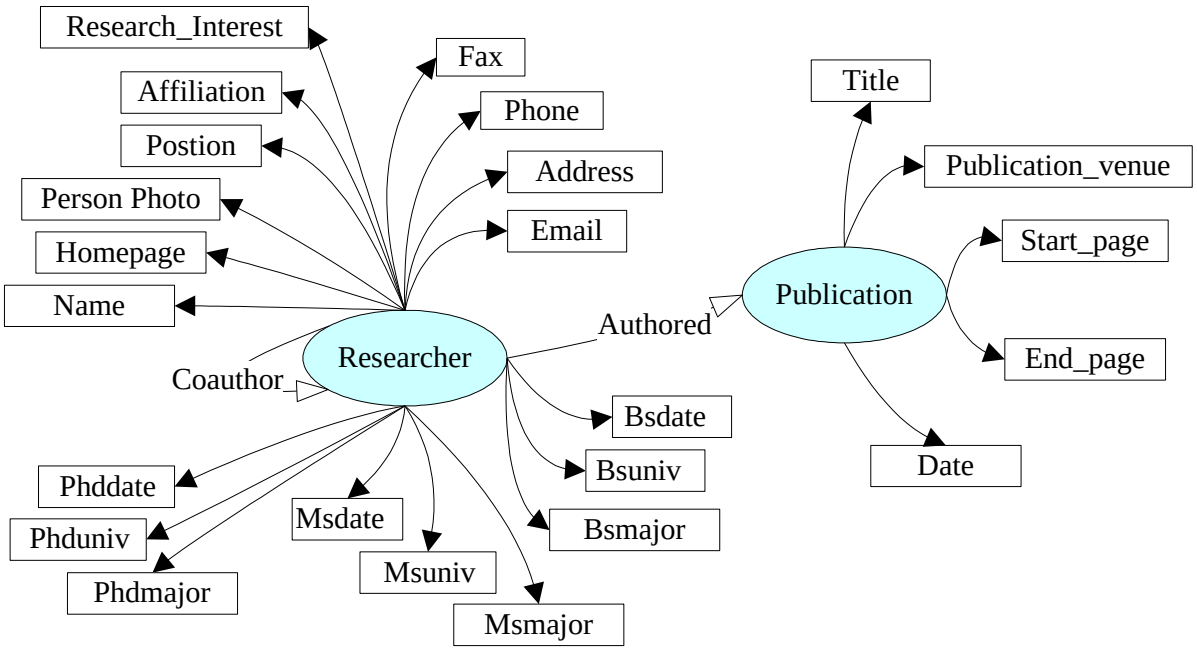
(ACM TKDD, ISWC'06, ICDM'07, ACL'07, CIKM'07-08)



Two questions:

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

Researcher Network Extraction



70.60% of the researchers have at least one homepage or an introducing page	
85.6% from universities	14.4% from companies
71.9% are homepages	28.1% are introducing pages
40% are in lists and tables	60% are natural language text

There are a large number of person names having the ambiguity problem
Even 3 “Yi Li” graduated from the author’s lab
70% moved at least one time

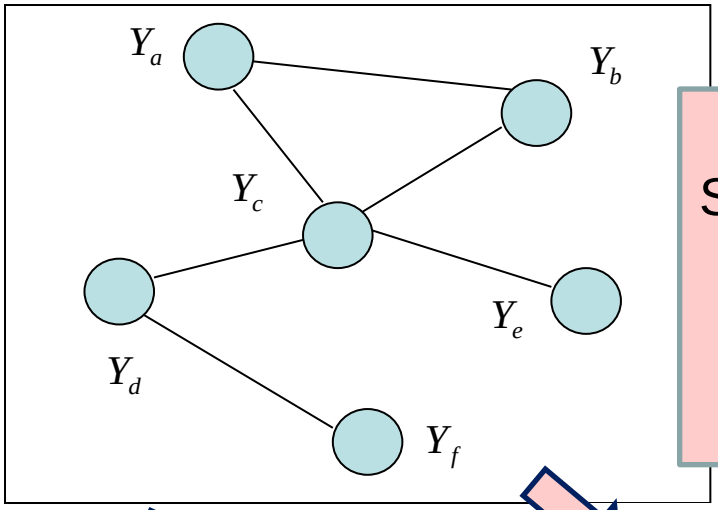
Our Approach Picture

– based on Markov Random Field

Markov Property:

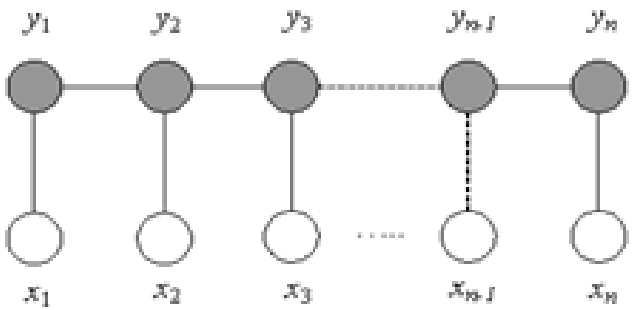
$$P(Y_i \mid \{Y_j \mid Y_j \neq Y_i\})$$

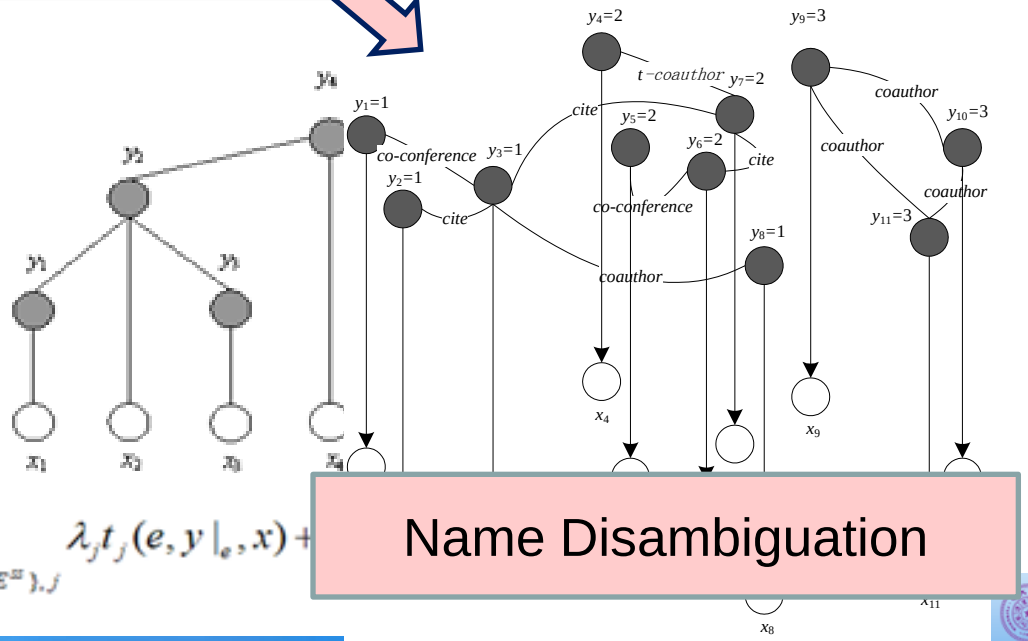
$$= P(Y_i \mid \{Y_j \mid Y_j \sim Y_i\})$$



Special cases:

- Conditional Random Fields
- Hidden Markov Random Fields





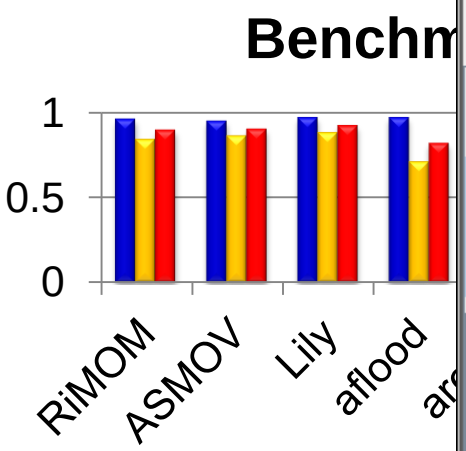
Researcher Profiling

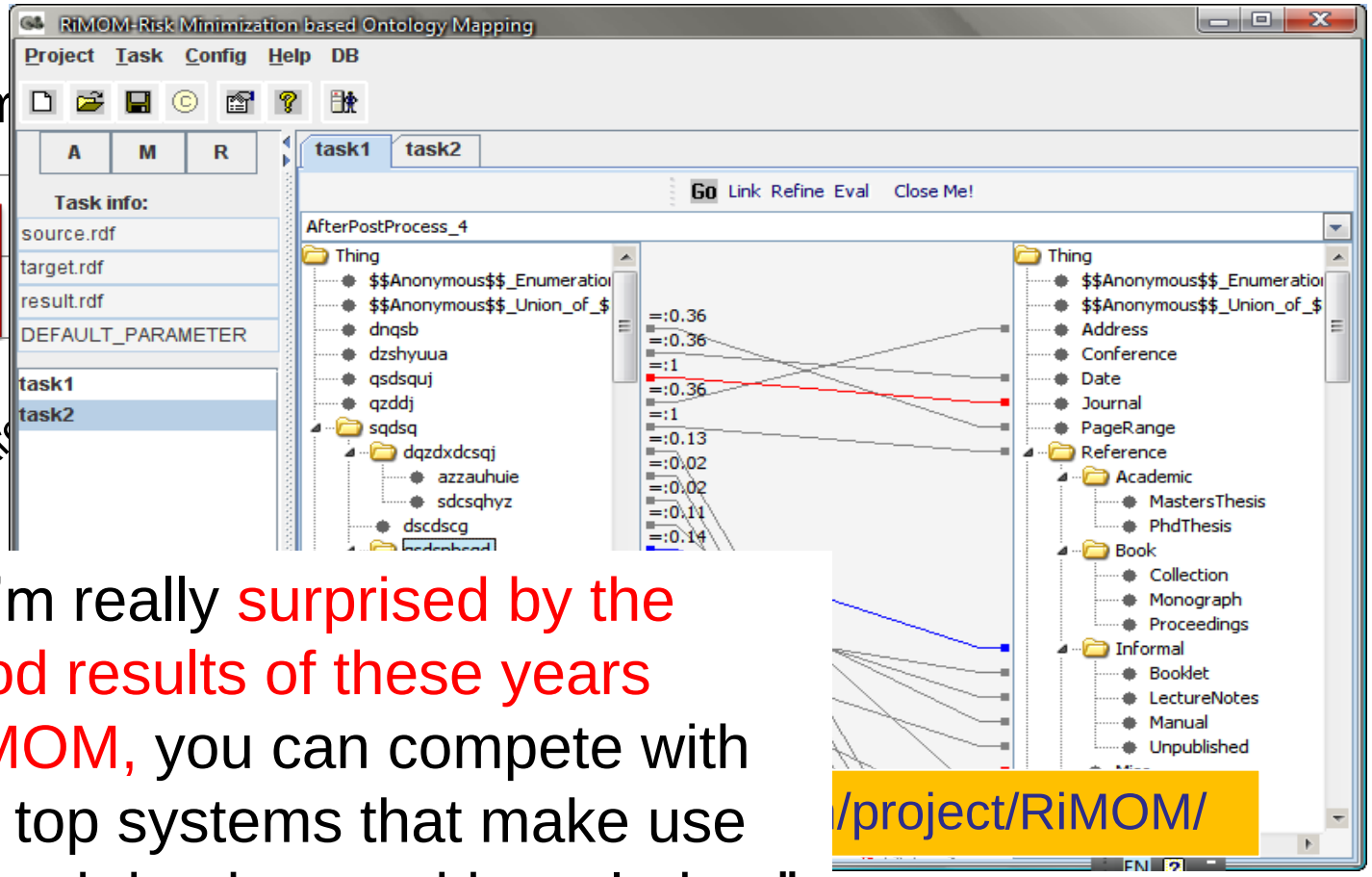
Name Disambiguation

CT2: Semantic Integration

(IEEE TKDE, SIGMOD'09, IJCAI'09, ISWC'09)

RiMOM-A Tool for Semantic Integration (OAEI'06-09)



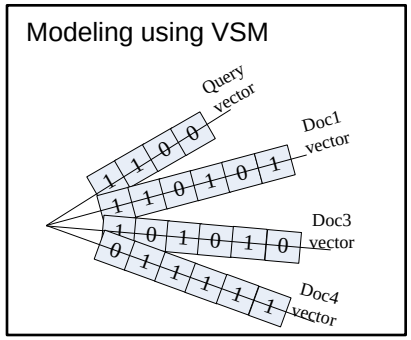


"I'm really surprised by the good results of these years RiMOM, you can compete with the top systems that make use of such background knowledge."

/project/RiMOM/

CT3: Topic-based Heterogeneous Ranking
(Machine Learn. J, KDD'08, ICDM'08, CIKM'09, DKE)

Search with keyword



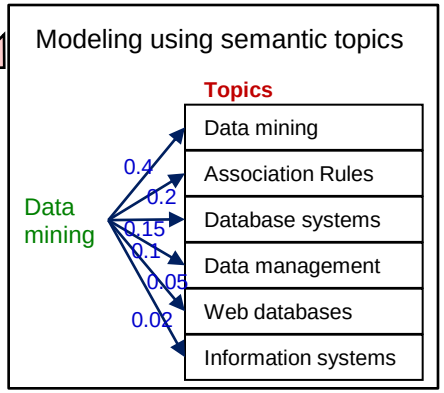
Return

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

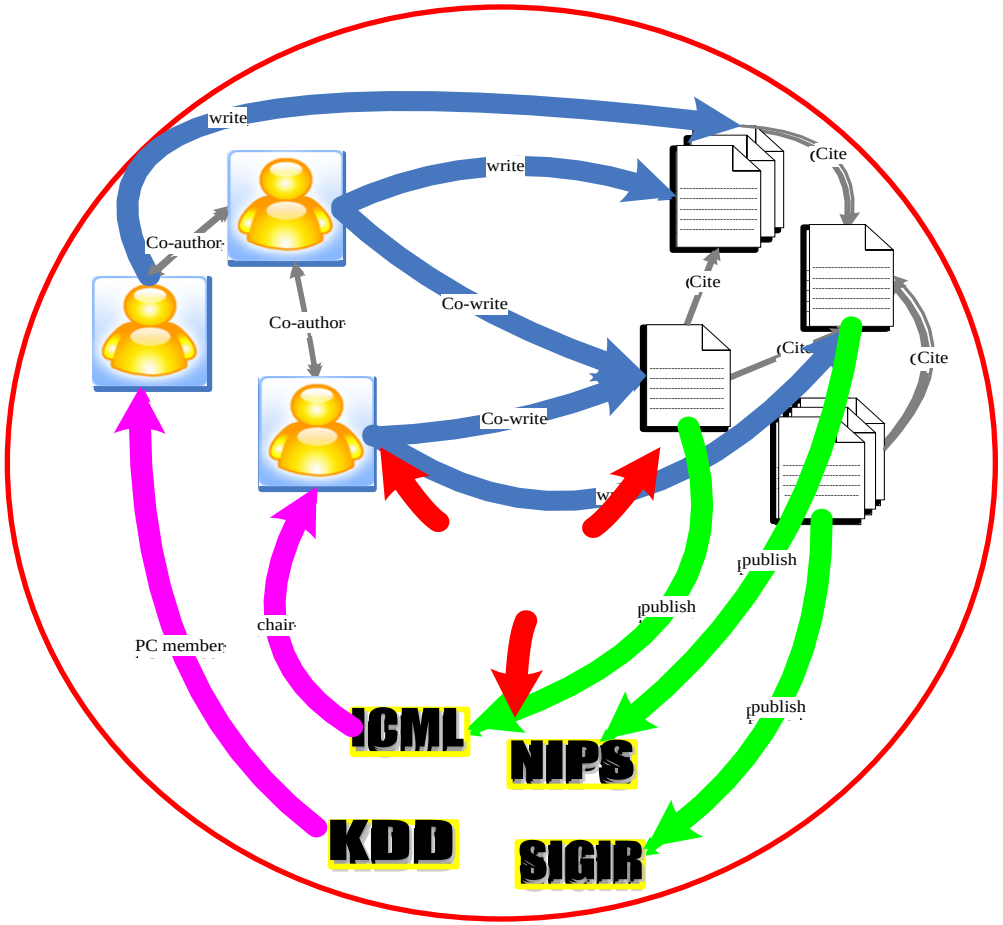


Return

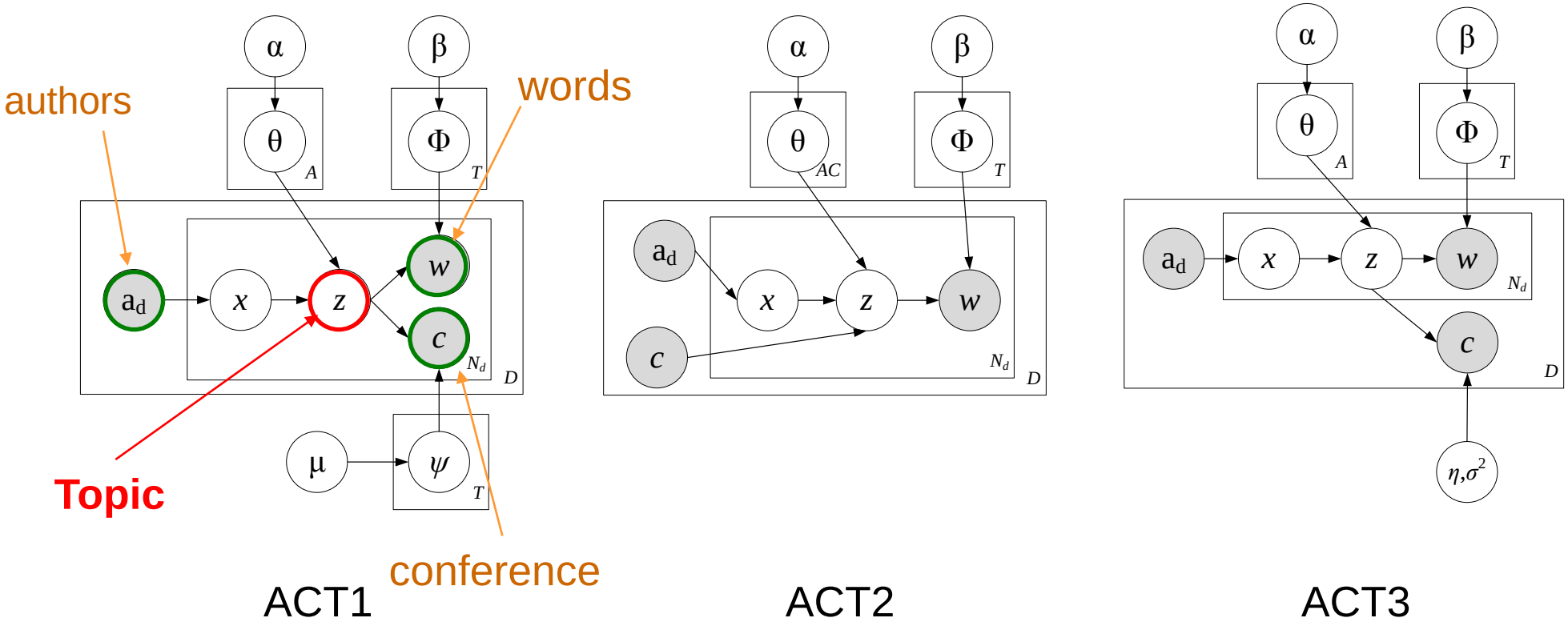
Challenges

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

Topic



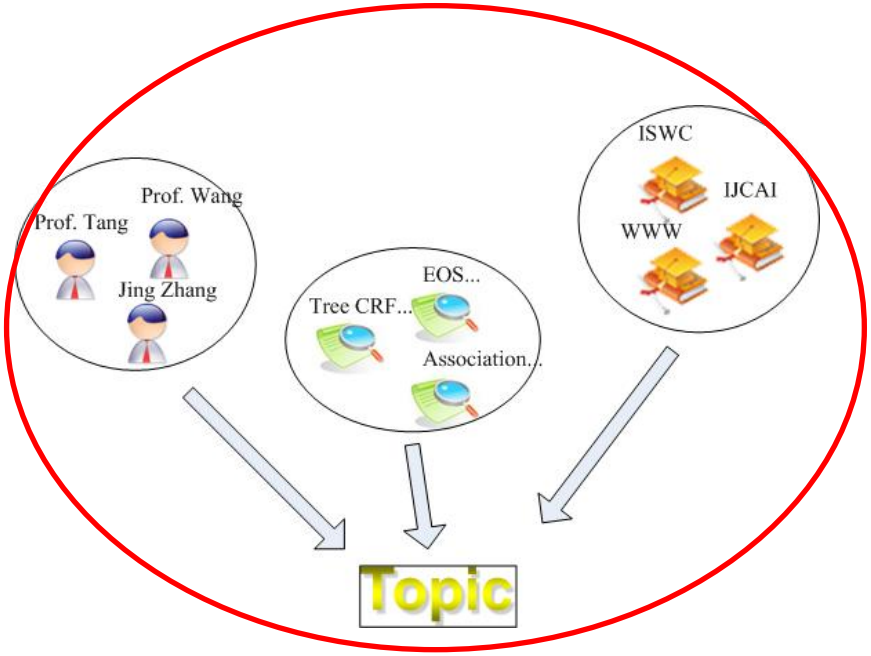
Modeling the Academic Network



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

Integrating Topic Model into Random Walk

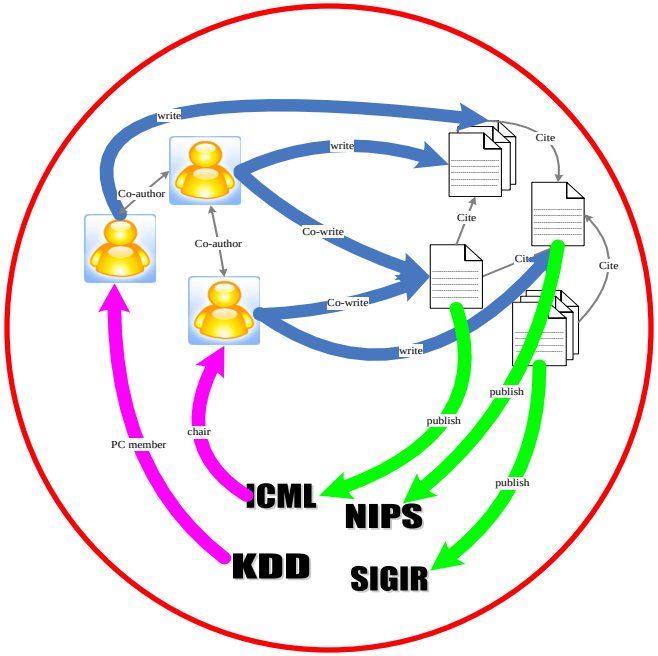
Modeling academic network with topics



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

+

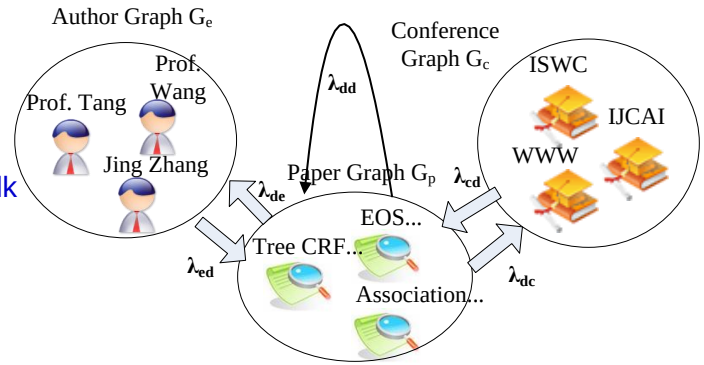
Random walk over the academic network



Combination Method 1

Stage 1:

Random walk



$$r[x] = \frac{\xi}{|V|} + (1-\xi) \times \sum_{(x,y)\in E} \lambda_{yx} r[y] P(x|y)$$

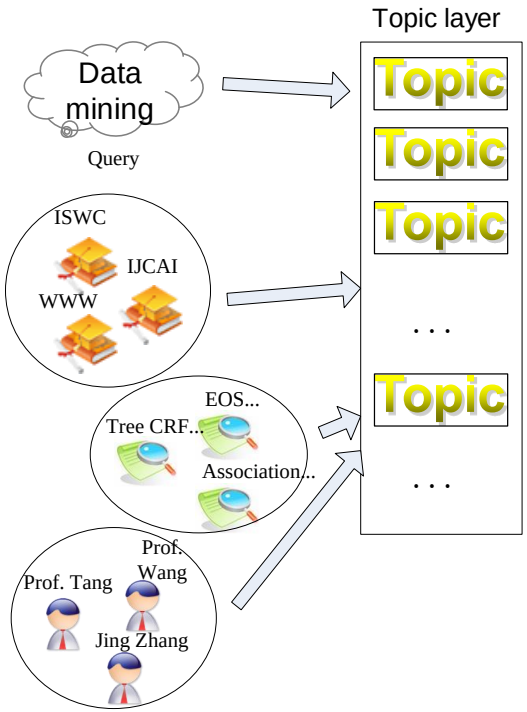
Ranking score

Combination by multiplication

$$R[d] = r[d] \times P(q|d)$$

Stage 2.

Topic-based relevance



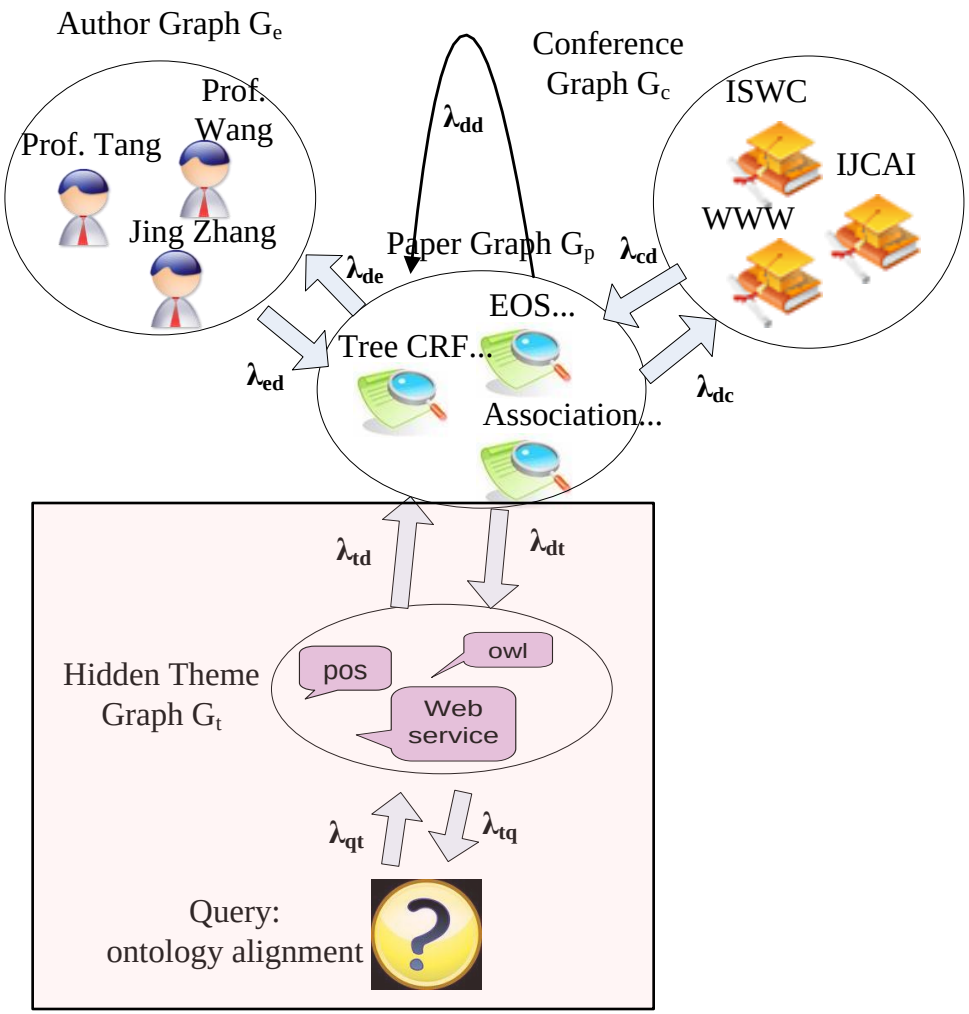
Topic-based relevance score

$$P_{LM}(q|d) = \prod_{w\in q} \frac{N_d}{N_d+\lambda} \cdot \frac{tf(w,d)}{N_d} + \left(1 - \frac{N_d}{N_d+\lambda}\right) \cdot \frac{tf(w,\mathbf{D})}{N_{\mathbf{D}}}$$

$$P_{ACT}(q|d,\theta,\phi) = \prod_{w\in q} \sum_{z=1}^{T} P(w|z,\phi_z) \sum_{a\in \mathbf{a}_d} P(z|a,\theta_a) P(a|d)$$

$$P(q|a) = P_{LM}(q|a) \times P_{ACT}(q|a)$$

Combination Method 2



Ranking score

$$r(d_i) = \lambda_{dd} \times \sum_{(d_j,d_i)\in E_{dd}} \boxed{p(d_i \mid d_j)} \cdot r(d_j) + \lambda_{ed} \times \sum_{(e_i,d_i)\in E_{de}} \boxed{P(d_i \mid e_i)} \cdot r(e_i)$$

$$+ \lambda_{cd} \times \sum_{(c_i,d_i)\in E_{dc}} \boxed{P(d_i \mid c_i)} \cdot r(c_i) + \lambda_{td} \times \sum_{(\theta_i,d_i)\in E_{dt}} \boxed{P(d_i \mid \theta_i)} \cdot r(\theta_i)$$

$$r(e_i) = \lambda_{de} \times \sum_{(d_j,e_i)\in E_{de}} \boxed{p(e_i \mid d_j)} \cdot r(d_j)$$

$$r(c_i) = \lambda_{dc} \times \sum_{(d_j,c_i)\in E_{dc}} \boxed{p(c_i \mid d_j)} \cdot r(d_j)$$

Transition probability

$$P(z_i|a_j) = \theta_{a_j z_i}$$

$$P(a_j|z_i) = \frac{P(z_i|a_j)P(a_j)}{P(z_i)}$$

$$P(z_i|d_j) = \frac{1}{A_d} \sum_{x \in \mathbf{a}_d} \theta_{x z_i}$$

$$P(d_j|z_i) = P(\mathbf{w}_d|z_i) = \prod_{i=1}^{N_d} P(w_{di}|z_i)$$

$$P(c_j|z_i) = \psi_{z_i c_j}$$

$$P(z_i|c_j) = \frac{P(c_j|z_i)P(z_i)}{P(c_j)}$$

$$P(q|z_i) = \prod_{w \in q} P(w|z_i)$$

$$P(z_i|q) \propto P(q|z_i)P(z_i)$$

Learning to Rank Experts

- Combining more information

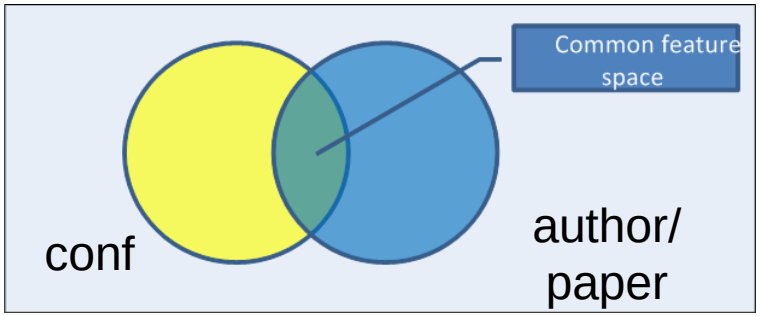
Empirical loss | weight | feature | Model penalty

$$\min_{w_T}\left\{\sum_{i=1}^{n_2}\left[1-z_{T_i}\left\langle w_T, x_{T_i}^a - x_{T_i}^b\right\rangle\right]_+ + \lambda\left\|w_T\right\|_2^2\right\}$$

Feature	Description
L1-L10	Low-level language model features
H1-H3	High-level language model features
S1	The year he/she published his/her first paper
S2	The number of papers of an expert
S3	The number of papers in recent 2 years
S4	The number of papers in recent 5 years
S5	The number of citations of all his/her papers
S6	The number of papers cited more than 5 times
S7	The numebr of papers cited more than 10 times
S8	PageRank score
SumCo1-8	The sum of coauthors' S1-S8 scores
AvgCo1-8	The average of coauthors' S1-S8 scores
SumStu1-8	The sum of his/her advisees' S1-S8 scores
AvgStu1-8	The average of his/her advisees' S1-S8 scores

Language model, BM25, tf*idf

Heterogeneous Cross-domain Ranking



Loss in one domain

Loss in another domain

$$\min_{w_S, w_T} \left\{ \sum_{i=1}^{n_1} \left[1 - z_{S_i} \left\langle w_S, x_{S_i}^a - x_{S_i}^b \right\rangle \right]_+ + C \sum_{i=1}^{n_2} \left[1 - z_{T_i} \left\langle w_T, x_{T_i}^a - x_{T_i}^b \right\rangle \right]_+ + \lambda \|W\|_{2,1}^2 \right\}$$

Common feature space

$$\min_{w_S, w_T, U} \left\{ \sum_{i=1}^{n_1} \left[1 - z_{S_i} \left\langle w_S, U^T \left(x_{S_i}^a - x_{S_i}^b \right) \right\rangle \right]_+ + C \sum_{i=1}^{n_2} \left[1 - z_{T_i} \left\langle w_T, U^T \left(x_{T_i}^a - x_{T_i}^b \right) \right\rangle \right]_+ + \lambda \|W\|_{2,1}^2 \right\}$$

Learning Algorithm

- Equivalent objective function:

$$\min_{D,M}\left\{\sum_{i=1}^{n_1}\left[1-z_{S_i}\left\langle\alpha_1, x_{S_i}^a - x_{S_i}^b\right\rangle\right]_+ + C\sum_{i=1}^{n_2}\left[1-z_{T_i}\left\langle a_2, x_{T_i}^a - x_{T_i}^b\right\rangle\right]_+ + \lambda\sum_{t=1}^{2}\left\langle\alpha_t, D^+\alpha_t\right\rangle\right\}$$

Algorithm Procedure

Input: Training set: $\mathcal{L}_S \bigcup \mathcal{L}_T$; Test set: $\mathcal{S}$
Output: Ranking function $f_T = \langle w_T^*, x\rangle$ and the predicted preferences over test data: $\{y_i\}_{i=1}^n$
Initialization: $D = \frac{I_{d\times d}}{d}$

Step 1: Latent Space Finding
while not reach maximal iteration number T do
$\alpha_1 = \mathtt{argmin}\{\sum_{i=1}^{n_1}[1 - z_{S_i}\langle\alpha, x_{S_i}^a - x_{S_i}^b\rangle]_+ + \lambda\langle\alpha, D^+\alpha\rangle\}$
$\alpha_2 = \mathtt{argmin}\{\sum_{i=1}^{n_2}[1 - z_{T_i}\langle\alpha, x_{T_i}^a - x_{T_i}^b\rangle]_+ + \lambda\langle\alpha, D^+\alpha\rangle\}$
$M = [\alpha_1, \alpha_2]$
set $D = \frac{(MM^\top)^{\frac{1}{2}}}{trace(MM^\top)^{\frac{1}{2}}}$
end while

Step 2: Learning in Latent Space
Apply SVD decomposition on D, $D = U\Sigma V^\top$
Construct U by the eigenvectors corresponding to the first and second biggest eigenvalues of D
$w_T^* = \mathtt{argmin}\{\sum_{i=1}^{n_1}[1 - z_{S_i}\langle w, U^\top(x_{S_i}^a - x_{S_i}^b)\rangle]_+ + C\sum_{i=1}^{n_2}[1 - z_{T_i}\langle w, U^\top(x_{T_i}^a - x_{T_i}^b)\rangle]_+ + \lambda||w||^2\}$
for i = 1 to n do
$y_i = \langle w_T^*, U^T x_i\rangle$
end for

Optimize the loss function for each domain

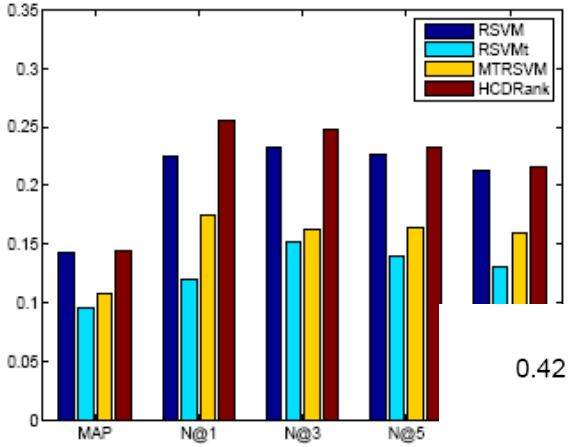
Common space discovery

Optimize the weight via the common space

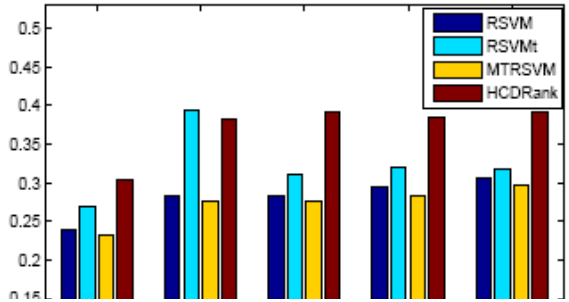
Experimental Results

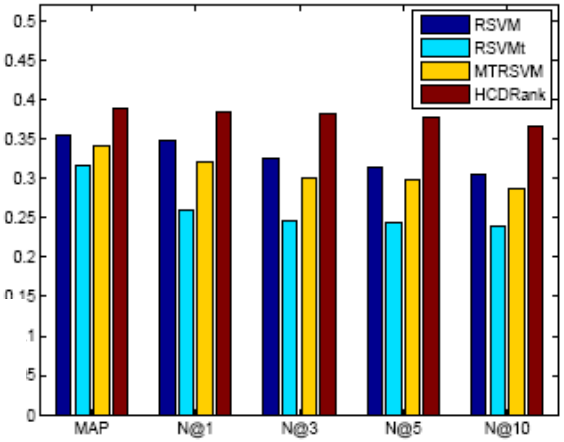
- Data sets
  - Homogeneous Data
    - LETOR 2.0: TREC2003, TREC2004, and OHSUMED
  - Heterogeneous Data
    - Academic network consisting of 14,134 authors, 10,716 papers, and 1,434 conferences.
  - Heterogeneous Tasks
    - Expert finding vs. Bole search
- Baselines
  - RSVM
  - Language model

Results on Homogeneous Data

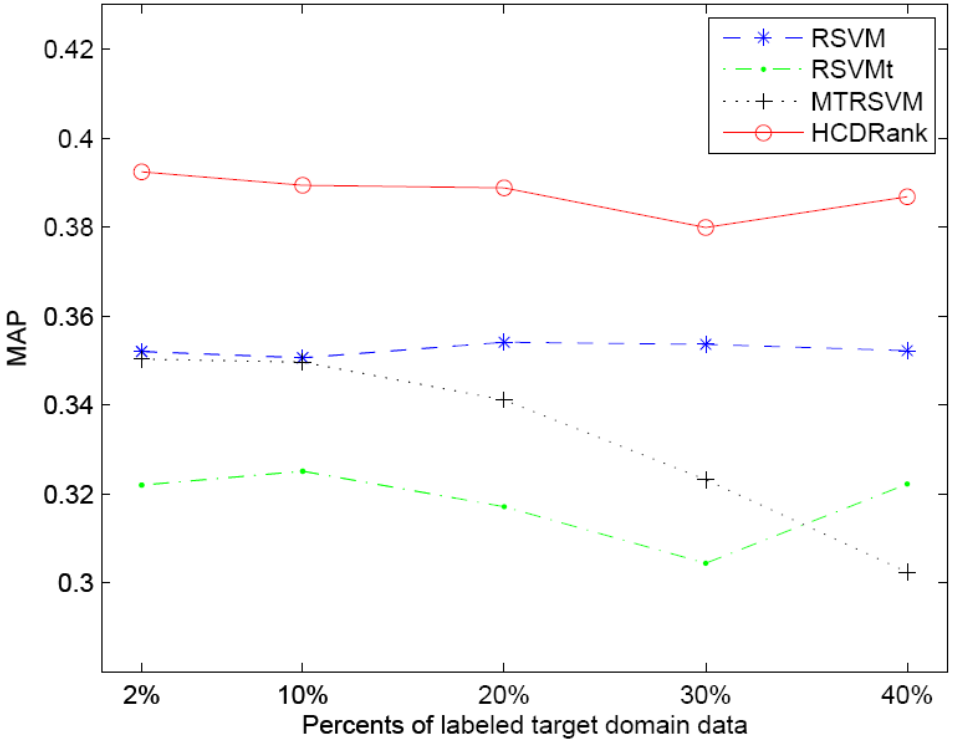


(a)TREC2003_TR
KL divergence=2.4





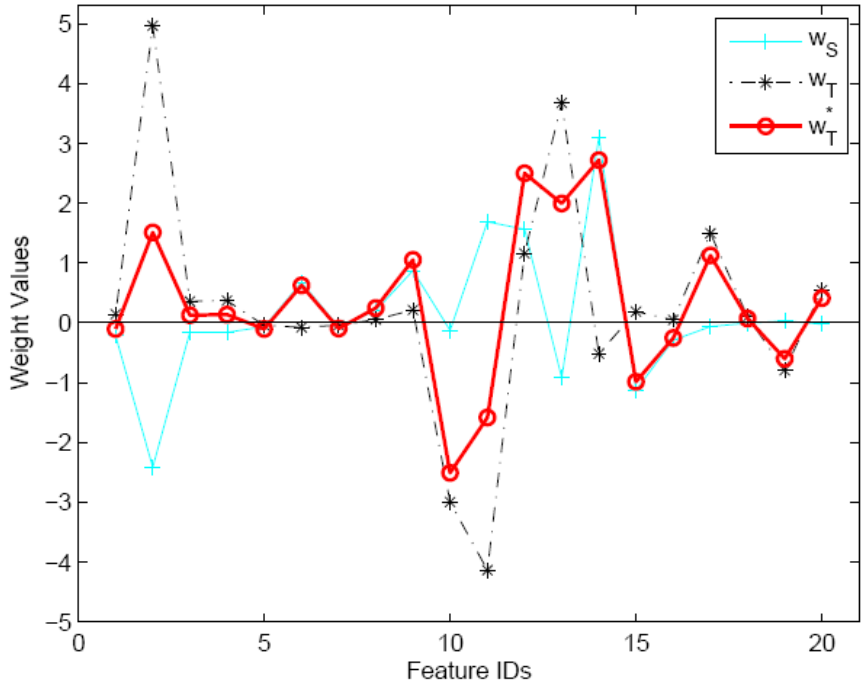
(c)OHSUMED_TR
KL divergence=2.20



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

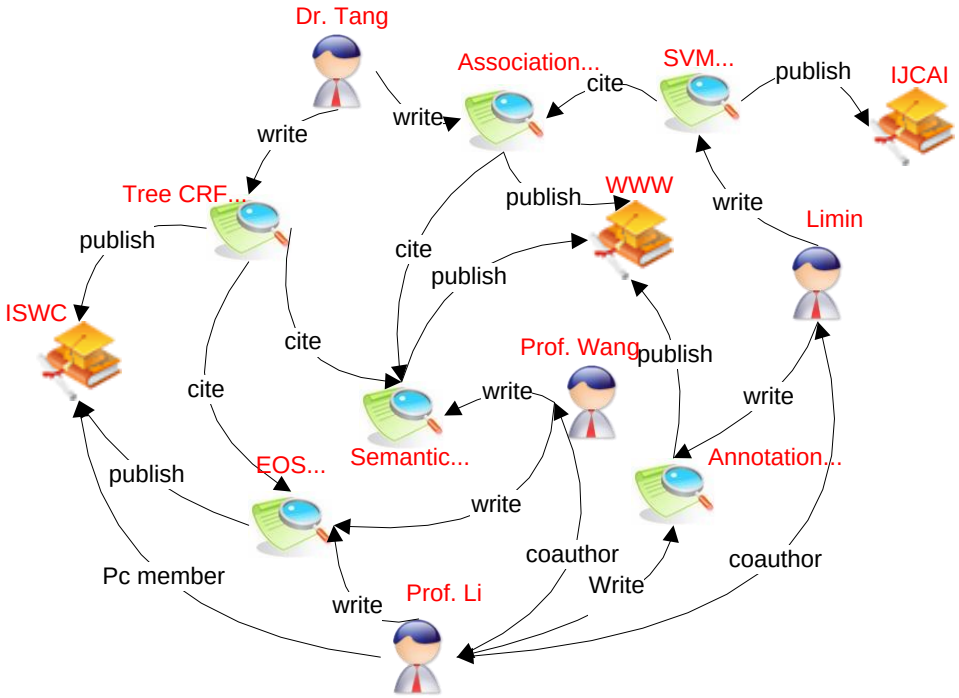
- Expert finding verse Bole search (finding best supervisor)
- To obtain ground truth of bole for each query
  - We sent emails to 50 senior researchers and 50 junior researchers (91.6% are post doc or graduates)
  - Average their feedbacks

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

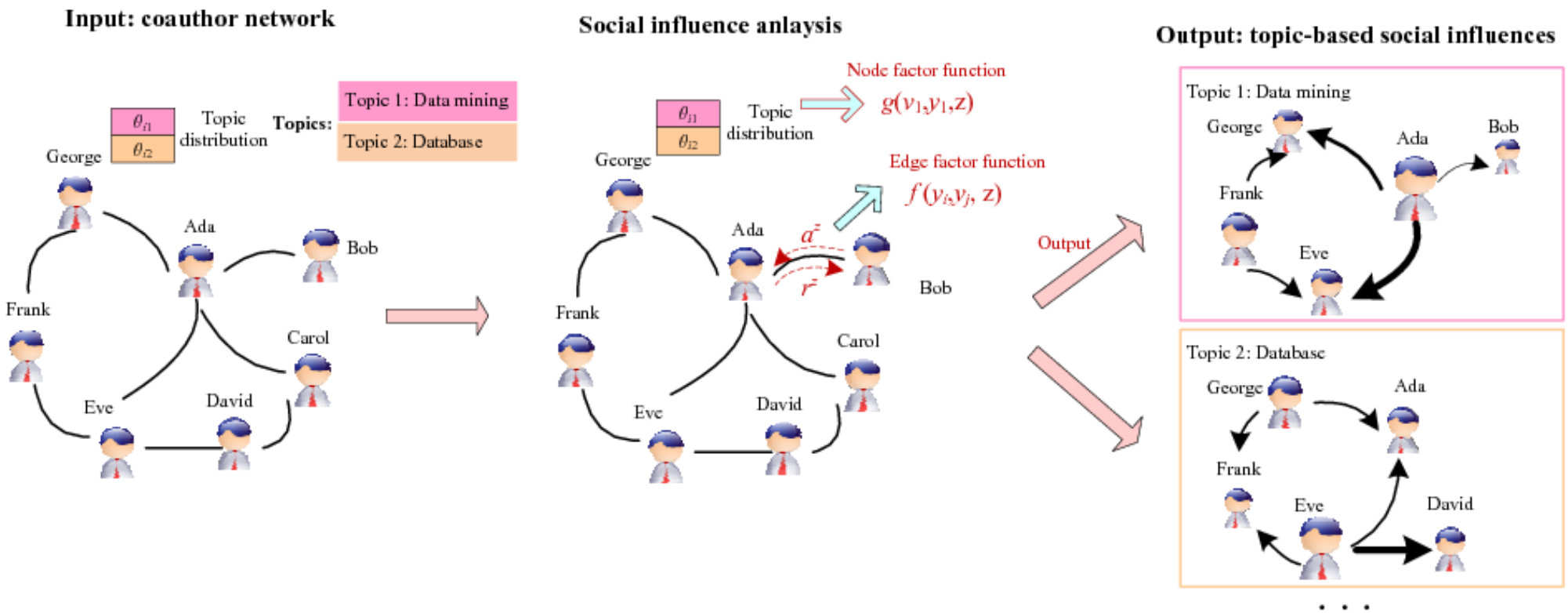
CT4: Social Influence Analysis
(KDD'10, KDD'10, KDD'09, ICDM'09, JIS)

- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user's actions?

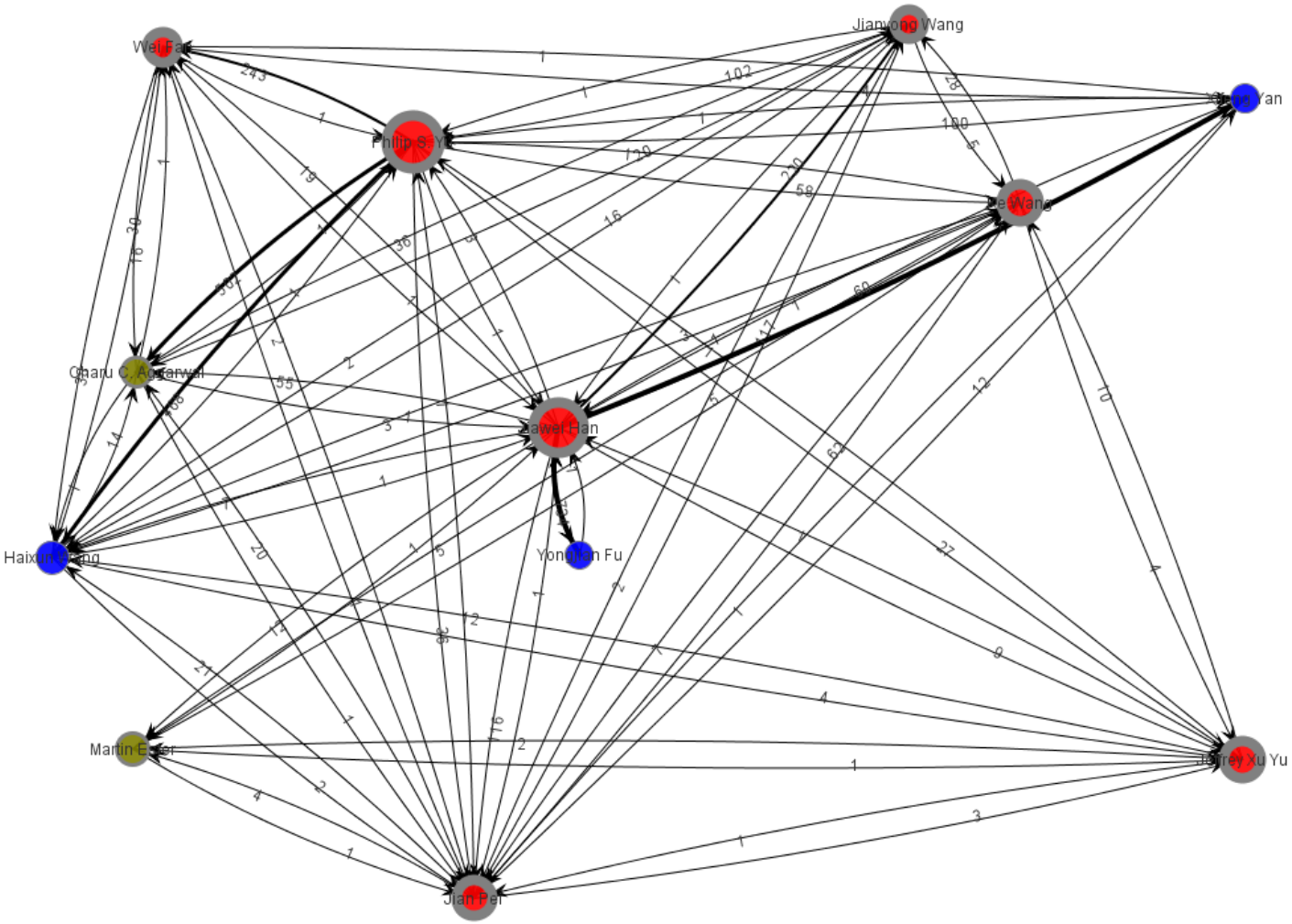


Topic-based Social Influence Analysis

- Social network -> Topical influence network



Social Influence Sub-graph on “Data mining”

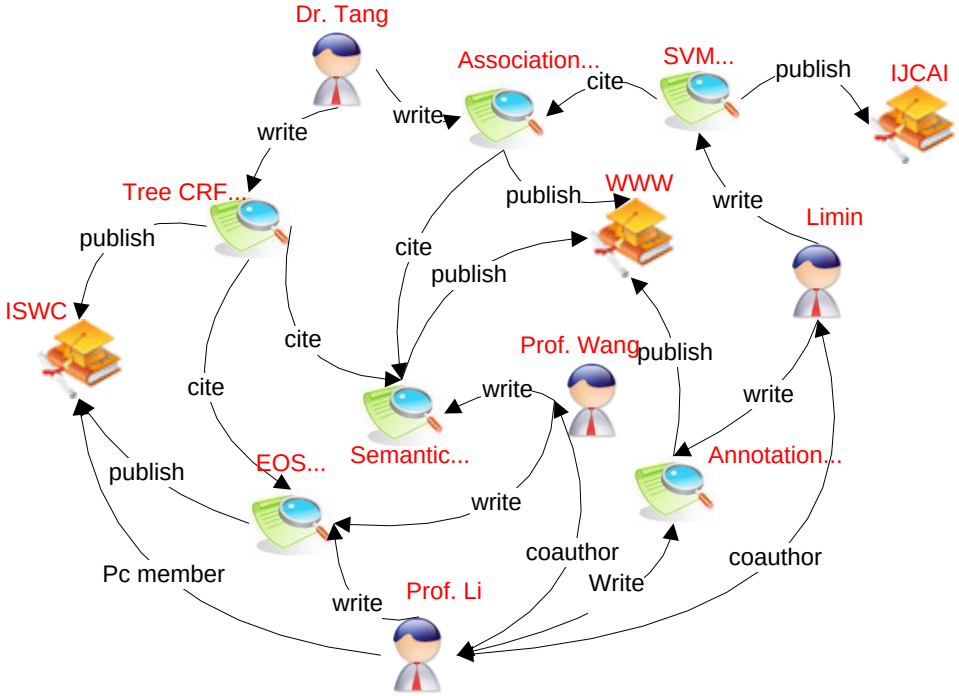


Influential nodes on different topics

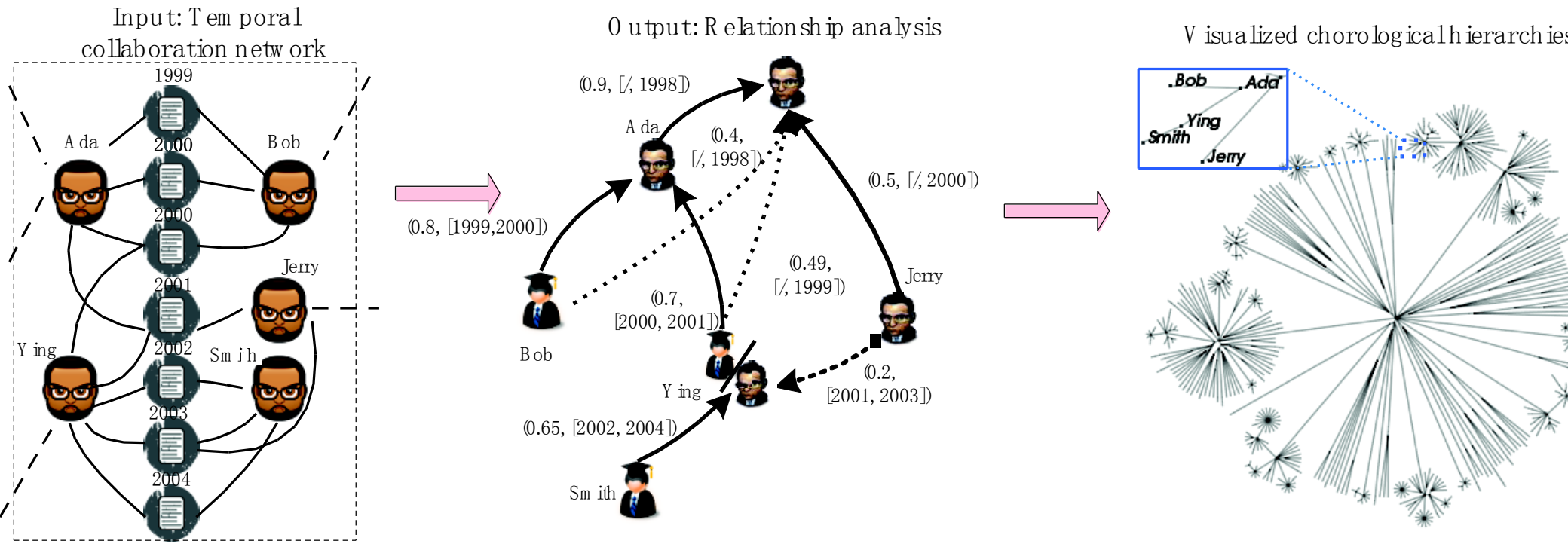
Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell, 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

CT4: Social Influence Analysis
(KDD'10, KDD'10, KDD'09, ICDM'09, JIS)

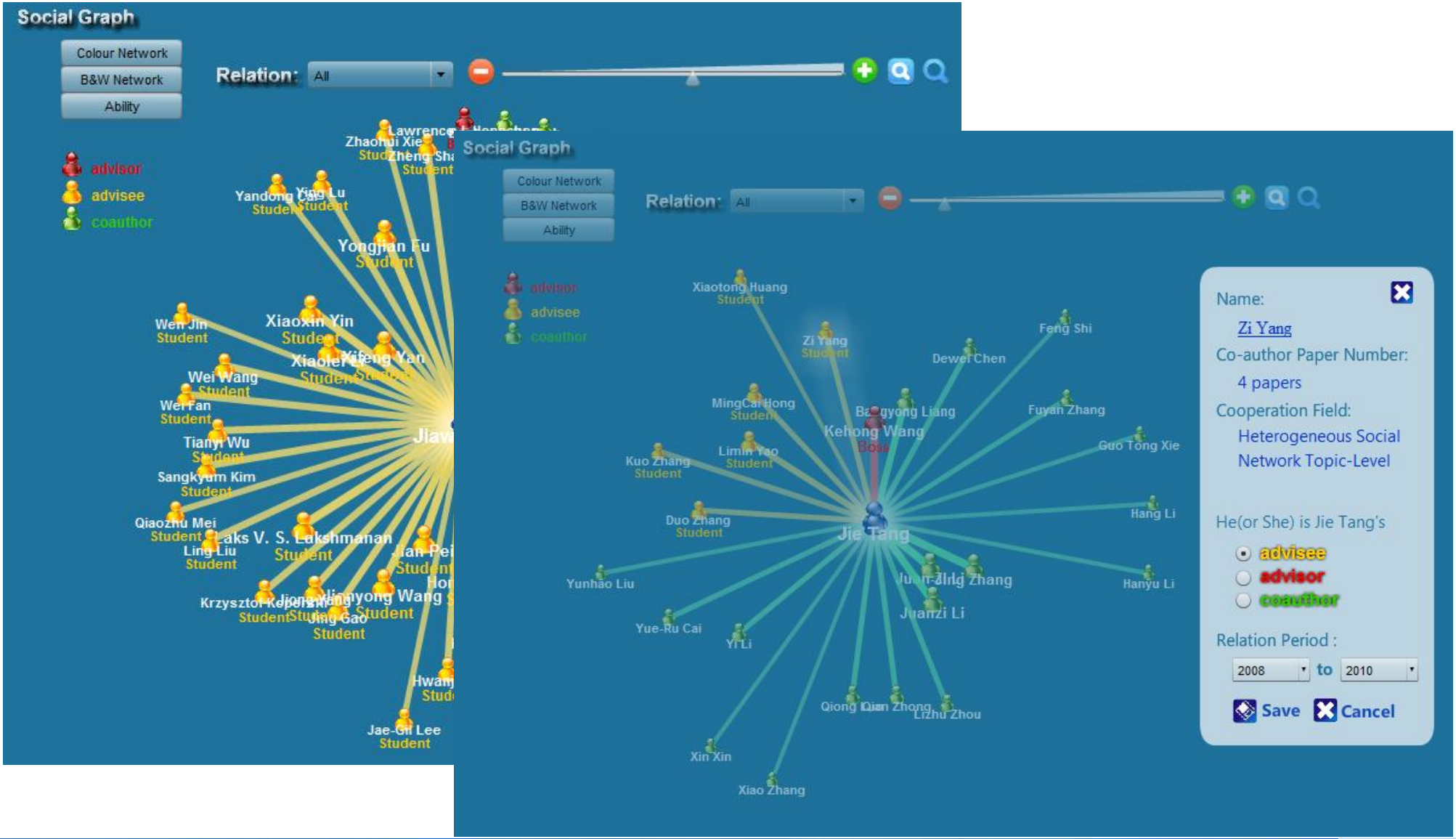
- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user's actions?



Mining Advisor-Advisee Relationship from Research Publication Networks

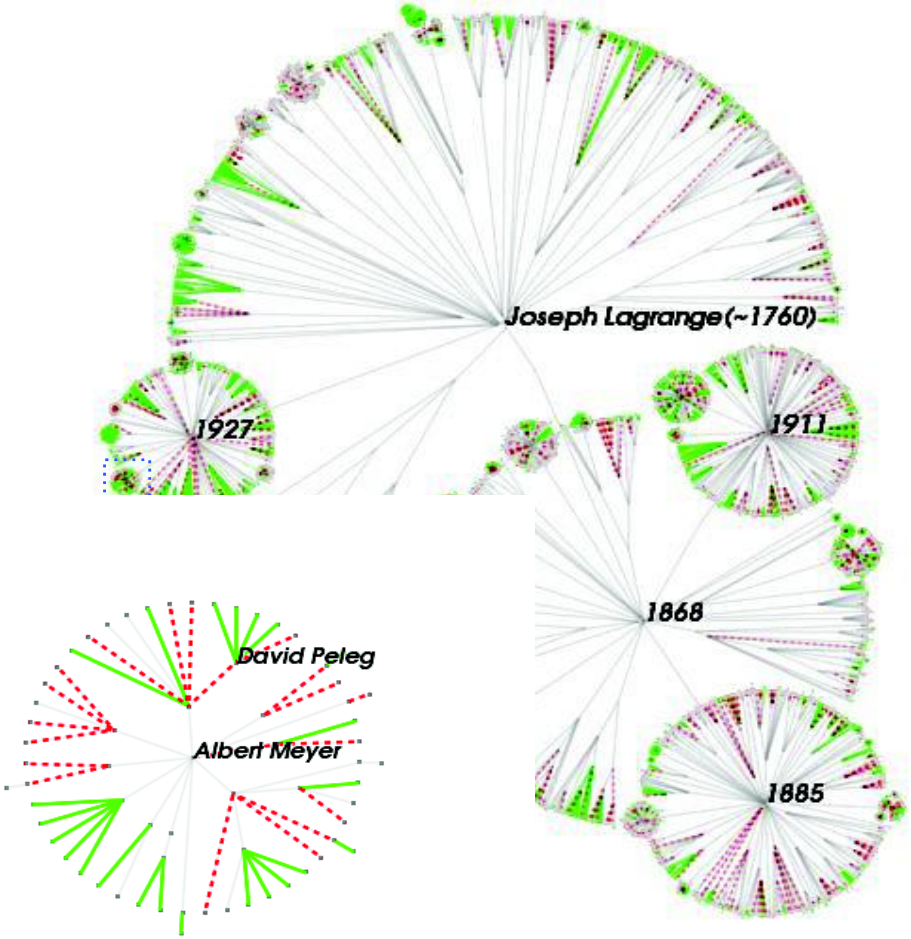


Results

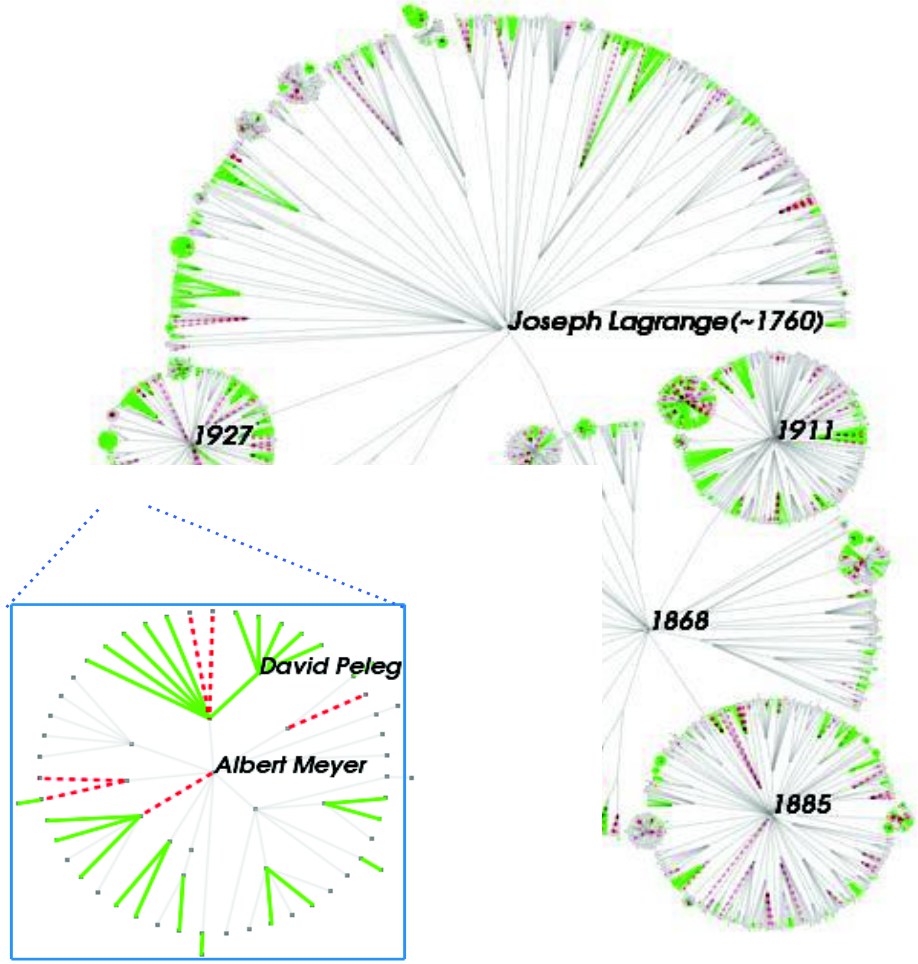

Social Graph
Colour Network
B&W Network
Ability
Relation: All
advisor
advisee
coauthor
Yongjian Fu
Student
Wei Jin
Student
Xiaoxin Yin
Student
Wei Wang
Student
Wei Fan
Student
Tianyi Wu
Student
Sangkyum Kim
Student
Qiaozhu Mei
Student
Ling Liu
Student
Laks V. S. Lakshmanan
Student
Jing Gao
Student
Jae-Gil Lee
Student
Social Graph
Colour Network
B&W Network
Ability
Relation: All
advisor
advisee
coauthor
Xiaotong Huang
Student
Zi Yang
Student
Dewei Chen
Feng Shi
MingCai Hong
Student
Kehong Wang
Boss
Fuyan Zhang
Guo Tong Xie
Kuo Zhang
Student
Limin Yao
Student
Duo Zhang
Student
Jie Tang
Hang Li
Yunhao Liu
Hanyu Li
Juanzi Li
Yue-Ru Cai
Yi Li
Qiong Qian Zhong
Lizhu Zhou
Xin Xin
Xiao Zhang
Name:
Zi Yang
Co-author Paper Number:
4 papers
Cooperation Field:
Heterogeneous Social Network Topic-Level
He(or She) is Jie Tang's
advisee
advisor
coauthor
Relation Period :
2008
to
2010
Save
Cancel

Application: visualization

RULE

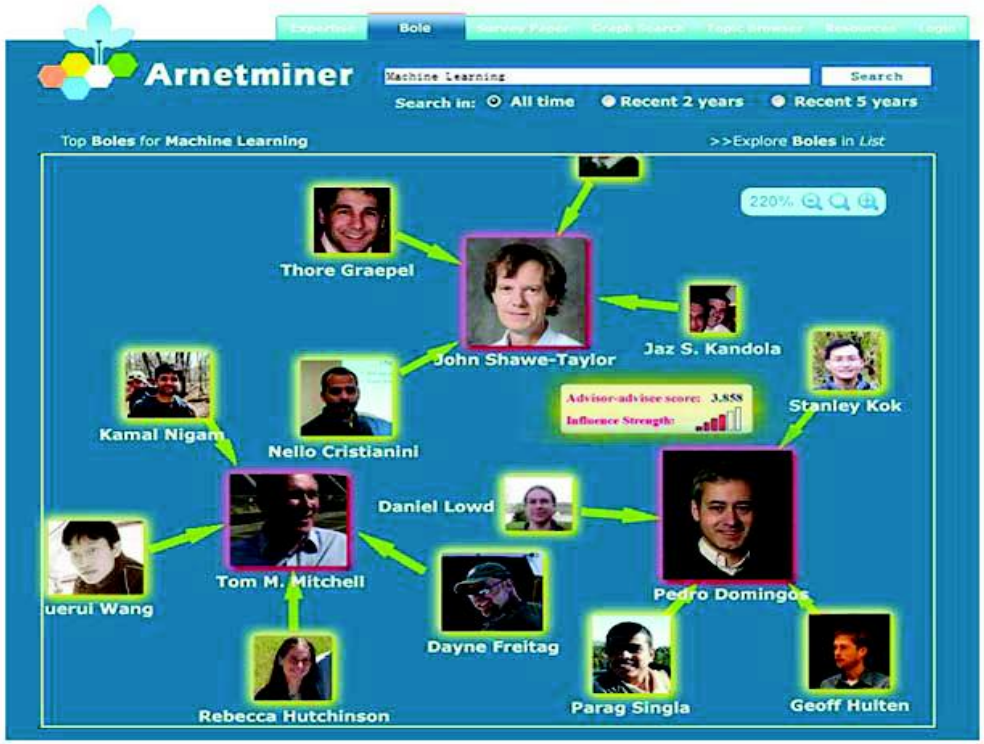


TPFG

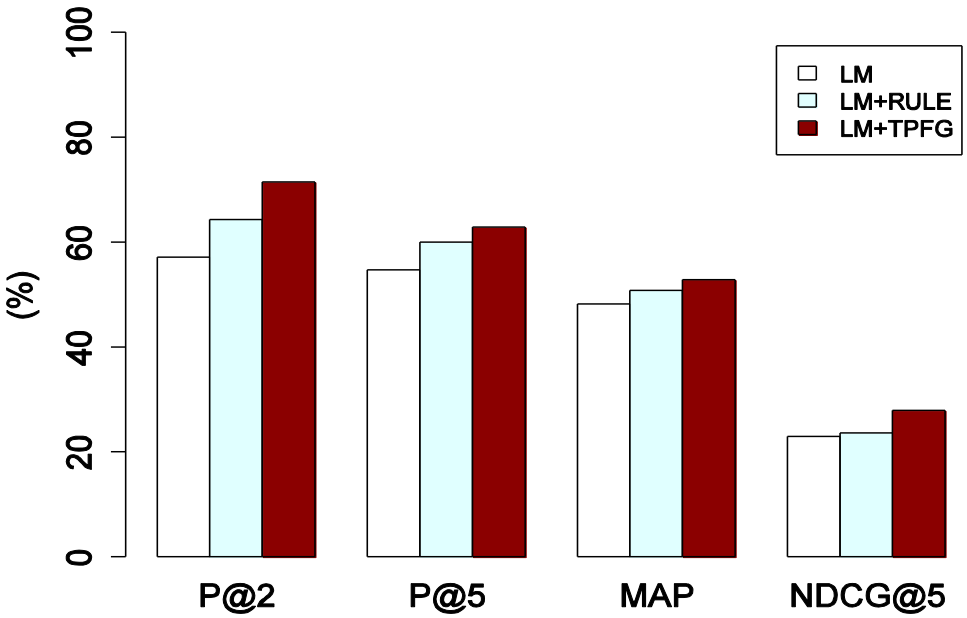


Bole Search In Arnetminer

An example on a real system: Arnetminer



Performance improvement



Results (cont.)

Table 1: Accuracy of prediction by P@(2,θ): $\frac{T}{T+F}$

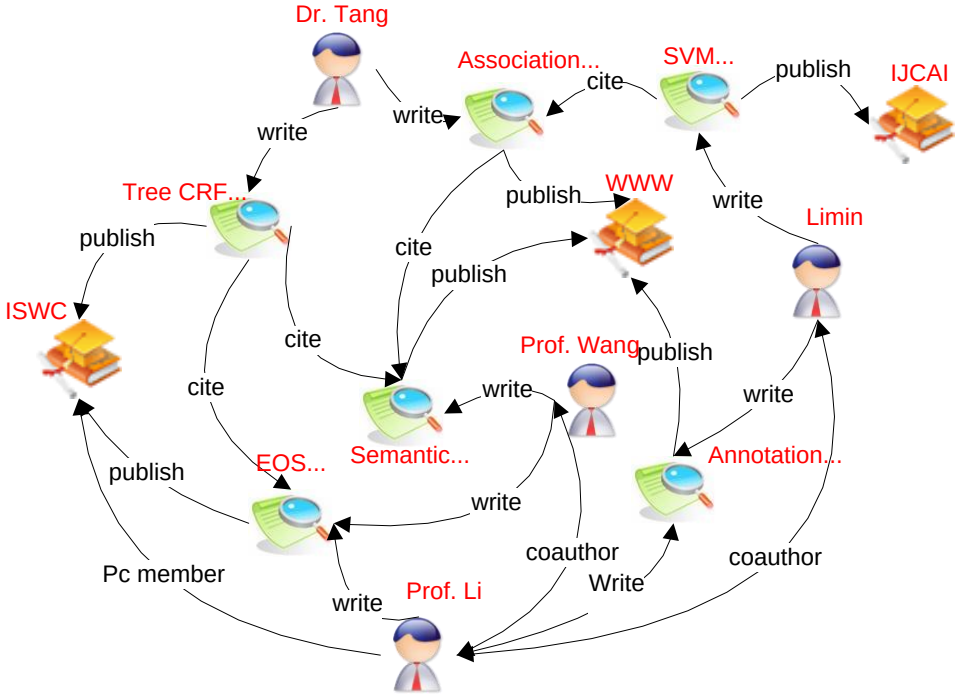
data set	RULE	SVM	IndMAX		TPFG	
TEST1	69.9%	73.4%	75.2%	78.9%	80.2%	84.4%
TEST2	69.8%	74.6%	74.6%	79.0%	81.5%	84.3%
TEST3	80.6%	86.7%	83.1%	90.9%	88.8%	91.3%

Advisee	Top Ranked Advisor	Time	Note
David M. Blei	1. Michael I. Jordan	2001-2003	PhD advisor, 2004
	2. John D. Lafferty	2005-2006	Postdoc, 2006
Hong Cheng	1. Qiang Yang	2002-2003	MS advisor, 2003
	2. Jiawei Han	2004-2008	PhD advsior, 2008
Surgey Brin	1. Rajeev Motwani	1997-1998	"Unofficial advisor"[3]

[3] cited from a blog of Surgey Brin, who left Stanford to found Google around 1998.

CT4: Social Influence Analysis
(KDD'10, KDD'10, KDD'09, ICDM'09, JIS)

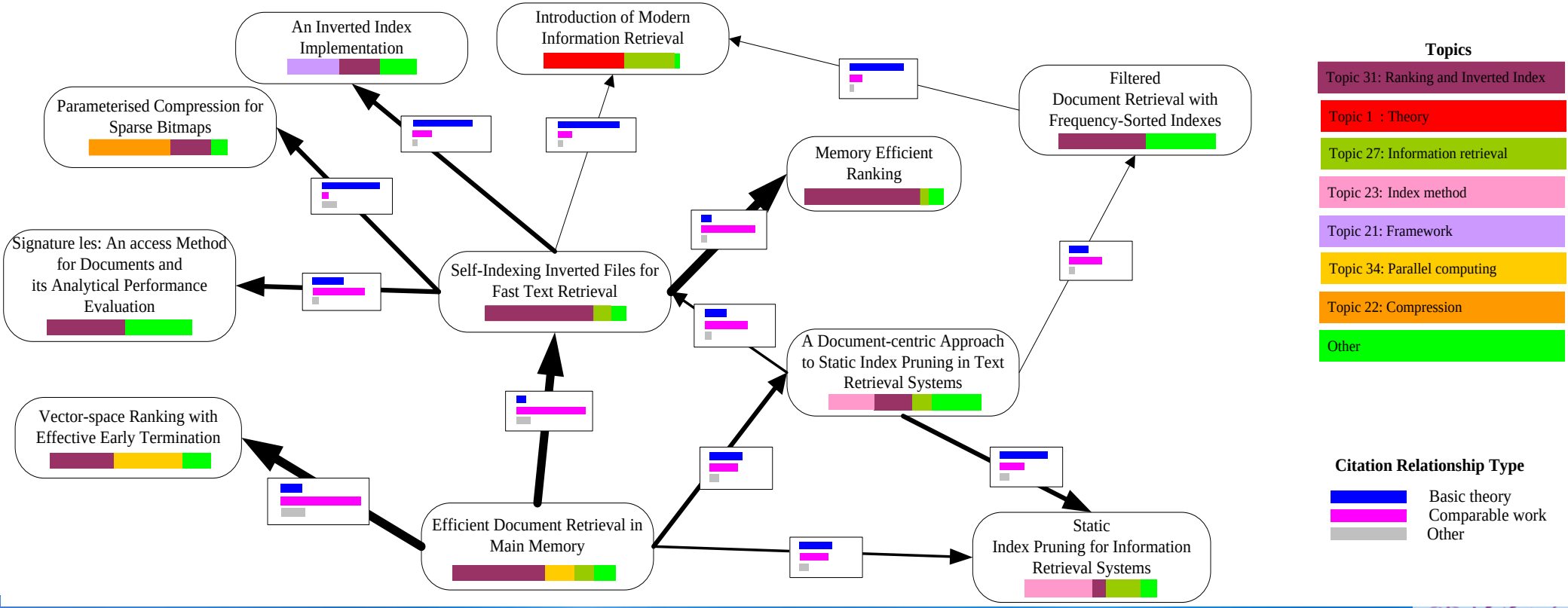
- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user's actions?



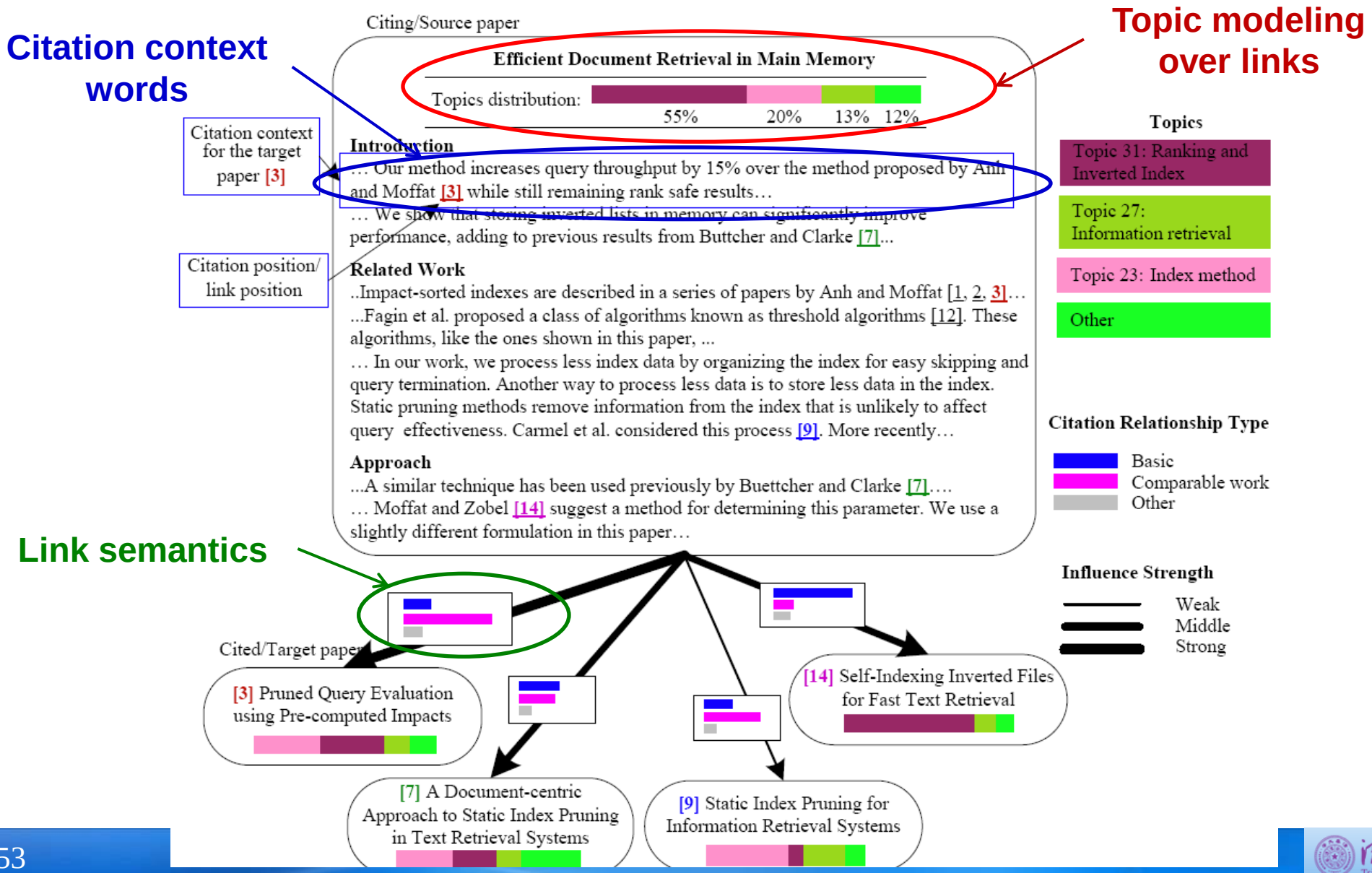
Examples – Topic distribution analysis over citations

Researcher A

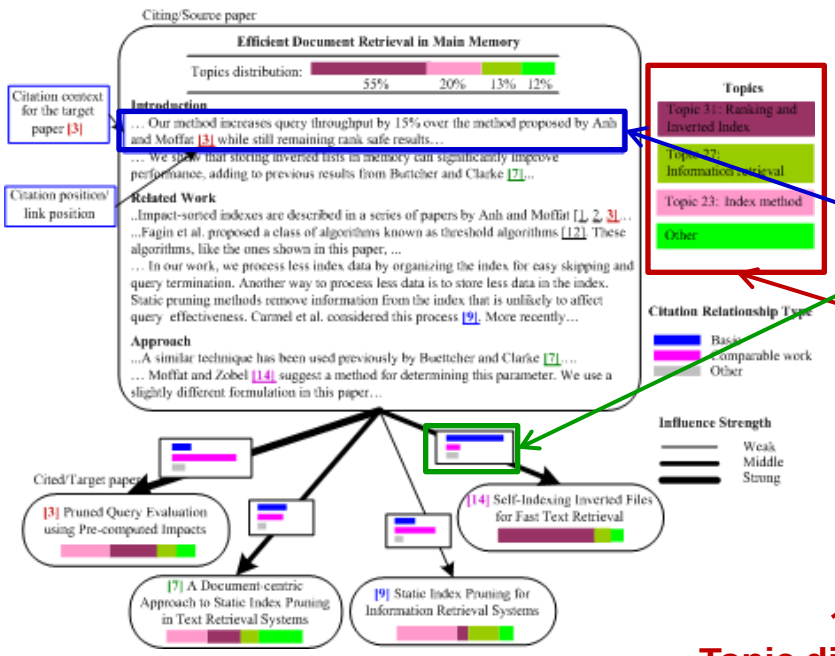
• an in-depth understanding of the research field?



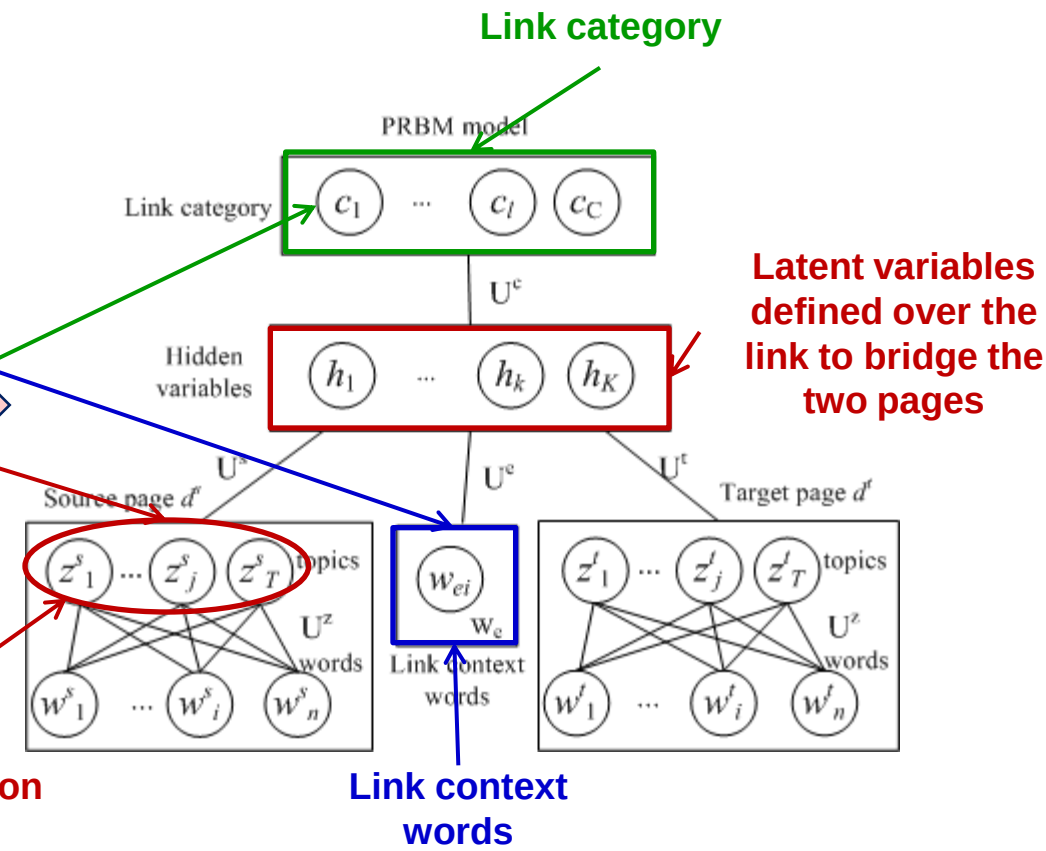
Problem: Link Semantic Analysis



Pairwise Restricted Boltzmann Machines (PRBMs)



Example



Pairwise Restricted Boltzmann Machines (PRBMs)

Accuracy of Link Categorization

Approach	Type	Precision	Recall	F1-measure	Accuracy
SVM	drill down	64.09	53.60	58.24	88.50
	similar	77.66	83.34	80.33	76.59
	other	64.63	59.58	61.77	79.71
	Avg.	68.79	65.51	66.78	81.60
SVM+LDA	drill down	66.31	58.56	61.71	89.24
	similar	80.69	85.47	82.93	79.71
	other	69.58	64.36	66.75	82.42
	Avg.	72.19	69.46	70.46	83.79
SVM+RBM	drill down	68.63	53.0	59.66	89.40
	similar	76.47	91.74	83.36	78.89
	other	76.65	53.28	62.54	82.58
	Avg.	73.92	66.02	68.52	83.62
gPRBM	drill down	69.77	65.22	67.42	88.07
	similar	77.33	87.22	81.98	79.01
	other	78.00	60.94	68.42	85.19
	Avg.	75.03	71.12	72.61	84.09
dPRBM	drill down	96.00	44.44	61.54	95.88
	similar	76.85	96.89	85.71	78.60
	other	75.00	37.50	50.00	80.25
	Avg.	**82.62**	59.61	65.75	84.91
hPRBM	drill down	65.38	68.00	66.67	93.00
	similar	83.54	92.31	87.71	84.77
	other	88.14	69.33	77.61	87.65
	Avg.	79.02	**76.55**	**77.33**	**88.48**

Tested on Arnetminer citation data

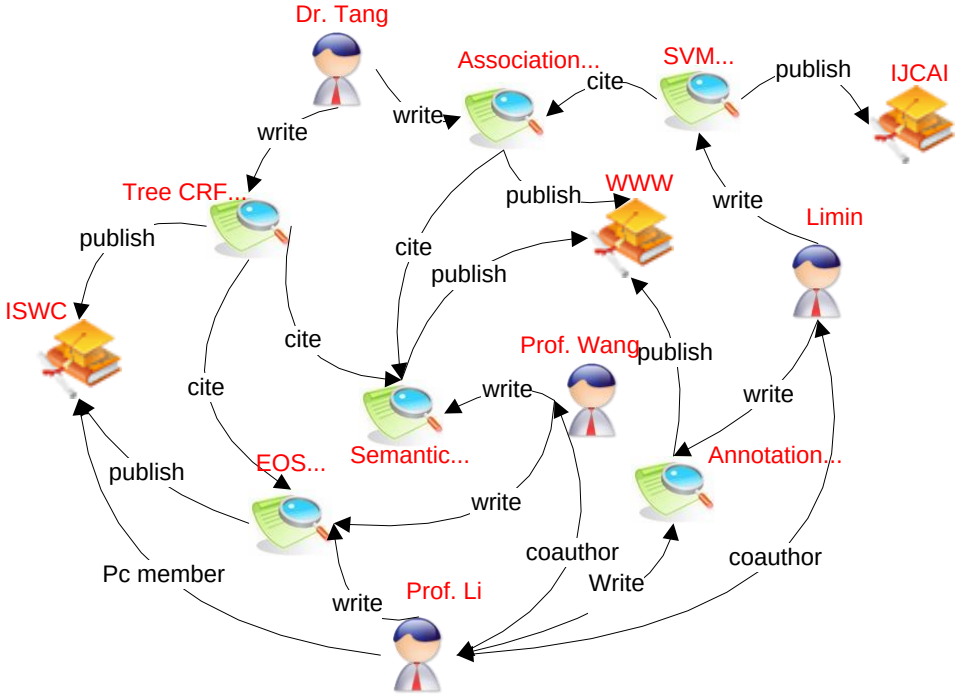
gPRBM: our approach with generative learning

dPRBM: our approach with discriminative learning

hPRBM: our approach with hybrid learning

CT4: Social Influence Analysis
(KDD'10, KDD'10, KDD'09, ICDM'09, JIS)

- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user's actions?



What can we do in SNS?

Social Action

Twitter

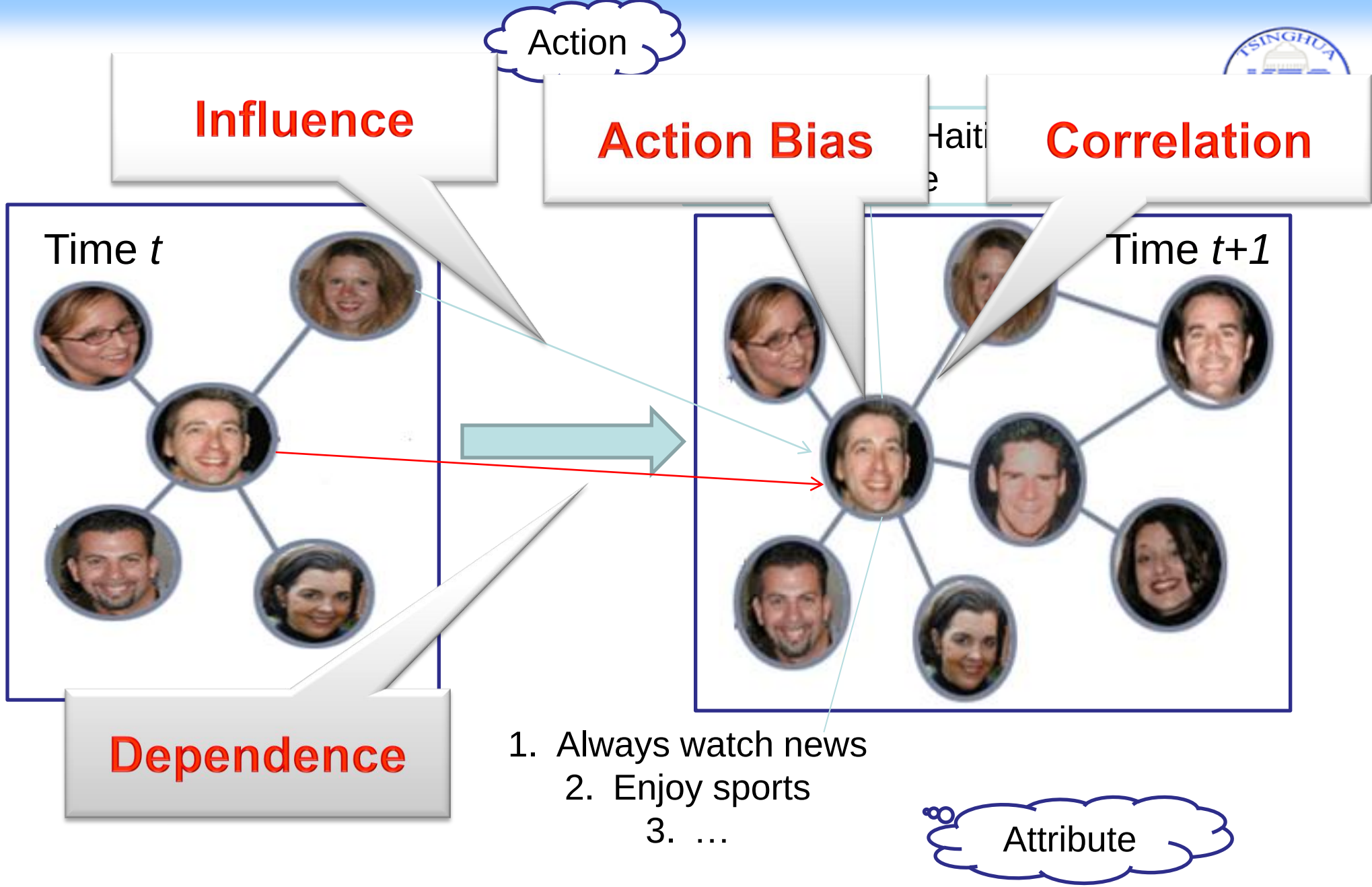
Flickr

KDD

Comment on Haiti Earthquake

Add favorites

Publish in KDD Conference


Action
Influence
Action Bias
Correlation
Time t
Time t+1
Dependence
1. Always watch news
2. Enjoy sports
3. …
Attribute

Results

Table 1: Performance of action prediction with different approaches (%).

Data set	Method	Recall	Precision	F1-Measure
Twitter	SVM	10.41	16.71	13.85
	wvRN	0.45	7.89	0.86
	NTT-FGM	26.40	21.14	23.47
Flickr	SVM	34.48	45	
	wvRN	60.02	48	
	NTT-FGM	56.18	45	
ArnetMiner	SVM	10.19	21	
	wvRN	14.83	16	
	NTT-FGM	31.14	44	

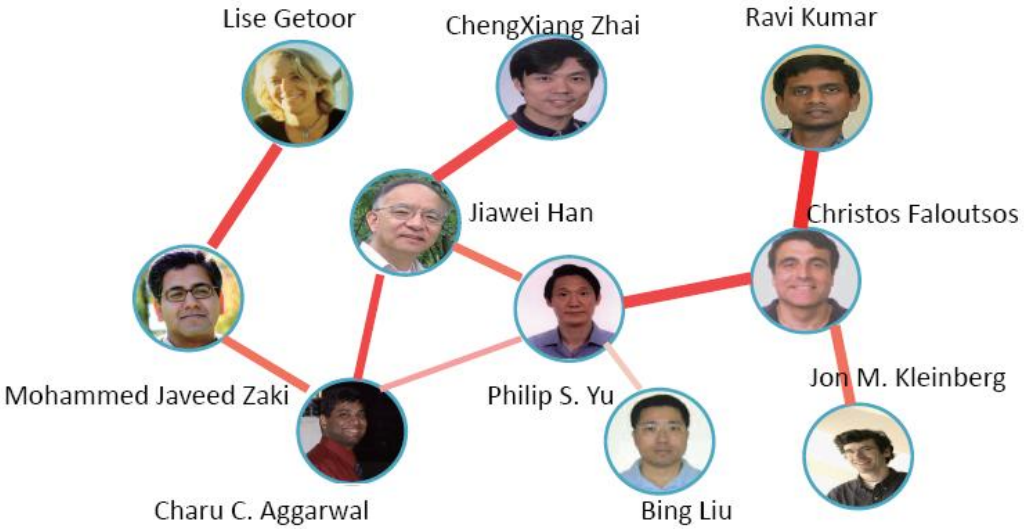


Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.

Arnetminer Today

— A brief summary

ArnetMiner's History

- 2006/5, V0.1 Perl-based CGI version
  - Profile extraction, person/paper/conf. search
- 2006/8, V1.0 Java (Demo @ ASWC)
  - Rewrite the above functions
- 2007/7, V2.0 (Demo @ KDD, ISWC)
  - New: survey search, research interest, association search
- 2008/4, V3.0 (Demo @ WWW)
  - Query understanding, New search GUI, log analysis
- 2008/11, V4.0 (Demo @ KDD, ICDM)
  - Graph search, topic mining, NSFC/NSF
- 2009/4, V5.0 (Demo @ KDD)
  - Bole/course search, profile editing, open resources, #citation
- 2009/12, V6.0
  - Academic statistics, user feedbacks, refined ranking
- V7.0, coming soon
  - Name disambiguation, reviewer assignment, supervisor suggestion, open API

ArnetMiner Today

* Arr
  > (
  > 3
  > 1
  > 5
  > 5

Messages from Users

… I've happened to visit your Arnetminer, and shocked. It was really impressive, its usefulness and your works!!! … [from …@selab.snu.ac.kr]

…I would first of all congratulate you on the excellent work you have done in Arnetminer and I am much inspired… [from …@nu.edu.pk]

a survey by UK Southampton

Title: Semantic Technologies for Learning and Teaching in Web 2.0.

— Thanassis Tiropanis, Hugh Davis, Dave Millard, Mark Weal

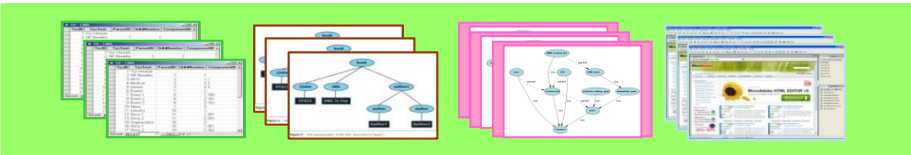
…Exposing the expertise of the institution to the outside world in order to attract funding and students. ArnetMiner is the most representative example of such tools at the moment…

Contextualised queries and searches, searches across repositories potentially in different departments or institutions, and matching of people for collaborative activities. Best example of the surveyed technologies to this end is ArnetMiner.

* V
co
* (
vi
* :

ind
m]
.org)
Miner.

Opportunity: exploiting semantic web and social network 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

Social search & mining

Social network Extraction
Social network Mining

IBM

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

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,

Arnetminer

Arnetminer

PatentMiner

ScopusMiner

PubmedMiner

CheMiner

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**.
- Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM **TKDD**, 2010.
- Juanzi Li, Jie Tang, Yi Li, Qiong Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE **TKDE**, 2009.
- 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, Duo Zhang, Yintao Yu, Jie Tang, 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**.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. **KDD'08**.
- Jie Tang, Hang Li, Yunbo Cao, and Zhaohui Tang. Email Data Cleaning. **KDD'05**.
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. **IJCAI'09**.
- Qian Zhong, Hanyu Li, Juanzi Li, Guotong Xie, Jie Tang, Lizhu Zhou. A Gauss Function based Approach for Unbalanced Ontology Matching. **SIGMOD'09**.
- Feng Shi, Juanzi Li, Jie Tang. Actively Learning Ontology Matching via User Interaction. **ISWC'09**.
- Chonghui Zhu, Jie Tang, Hang Li, Hwee Tou Ng, and Tiejun Zhao. A Unified Tagging Approach to Text Normalization. **ACL'07**.

Others: **ICDM'07-09**, **CIKM'07-09**, **SDM'09**, **ISWC'06**, **DKE**, **JIS**, etc.

Thanks!

Demo: http://arnetminer.org

HP: http://keg.cs.tsinghua.edu.cn/persons/tj/