Arnetminer: Deep Analysis and Mining for Academic Social Networks

Jie Tang
Knowledge Engineering Group
Department of Computer Science and Technology
Tsinghua University
Email: jietang@tsinghua.edu.cn
Web-based Social Network Mining

1. Social knowledge acquisition
   - ACM TKDD, IEEE TKDE
   - SIGMOD’09, SIGKDD’08
   - SIGIR’11, IJCAI’09
   - SDM’09, ICDM’07

2. Social search
   - Mach. Learn. J
   - ICDM’08
   - CIKM’09

3. Relation mining
   - SIGKDD’10
   - ICDM’09
   - PKDD’11

4. Social influence analysis
   - SIGKDD’09
   - CIKM’10
   - IEEE Inte. Sys.

5. Social prediction
   - SIGKDD’10
   - SIGKDD’11
   - CIKM’10

6. Mobile Social Network
   - IEEE TAC
   - ICDM’10

Complex Social Network

Theoretical layer

Collective learning | Learning from users | Social theory | Graphical models
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/
Person Search

Basic Info.

Citation statistics

Research Interests

Social Network

Fundings

Publications
Expertise Search

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

Finding courses for “data mining”
Association Search

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

Usage:
- to find a partner
- to find a person with same interests
200 topics have been discovered automatically from the academic network.
Academic Statistics

Top 3 by H-index:

1. Hector Garcia-Molina (Prof)
   H-index 95, Papers 37
   Expertise: Semantic web / Hypermedia Systems, Multimedia Applications / Context-Based Video Retrieval, Intelligent User Interfaces / Information Agents
   Homepage: http://www.cs.columbia.edu/~hec
   
2. Christos H. Papadimitriou
   H-index: 90, Papers: 77
   Expertise: Optimization, Parallel Algorithms, Complexity Classes, Graph Theory
   Homepage: http://www.ccs.neu.edu/home/papadim/
   
3. Arvind K. Jain (Distinguished Prof)
   H-index: 93, Papers: 78
   Expertise: Object Recognition / MfI, Image registration, Character Recogn...
Outline

• ArnetMiner: Academic Social Network

• Core Techniques
  – Knowledge Acquisition
  – Semantic Integration
  – Heterogeneous Ranking
  – Social Influence Analysis

• Demo
ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

ArnetMiner: Overview

1. Social Network Storage
   - Access interface
   - Indexing
   - Social Network Storage

2. Modeling and Search Network
   - Topic model
   - Expertise search

3. Social Network Analysis
   - Citation tracing analysis
   - Social influence analysis
   - Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding

Social Network Storage
- Access interface
- Indexing
- Social Network Storage

Social Network Analysis
- Citation tracing analysis
- Social influence analysis
- Social involution analysis

Social Network Extraction
- Profiling
- Name disambiguation
- Publication extraction
- Homepage finding
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?
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)

Benchmark Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiMOM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASMOV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lily</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aflode</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“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.”
**CT3: Topic-based Heterogeneous Ranking**
(KDD’08, ICDM’08, CIKM’09, DKE)

Data mining

Search with keyword

Search with semantic modeling

Modeling using VSM

Modeling using semantic topics

**Topics**
- Data mining
- Association Rules
- Database systems
- Data management
- Web databases
- Information systems

**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…

**Experts**

**Expertise conferences**

**Expertise papers**

**1 - 10 of 4972 experts for data mining. (9.125 seconds)**

**1 - 10 of 2053 publications (30.352 seconds)**

**Active Data Mining**
Authors: Jiaji Wu.
In Proceedings: SIGMOD Conference.
Published year: 1996.

**Active Data Mining: What’s It Handling? ( Fairfax)**
Authors: Jiaji Wu, Hain.
In Proceedings: KDD.
Published year: 1998.

**A Perspective on Databases and Data Mining**
Authors: M. Mohai, M. L. Kwan.
In Proceedings: SIGMOD.
Published year: 1995.

**Theoretical Frameworks for Data Mining**
Authors: Hails, Harsh.
In Proceedings: SIGKDD Explorations.
1. How to **model** the heterogeneous academic network?

2. How to **capture** the link information for ranking objects in the academic network?
Modeling the Academic Network

Author-Conference-Topic Model [Tang et al., 08]
Integrating Topic Model into Random Walk

Modeling academic network with topics

Random walk over the academic network

Author-Conference-Topic Model [Tang et al., 08]
Combination Method 1

Stage 1:
Random walk

Stage 2:
Topic-based relevance

Ranking score

Combination by multiplication

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

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

is \( P_{LM}(q|d) \) and \( P_{ACT}(q|d, \theta, \phi) \) are defined as:

\[ P_{LM}(q|d) = \prod_{w \in q} \frac{N_d}{N_d + \lambda} \cdot \frac{t_f(w, d)}{N_d} + (1 - \frac{N_d}{N_d + \lambda}) \cdot \frac{t_f(w, D)}{N_D} \]

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

P(q|a) = P_{LM}(q|a) \times P_{ACT}(q|a)
Combination Method 2

Ranking score

\[ r(d_i) = \lambda_{dd} \times \sum_{(d_j, d_y) \in E_{de}} p(d_i | d_j) \cdot r(d_j) + \lambda_{cd} \times \sum_{(e_i, e_y) \in E_{de}} P(d_i | e_i) \cdot r(e_i) \]
\[ + \lambda_{cd} \times \sum_{(c_i, d_y) \in E_{de}} P(d_i | c_i) \cdot r(c_i) + \lambda_{cd} \times \sum_{(\theta_i, d_y) \in E_{de}} P(d_i | \theta_i) \cdot r(\theta_i) \]
\[ r(e_i) = \lambda_{de} \times \sum_{(d_y, e_y) \in E_{de}} p(e_i | d_j) \cdot r(d_j) \]
\[ r(c_i) = \lambda_{dc} \times \sum_{(d_y, c_y) \in E_{de}} p(c_i | d_j) \cdot r(d_j) \]

Transition probability

\[ P(z_i | a_j) = \theta_{a_j z_i} \]
\[ P(a_j | z_i) = \frac{P(z_i | a_j) P(a_j)}{P(z_i)} \]
\[ P(z_i | d_j) = \frac{1}{A_d} \sum_{z \in a_d} \theta_{x z_i} \]
\[ P(d_j | z_i) = P(w_d | z_i) = \prod_{i=1}^{N_d} P(w_{di} | z_i) \]
\[ P(c_j | z_i) = \psi_{z_i c_j} \]
\[ P(z_i | c_j) = \frac{P(c_j | z_i) P(z_i)}{P(c_j)} \]
\[ P(q | z_i) = \prod_{w \in q} P(w | z_i) \]
\[ P(z_i | q) \propto P(q | z_i) P(z_i) \]
Learning to Rank Experts

- Combining more information

\[ \min_{w_T} \left\{ \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \left( w_T \cdot x_{T_i}^a - x_{T_i}^b \right) \right] + \lambda \left\| w_T \right\|^2 \right\} \]

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

Language model, BM25, tf*idf
Heterogeneous Cross-domain Ranking

Query: “data mining”

Papers
- Principles of Data Mining
- Data Mining: Concepts and Techniques

Conferences
- KDD
- SDM
- ICDM
- PAKDD

Authors
- P. Yu

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

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

Common feature space
Learning Algorithm

• Equivalent objective function:

\[
\min_{D,M} \left\{ \sum_{i=1}^{n} \left[ 1 - z_{S_i} \left\langle \alpha_1, x_{S_i}^a - x_{S_i}^b \right\rangle \right]_+ + C \sum_{i=1}^{n} \left[ 1 - z_{T_i} \left\langle \alpha_2, x_{T_i}^a - x_{T_i}^b \right\rangle \right]_+ + \lambda \sum_{i=1}^{2} \left\langle \alpha_i, D^+ \alpha_i \right\rangle \right\}
\]

Algorithm Procedure

Input: Training set: \( L_S \cup L_T \); Test set: \( S \)
Output: Ranking function \( f_T = \left\langle w_T^*, x \right\rangle \) 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
1: while not reach maximal iteration number \( T \) do
2: \( \alpha_1 = \arg\min \left\{ \sum_{i=1}^{n_1} \left[ 1 - z_{S_i} \left\langle \alpha, x_{S_i}^a - x_{S_i}^b \right\rangle \right]_+ + \lambda \left\langle \alpha, D^+ \alpha \right\rangle \right\} \)
3: \( \alpha_2 = \arg\min \left\{ \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \left\langle \alpha, x_{T_i}^a - x_{T_i}^b \right\rangle \right]_+ + \lambda \left\langle \alpha, D^+ \alpha \right\rangle \right\} \)

Step 2: Learning in Latent Space
7: Apply SVD decomposition on \( D, D = U\Sigma V^T \)
8: Construct \( U \) by the eigenvectors corresponding to the first and second biggest eigenvalues of \( E \)
9: \( w_T^* = \arg\min \left\{ \sum_{i=1}^{n_1} \left[ 1 - z_{S_i} \left\langle w, U^T (x_{S_i}^a - x_{S_i}^b) \right\rangle \right]_+ + C \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \left\langle w, U^T (x_{T_i}^a - x_{T_i}^b) \right\rangle \right]_+ + \lambda \|w\|^2 \right\} \)
10: for \( i = 1 \) to \( n \) do
11: \( y_i = \left\langle w_T^*, U^T x_i \right\rangle \)
12: 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
Performance of different approaches for expert finding.

<table>
<thead>
<tr>
<th>Approach</th>
<th>MAP</th>
<th>N@1</th>
<th>N@3</th>
<th>N@5</th>
<th>N@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libra</td>
<td>0.5104</td>
<td>0.4800</td>
<td>0.4634</td>
<td>0.4467</td>
<td>0.4978</td>
</tr>
<tr>
<td>Rexa</td>
<td>0.4621</td>
<td>0.4512</td>
<td>0.4332</td>
<td>0.4236</td>
<td>0.4798</td>
</tr>
<tr>
<td>PSVM</td>
<td>0.9374</td>
<td>0.6071</td>
<td>0.4790</td>
<td>0.3934</td>
<td>0.3385</td>
</tr>
</tbody>
</table>

### Features and Weights

<table>
<thead>
<tr>
<th>Features</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>2.7201</td>
</tr>
<tr>
<td>L10</td>
<td>-2.5080</td>
</tr>
<tr>
<td>H2</td>
<td>2.5018</td>
</tr>
<tr>
<td>H3</td>
<td>1.9956</td>
</tr>
<tr>
<td>H1</td>
<td>-1.5827</td>
</tr>
<tr>
<td>L2</td>
<td>1.5122</td>
</tr>
<tr>
<td>S4</td>
<td>1.1284</td>
</tr>
<tr>
<td>L9</td>
<td>1.0525</td>
</tr>
<tr>
<td>S2</td>
<td>-0.9863</td>
</tr>
<tr>
<td>L6</td>
<td>0.6276</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Approach</th>
<th>P@5</th>
<th>P@10</th>
<th>P@15</th>
<th>MAP</th>
<th>N@5</th>
<th>N@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSVM</td>
<td>.7714</td>
<td>.8429</td>
<td>.8285</td>
<td>.7756</td>
<td>.5545</td>
<td>.5947</td>
</tr>
<tr>
<td>RSVMt</td>
<td>.8000</td>
<td>.8286</td>
<td>.8476</td>
<td>.7837</td>
<td>.5923</td>
<td>.5999</td>
</tr>
<tr>
<td>MTRSVM</td>
<td>.8000</td>
<td>.8286</td>
<td>.8476</td>
<td>.7875</td>
<td>.6140</td>
<td>.6075</td>
</tr>
<tr>
<td>HCDRank</td>
<td>.8285</td>
<td>.7857</td>
<td>.8571</td>
<td>.7971</td>
<td>.6189</td>
<td>.6112</td>
</tr>
<tr>
<td>Language model</td>
<td>.6250</td>
<td>.6875</td>
<td>.6500</td>
<td>.6726</td>
<td>.3343</td>
<td>.3809</td>
</tr>
</tbody>
</table>
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?
- Can we predict the user’s actions?
Topic-based Social Influence Analysis

- Social network -> Topical influence network
### Influential nodes on different topics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Topic</th>
<th>Representative Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Pat Langley, Alex Warbel, Trevor Darrell, C. Lee Giles, Terrence J. Sejnowski, Samy Bengio, Daphne Koller, Luc De Raedt, Vasant Honavar, Floriana Esposito, Bernhard Scholzpf</td>
<td>-</td>
</tr>
<tr>
<td>Database System</td>
<td>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</td>
<td></td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Frieder, Alan F. Smeaton, Hong Jin</td>
<td></td>
</tr>
<tr>
<td>Web Services</td>
<td>Yan Wang, Liang-je Zhang, Saharan Dustdar, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah</td>
<td></td>
</tr>
<tr>
<td>Semantic Web</td>
<td>Wolfgang Nejdl, Daniel Schwabe, Sefwen Staab, Mark A. Musen, Andrew Tomkins, Julianna Freire, Carole A. Goble, James A. Hendler, Rudi Studer, Enrico Motta</td>
<td></td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets</td>
<td></td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td><strong>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 Multi-level Association Rules from Large Databases, Interleaving a Join Sequence with Semijoins in Distributed Query Processing</strong></td>
<td></td>
</tr>
<tr>
<td>Web Mining</td>
<td><strong>faCT and faCT, 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 Ds</strong></td>
<td></td>
</tr>
</tbody>
</table>
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?
- Can we predict the user’s actions?
Mining Advisor-Advisee Relationship from Research Publication Networks

Input: Temporal collaboration network

Output: Relationship analysis

(0.9, [/, 1998])
(0.4, [/, 1998])
(0.2, [2001, 2003])
(0.65, [2002, 2004])
(0.49, [1999, 2000])
(0.7, [2000, 2001])
(0.5, [/, 2000])
(0.9, [/, 1998])
(0.8, [1999, 2000])
Results
Results (cont.)

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

<table>
<thead>
<tr>
<th>data set</th>
<th>RULE</th>
<th>SVM</th>
<th>IndMAX</th>
<th>TPFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST1</td>
<td>69.9%</td>
<td>73.4%</td>
<td>75.2%</td>
<td>78.9%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>74.6%</td>
<td>79.0%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>83.1%</td>
<td>90.9%</td>
</tr>
</tbody>
</table>

Advisee | Top Ranked Advisor | Time | Note
--------|---------------------|------|------
| 2. Jiawei Han | 2004-2008 | PhD advisor, 2008 |

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?
• Can we predict the user’s actions?
What can we do in SNS?

facebook

News Feed

flickr

mike.panchenko’s photostream

twitter

Bill Gates
1. Always watch news
2. Enjoy sports
3. …
# Results

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

<table>
<thead>
<tr>
<th>Data set</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>SVM</td>
<td>10.41</td>
<td>16.71</td>
<td>13.85</td>
</tr>
<tr>
<td>Twitter</td>
<td>wvRN</td>
<td>0.45</td>
<td>7.89</td>
<td>0.86</td>
</tr>
<tr>
<td>Twitter</td>
<td>NTT-FGM</td>
<td>26.40</td>
<td>21.14</td>
<td>23.47</td>
</tr>
<tr>
<td>Flickr</td>
<td>SVM</td>
<td>34.48</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Flickr</td>
<td>wvRN</td>
<td>60.02</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Flickr</td>
<td>NTT-FGM</td>
<td>56.18</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>ArnetMiner</td>
<td>SVM</td>
<td>10.19</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>ArnetMiner</td>
<td>wvRN</td>
<td>14.83</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>ArnetMiner</td>
<td>NTT-FGM</td>
<td>31.14</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.
Arnetminer Today
— A brief summary
# ArnetMiner’s History

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>New Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006/5</td>
<td>V0.1</td>
<td>Profile extraction, person/paper/conf. search</td>
</tr>
<tr>
<td>2006/8</td>
<td>V1.0</td>
<td>Rewritten all codes in Java</td>
</tr>
<tr>
<td>2007/7</td>
<td>V2.0</td>
<td>Survey search, research interest, association search</td>
</tr>
<tr>
<td>2008/4</td>
<td>V3.0</td>
<td>Query understanding, New search GUI, log analysis</td>
</tr>
<tr>
<td>2008/11</td>
<td>V4.0</td>
<td>Graph search, topic mining, NSFC/NSF</td>
</tr>
<tr>
<td>2009/4</td>
<td>V5.0</td>
<td>Bole/course search, profile editing, open resources, #citation</td>
</tr>
<tr>
<td>2009/12</td>
<td>V6.0</td>
<td>Academic statistics, user feedbacks, refined ranking</td>
</tr>
<tr>
<td>2010/5</td>
<td>V7.0</td>
<td>Name disambiguation, reviewer assignment, supervisor suggestion, open API</td>
</tr>
<tr>
<td>2010</td>
<td>V8.0</td>
<td>ArnetApp Platform</td>
</tr>
</tbody>
</table>
ArnetMiner Today

* Arnetminer data:
  > 1.5 M researcher profiles
  > 3M papers
  > 34M citation relationships
  > 8K conferences
  > 80M logs

* > 1,000,000 users from more than 200 countries
* Continuously 20+% increase of visits per month

* >300,000 page views per day
**ArnetMiner Today**

* Arnetminer data:
  > 1.5 M researcher profiles
  > 3M papers
  > 34M citation relationships
  > 8K conferences
  > 80M logs

* Continuously 20+% increase of visits per month
* >300,000 page views per day

---

**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]

---

**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.
Arnetminer’s User Distribution

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

- The largest publisher: Elsevier
- Conferences
  - KDD 2010
  - KDD 2011
  - WSDM 2011
  - ICDM 2011
  - 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…

**Arnetminer**

- **PatentMiner**
  - Mining knowledge from patents:
    - competitor analysis
    - company search
    - patent summarization

- **QQMiner**
  - Mining “QQ”
    - Association search
    - Influence analysis
    - Hot topic detection

- **PubmedMiner**
  - Mining Pubmed data
    - Expert finding
    - Ranking subgraphs
    - Novel search
    - Instant search

- **...**
  - Mining more data…
PatentMiner
Patent Search

Summary

Inventor

Company

Patents on “3d seismic data”
Patent Miner

Microsoft Corporation

Trend Analysis:

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

Patent Trend

- patent trends

Competitors:

- Research in Motion Limited
  # patents: 827
- Silverbrook Research Pty Ltd
  # patents: 2831
- Nokia Corporation
  # patents: 3069
- International Business Machines Corporation
  # patents: 50920
- Hewlett-Packard Development Company, L.P.
  # patents: 11033

1990-2000

- Research in Motion Limited
  # patents: 887
- Silverbrook Research Pty Ltd
  # patents: 3069
- Nokia Corporation
  # patents: 50920
- International Business Machines Corporation
  # patents: 11033

2001-2005

- Research in Motion Limited
  # patents: 603
- NEC Corporation
  # patents: 21653
- Samsung Electronics Co., Ltd.
  # patents: 24001
- Sony Ericsson Mobile Communications AB
  # patents: 423
- International Business Machines Corporation
  # patents: 50920

2006-2010

Top 5 Inventor Trend

- Hong Xiang Zhang
- John E. Dourcet
- Heung Yeung Shum
- Paul England
- Ramarathnam Venkatesan
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.
QQMIner
蔡卓妍

兴趣：成长 / 恋爱 / 生活 / 日常

陈果

兴趣：成长 / 恋爱 / 生活 / 动物 / 生活 / 学习 / 社会 / 美食

Social Graph

Paths
Topic 0: 小说 / 历史
热门作者
蔡琴 邓紫棋 晋美高 陈若兰 蒋诗敏 萧安泰
热门Blog

Topic 1: 音乐 / 食物
热门作者
陈慧娴 蔺海冰 多啦才吉 东来 邓雅芝 陈锦贤
热门Blog

Topic 2: 美容 / 女性
热门作者
星琴 星依林 邓倩倩 陈彩君 蒋雅琪 童娜
热门Blog

Topic 3: 男女 / 爱情
热门作者
范广惠 范方 阮韩慧 邓丽欣 陈悠然 范可欣
热门Blog

Topic 4: 生活 / 邮寄

Topic Trend

活跃用户 in Topic 29: 地震 / 救灾

图示显示了Topic 29: 地震 / 救灾下活跃用户的分布。
Opportunity: exploiting social network and semantic web in the real-world

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

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

- 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 3 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, 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.*

*Others: ICDM’07-10, CIKM’07-11, SDM’09, ISWC’06&09, DKE, JIS, etc.*
Thanks!

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