Colorful Social Networks: Relationship Mining in Social Networks

Jie Tang
Tsinghua University, China

Collaborate with
Chi Wang and Jiawei Han (UIUC), Jeffrey Xu Yu (CUHK)
Zhong Su, et al. (IBM)
Social Networks

Web 1.0 (1989)
Pages, hyperlinks
Relevance search

Web 2.0 (2004)
Social networks

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.
Colorful Social Relationship?

- **Social link** (relationship) is one of the most important information in the social network.
- However, although one’s relationship should be *colorful*, e.g., “family”, “co-workers”, “classmates”, **users do not** take the time to create it.
  - E.g., only 16% of mobile phone users in Europe have created custom contact groups

- The fact is that our social network is **black** white...
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
Questions:  
- How to combine social links, contents, and prior knowledge for discovering the colorful social relationship?
Mining Advisor-advisee Relationship from Research Publication Network
This Work: Advisor-advisee

- Input: research publication network.
- Output: potential advising relationships and their probabilities—(prob, [s_time, e_time])

---

Visualized chorological hierarchies

Diagram showing temporal collaboration network and relationship analysis.
Objective: extract semantic meaning from plain links to finely model and better organize information networks

Challenge?

- Latent semantic knowledge
- Volatility
- Interdependency
- Scalability

Opportunity?

- Human intuition
- Propagate constraint
- Crosscheck using network
- Collective intelligence

Methodology: propagate simple intuitive rules and constraints over the whole network
Overall Framework

- $a_i$: author $i$
- $p_j$: paper $j$
- $py$: paper year
- $pn$: paper#  
- $st_{i,yi}$: starting time
- $ed_{i,yi}$: ending time
- $r_{i,yi}$: probability
Local features - preprocess

- For every pair of coauthors $a_i$ and $a_j$
  - Create a potential link from $a_i$ to $a_j$ if $a_j$ has a longer publication history than $a_i$
  - Compute Kulczynski and Imbalance Ratio measure for the coauthored publications at different time $t$
  - Estimate the advising time
    - $S_{tij} = \text{the start time of coauthorship}$
    - $E_{tij} = \text{the time point when correlation drops}$
      - $\text{YEAR1: } \text{Kul}_{ij}^t > \text{Kul}_{ij}^{t+1}$
      - $\text{YEAR2: } \max(\text{Kul}_{ij}^t - \text{Kul}_{ij}^{t+1})$
  - Remove the link if certain rules apply, o.w. sum average Kul and IR as a rough likelihood
Why is network structure helpful?

• More than pairwise features: interdependency

one's advisor depends on others’ advisor!
Basic constraints

- $a_y$ can only advise $a_x$ after it graduated
  \[ ed_y < st_x < ed_x \] If $a_y$ advises $a_x$
- If $a_y$ advises $a_x$ since the year $st_x$, $a_y$ must have a longer history of publication than $a_x$ before $st_x$.
  \[ \Rightarrow \text{The candidate graph H’ is a DAG.} \]

The model can incorporate other intuitions as factor functions.
Time-constrained Probabilistic Factor Graph (TPFG)

- Hidden variable $y_x$: $a_x$’s advisor
- $st_{x,y_x}$: starting time
- $ed_{x,y_x}$: ending time
- $g(y_x, st_x, ed_x)$ is pairwise local feature
- $f_x(y_x, Z_x) = \max g(y_x, st_x, ed_x)$ under time constraint
- Objective function $P(\{y_x\}) = \prod_x f_x(y_x, Z)$
- $Z = \{z| x \in Y_x\}$
- $Y_x$: set of potential advisors of $a_x$
Inference algorithm of TPFG

- \( r_{ij} = \max P(y_1, \ldots, y_{na}|y_i = j) = \exp (\text{sent}_{ij} + \text{recv}_{ij}) \)
Experiment Results

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

<table>
<thead>
<tr>
<th>Datasets</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>80.2%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>74.6%</td>
<td>81.5%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>83.1%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

- Heuristics
- Supervised learning
- Empirical parameter
- Optimized parameter

Supervised learning
## Case Study & Scalability

<table>
<thead>
<tr>
<th>Advisee</th>
<th>Top Ranked Advisor</th>
<th>Time</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>David M. Blei</td>
<td>1. Michael I. Jordan</td>
<td>01-03</td>
<td>PhD advisor, 2004 grad</td>
</tr>
<tr>
<td></td>
<td>2. John D. Lafferty</td>
<td>05-06</td>
<td>Postdoc, 2006</td>
</tr>
<tr>
<td>Hong Cheng</td>
<td>1. Qiang Yang</td>
<td>02-03</td>
<td>MS advisor, 2003</td>
</tr>
<tr>
<td></td>
<td>2. Jiawei Han</td>
<td>04-08</td>
<td>PhD advisor, 2008</td>
</tr>
<tr>
<td>Sergey Brin</td>
<td>1. Rajeev Motawani</td>
<td>97-98</td>
<td>“Unofficial advisor”</td>
</tr>
</tbody>
</table>

![Graphs](image)
Effect of rules - ROC curve

- Filtering rules in TPFG

  \( R1: IR_{ij}^t < 0 \) in the sequence \( \{IR_{ij}^t\}_t \) during the collaboration period of \( a_i \) and \( a_j \),

  \( R2: \) there is no increase in the sequence \( \{kulc_{ij}^t\}_t \) during the collaboration period,

  \( R3: \) the collaboration period of \( a_i \) and \( a_j \) lasts only for one year,

  \( R4: py_{ij}^1 + 2 > py_{ij}^1 \),

- Local feature measure:
  KULC and IR
Effect of network depth

- Different closures of given set of nodes
  - DFS with bounded maximal depth $d$: $d$-closure
Exact VS approximate inference

• Exact inference of TPFG
  – JuncT: Junction Tree + Sum-Product

• Approximate inference
  – LBP: Loopy Belief Propagation
  – TPFG: the proposed message passing algorithm
Application: visualization

RULE

TPFG
Bole Search In Arnetminer

An example on a real system: Arnetminer

Performance improvement

![Graph showing performance improvement with different metrics: P@2, P@5, MAP, NDCG@5 for three different systems: LM, LM+RULE, LM+TPFG.](image)
Ongoing and future work

• “Replying” relationship of forum posts
  – Input: forum posts without replying relation.
  – Output: a thread tree organized by their replying hierarchy

• More General: hierarchical relationship
  – Automatic feature/rule extraction
  – Generic form of constraints
  – Supervised information incorporation

• Applying to other problems
  – Hierarchical clustering; network summarization; etc.
Topic Distributions over Links on Web
Motivation

• Web users create links with significantly different intentions
• Understanding of the category and the influence of each link can benefit many applications, e.g.,
  – Expert finding
  – Collaborator finding
  – New friends recommendation
  – …
Examples – Topic distribution analysis over citations

- an in-depth understanding of the research field?

Researcher A

- An Inverted Index Implementation
- Introduction of Modern Information Retrieval
- Self-Indexing Inverted Files for Fast Text Retrieval
- Signature les: An access Method for Documents and its Analytical Performance Evaluation
- Parameterised Compression for Sparse Bitmaps
- Vector-space Ranking with Effective Early Termination
- Efficient Document Retrieval in Main Memory
- Filtered Document Retrieval with Frequency-Sorted Indexes
- Memory Efficient Ranking
- A Document-centric Approach to Static Index Pruning in Text Retrieval Systems
- Static Index Pruning for Information Retrieval Systems

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

Citation Relationship Type
- Basic theory
- Comparable work
- Other
Problem: Link Semantic Analysis

Citation context words

Introduction

...Our method increases query throughput by 15% over the method proposed by Anh and Moffat [3] while still retaining rank safe results...

...We show that storing inverted lists in memory can significantly improve performance, adding to previous results from Butcher and Clarke [7].

Related Work

Impact-sorted indexes are described in a series of papers by Anh and Moffat [1, 2, 3]...

Fagin et al. proposed a class of algorithms known as threshold algorithms [12]. These algorithms, like the ones shown in this paper, ...

In our work, we process less index data by organizing the index for easy skipping and query termination. Another way to process less data is to store less data in the index. Static pruning methods remove information from the index that is unlikely to affect query effectiveness. Carmel et al. considered this process [9]. More recently...

Approach

A similar technique has been used previously by Buettcher and Clarke [7]...

Moffat and Zobel [14] suggest a method for determining this parameter. We use a slightly different formulation in this paper...

Cited/Target paper


[9] Static Index Pruning for Information Retrieval Systems

[14] Self-Indexing Inverted Files for Fast Text Retrieval

Topic modeling over links

Citation Relationship Type

Influence Strength

Weak

Middle

Strong

Efficient Document Retrieval in Main Memory

Topics distribution: 55% 20% 13% 12%

Topics

- Topic 31: Ranking and Inverted Index
- Topic 27: Information retrieval
- Topic 23: Index method
- Other

Citation context for the target paper [3]
Outline

• Our Approach
  – Pairwise Restricted Boltzmann Machines (PRBMs)

• Experimental Results

• Conclusion & Future Work
Pairwise Restricted Boltzmann Machines (PRBMs)

- Pairwise Restricted Boltzmann Machines (PRBMs)
- Link category
- Latent variables defined over the link to bridge the two pages
- Example
Formalization of PRBMs

**Formalization**

**Obj. Func:** \[ \mathcal{L}_{gen}(D) = - \sum_{e=1}^{N} \log p(w_e, c_e) \]

with

\[ p(w_e, c_e) = \frac{1}{Z} \sum_{h_e} \sum_{z^s} \sum_{z^t} \exp(-E(z^s_e, z^t_e, w_e, c_e, h_e)) \]

\[ E(z^s_e, z^t_e, w_e, c_e, h_e) = - \sum_{j=1}^{T} \sum_{k=1}^{K} U^t_{jk} z^t_e h_{ek} - \sum_{i=1}^{W} \sum_{k=1}^{K} U^e_{ik} w_{ei} h_{ek} - \sum_{l=1}^{C} \sum_{k=1}^{K} U^c_{lk} c_{el} h_{ek} \]

\[ - \sum_{j=1}^{T} b_j z^s_e - \sum_{j=1}^{T} b_j z^t_e - \sum_{i=1}^{W} o_i w_{ei} - \sum_{k=1}^{K} g_k h_{ek} - \sum_{l=1}^{C} s_l c_{el} \]
Model Learning

Generative learning

\[ \frac{\partial L_{\text{gen}}(D)}{\partial \theta} = \mathbb{E}_{P_0}[\frac{\partial}{\partial \theta} E(D)] - \mathbb{E}_{P_M}[\frac{\partial}{\partial \theta} E(\hat{D})] \]

Discriminative learning

Obj. Func:

\[ L_{\text{dis}}(D) = -\sum_{d=1}^{M} \ln p(w_d) - \sum_{e=1}^{N} \log p(c_e | z^s_e, z^t_e, w_e) \]

\[ \frac{\partial L_{\text{dis}}(D)}{\partial U^s_{jk}} = \mathbb{E}_{P_0}[z^s_{ej} \hat{h}_{ek}] - \mathbb{E}_{P_T}[z^s_{ej} \hat{h}_{ek}] \]

\[ \frac{\partial L_{\text{dis}}(D)}{\partial U^t_{jk}} = \mathbb{E}_{P_0}[z^t_{ej} \hat{h}_{ek}] - \mathbb{E}_{P_T}[z^t_{ej} \hat{h}_{ek}] \]

\[ \frac{\partial L_{\text{dis}}(D)}{\partial U^e_{jk}} = \mathbb{E}_{P_0}[w_{ei} \hat{h}_{ek}] - \mathbb{E}_{P_T}[w_{ei} \hat{h}_{ek}] \]

Hybrid learning

\[ L_{\text{hybrid}}(D) = \alpha L_{\text{gen}}(D) + L_{\text{dis}}(D) \]

We use the Contrast Divergence to learn the model distribution \( P_M \).
Link Semantic Analysis

• Link category annotation
  – First we calculate $p(z_j = 1|w) = \sigma(\sum_i U_{ij} w_i + a_j)$
  – Then we estimate the probability $p(c|e)$ by a mean field algorithm

• Link influence estimation
  – Estimate influence by KL divergence
    \[
    f = KL(\phi^d \parallel \phi^t) = \sum_{k=1}^{K} \phi_k^d \log \frac{\phi_k^d}{\phi_k^t}
    \]
  – An alternative way is to generate the influence score by a Gaussian distribution, thus
    \[
    p(f|h_e) = \text{Gaussian}(f|\sum_{k=1}^{K} U_k^f h_{ek} + r, 1)
    \]
Experimental Setting

• Data sets
  – Arnetminer data: 978,504 papers, 14M citations
  – Wikipedia: 14K “article” pages and 25K links

• Evaluation measures
  – Link categorization accuracy
  – Topical analysis

• Baselines:
  – SVM+LDA
  – SVM+RBM
## Accuracy of Link Categorization

<table>
<thead>
<tr>
<th>Approach</th>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>drill down</td>
<td>64.09</td>
<td>53.60</td>
<td>58.24</td>
<td>88.50</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>77.66</td>
<td>83.34</td>
<td>80.33</td>
<td>76.59</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>64.63</td>
<td>59.58</td>
<td>61.77</td>
<td>79.71</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>68.79</td>
<td>65.51</td>
<td>66.78</td>
<td>81.60</td>
</tr>
<tr>
<td>SVM+LDA</td>
<td>drill down</td>
<td>66.31</td>
<td>58.56</td>
<td>61.71</td>
<td>89.24</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>80.69</td>
<td>85.47</td>
<td>82.93</td>
<td>79.71</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>69.58</td>
<td>64.36</td>
<td>66.75</td>
<td>82.42</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>72.19</td>
<td>69.46</td>
<td>70.46</td>
<td>83.79</td>
</tr>
<tr>
<td>SVM+RBM</td>
<td>drill down</td>
<td>68.63</td>
<td>53.00</td>
<td>59.66</td>
<td>89.40</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>76.47</td>
<td>91.74</td>
<td>83.36</td>
<td>78.89</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>76.65</td>
<td>53.28</td>
<td>62.54</td>
<td>82.58</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>73.92</td>
<td>66.02</td>
<td>68.52</td>
<td>83.62</td>
</tr>
<tr>
<td>gPRBM</td>
<td>drill down</td>
<td>69.77</td>
<td>65.22</td>
<td>67.42</td>
<td>88.07</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>77.33</td>
<td>87.22</td>
<td>81.98</td>
<td>79.01</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>78.00</td>
<td>60.94</td>
<td>68.42</td>
<td>85.19</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>75.03</td>
<td>71.12</td>
<td>72.61</td>
<td>84.09</td>
</tr>
<tr>
<td>dPRBM</td>
<td>drill down</td>
<td>96.00</td>
<td>44.44</td>
<td>61.54</td>
<td>95.88</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>76.85</td>
<td>96.89</td>
<td>85.71</td>
<td>78.60</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>75.00</td>
<td>37.50</td>
<td>50.00</td>
<td>80.25</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td><strong>82.62</strong></td>
<td>59.61</td>
<td>65.75</td>
<td>84.91</td>
</tr>
<tr>
<td>hPRBM</td>
<td>drill down</td>
<td>65.38</td>
<td>68.00</td>
<td>66.67</td>
<td>93.00</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>83.54</td>
<td>92.31</td>
<td>87.71</td>
<td>84.77</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>88.14</td>
<td>69.33</td>
<td>77.61</td>
<td>87.65</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>79.02</td>
<td><strong>76.55</strong></td>
<td><strong>77.33</strong></td>
<td><strong>88.48</strong></td>
</tr>
</tbody>
</table>

**gPRBM**: our approach with generative learning  
**dPRBM**: our approach with discriminative learning  
**hPRBM**: our approach with hybrid learning
## Category-Topic Mixture

<table>
<thead>
<tr>
<th>Topics ((z^s))</th>
<th>Topics ((z^t))</th>
<th>Influential Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>of Citing Papers</td>
<td>of Cited Papers</td>
<td></td>
</tr>
<tr>
<td><strong>Drill down (Basic theory)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 37</td>
<td>Topic 51</td>
<td>5.05</td>
</tr>
<tr>
<td>(Probabilistic model)</td>
<td>(Maximum entropy)</td>
<td></td>
</tr>
<tr>
<td>Topic 13</td>
<td>Topic 26</td>
<td>4.36</td>
</tr>
<tr>
<td>(Semantic Web)</td>
<td>(Ontology)</td>
<td></td>
</tr>
<tr>
<td><strong>Similar (Comparable work)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 7</td>
<td>Topic 7</td>
<td>0.00</td>
</tr>
<tr>
<td>(Recommendation)</td>
<td>(Recommendation)</td>
<td></td>
</tr>
<tr>
<td>Topic 21</td>
<td>Topic 21</td>
<td>0.00</td>
</tr>
<tr>
<td>(Frequent pattern learning)</td>
<td>(Frequent pattern learning)</td>
<td></td>
</tr>
<tr>
<td>Topic 26</td>
<td>Topic 26</td>
<td>0.00</td>
</tr>
<tr>
<td>(Ranking &amp; Inverted Index)</td>
<td>(Ranking &amp; Inverted Index)</td>
<td></td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 26</td>
<td>Topic 43</td>
<td>4.87</td>
</tr>
<tr>
<td>(Ranking &amp; Inverted Index)</td>
<td>(Information retrieval)</td>
<td></td>
</tr>
<tr>
<td>Topic 26</td>
<td>Topic 48</td>
<td>4.24</td>
</tr>
<tr>
<td>(Ranking &amp; Inverted Index)</td>
<td>(Web mining)</td>
<td></td>
</tr>
<tr>
<td>Topic 9</td>
<td>Topic 18</td>
<td>4.69</td>
</tr>
<tr>
<td>(Clustering)</td>
<td>(Natural Language Processing)</td>
<td></td>
</tr>
</tbody>
</table>
## Example Analysis

<table>
<thead>
<tr>
<th>Source Paper</th>
<th>Target Paper</th>
<th>Influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drill Down (Basic theory)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent dirichlet allocation</td>
<td>A variational bayesian framework for graphical models</td>
<td>2.36</td>
</tr>
<tr>
<td>A sentiment-aware model for predicting sales</td>
<td>Probabilistic latent semantic analysis</td>
<td>1.93</td>
</tr>
<tr>
<td>Mining and summarizing customer reviews</td>
<td>Foundations of statistical natural language processing</td>
<td>3.25</td>
</tr>
<tr>
<td>Similar (Comparable work)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constraint-driven clustering</td>
<td>Max-min d-cluster formation in wireless ad hoc networks</td>
<td>2.00</td>
</tr>
<tr>
<td>Hmms and coupled hmms for multi-channel eeg classification</td>
<td>Gaussian observation hidden markov models for eeg analysis</td>
<td>0.51</td>
</tr>
<tr>
<td>Scalable collaborative filtering using cluster-based smoothing</td>
<td>Evaluating collaborative filtering recommender systems</td>
<td>1.51</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent dirichlet allocation</td>
<td>Overview of the first text retrieval conference</td>
<td>3.93</td>
</tr>
<tr>
<td>Utility scoring of product reviews</td>
<td>Introduction to modern information retrieval</td>
<td>3.16</td>
</tr>
<tr>
<td>Sentiment analyzer: extracting sentiments about a given topic</td>
<td>Finding parts in very large corpora</td>
<td>2.71</td>
</tr>
</tbody>
</table>
Conclusion & Future Work

• Concluding remarks
  – Investigate the problem of quantifying link semantics on the Web
  – Propose a Pairwise Restricted Boltzmann Machines to solve this problem

• Future Work
  – Correlation between the link semantics and the information propagation
  – Apply the link semantics to social prediction
**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, MLJ, DKE, JIS; KDD’08-10, SDM’09, ICDM’07-09, CIKM’07-10

http://arnetminer.org/
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

* ArnetMiner data:
  > 1 M researcher profiles
  > 3M papers
  > 34M citation relationships
  > 8K conferences
  > 50M logs

* Visits come from more than 190 countries
* Continuously +20% increase of visits per month
* More than 2,000 unique IP visits 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]

... ArnetMiner is one of my favorite tools to find folk and academic relatives… [from ...@qlink.com]

Dear Dr. Jie Tang,
Can you include our papers (http://www.waset.org) in your Arnetminer?...
[from ...@waset.org]

... top top! I am very interested in your ArnetMiner. Is that possible give me a bit of your social network data… [from ...@cse.ust.hk]

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

Data Mining and Social Network techniques

Scientific Literature
Users cover >180 countries
>600K researchers
>3M papers
Arnetminer.org (NSFC, 863)

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,
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, 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.
- 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!

Q&A

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