AMiner II — Toward Understanding
Big Scholar Data

Jie Tang
AMiner II — Toward Understanding Big Scholar Data

@2006-2015, http://aminer.org

Jie Tang
Tsinghua University
Mining Knowledge from Big Data

Drown or Survive?
- Researcher profile extraction
- Expert finding
- Social network search
- Topic browser
- Conference analysis
- ArnetApp platform
Person Search

Basic Info.

Citation statistics

Research Interests

Ego network
Expert Search

Finding experts, for “data mining”

Demographics: gender, language, location, etc.

Knowledge about “data mining”

Similar authors
Reviewer Suggestion

- Interest matching
- COI avoiding
- Load balancing
- Forecast review quality
Reviewer Suggestion
AMiner II (ArnetMiner)

- Academic Social Network Analysis and Mining system—AMiner (http://aminer.org)
  - Online since 2006
  - >38 million researcher profiles
  - >76 million publication papers
  - >241 million requests
  - >12.35 Terabyte data
  - 100K IP access from 170 countries per month
  - 10% increase of visits per month
- Deep analysis, mining, and search
User Distribution

6.32 million IP from 220 countries/regions
User Distribution

6.32 million IP from 220 countries/regions

Top 10 countries
1. USA 6. Canada
2. China 7. Japan
3. Germany 8. Spain
4. India 9. France
5. UK 10. Italy
Technologies
—Toward understanding big scholar data

Recent progress…
Knowledge Acquisition from the Web

(ACM TKDD, WWW’12, ISWC’06, ICDM’07, ACL’07)
Researcher Profile Database[1]

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

Is this Enough?
Required Semantics are distributed in Multiple Sources

LinkedIn

Wikipedia

Jeannette Wing

From Wikipedia, the free encyclopedia

Pittsburgh

Jeannette Marie Wing is Corporate Vice President of Microsoft Research with oversight of its core research laboratories around the world and Microsoft Research Connections.[1][2] Prior to 2013, she was the President’s Professor of Computer Science at Carnegie Mellon University, Pittsburgh, Pennsylvania, United States. She also served as assistant director for Computer and Information Science and Engineering at the NSF from 2007 to 2010.[3][4][5][6][7][8][9][10][11][12][13]

Contents

- Education
- Career and research
- References
- External links

Background

Experience

Corporate Vice President
Microsoft
July 2013 – Present (2 years 1 month) | Redmond
In charge of core research labs at Microsoft Research.

President's Professor of Computer Science
Carnegie Mellon University, Computer Science
1985 – Present (30 years)

Education

Wing earned her B.S. and S.M. in Electrical Engineering and Computer Science at MIT in June 1979. Her advisers were Ronald Rivest and John Reiser. In 1983, she earned her Ph.D. in Computer Science at MIT under John Guttag.[11]

Born
Jeannette Marie Wing

Nationality
American

Fields
Computer science

Institutions
Carnegie Mellon University

Aims mater
Massachusetts Institute of Technology

Thesis
A Two-Tiered Approach to Specifying Programs (1983)
Network Integration

• Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.
COSNET: Connecting Social Networks with Local and Global Consistency

• **Input:** $G=\{G^1, G^2, \ldots, G^m\}$, with $G^k=(V^k, E^k, R^k)$

• **Formalization:** $X=\{x_i\}$, all possible pairwise matchings and each corresponds to $y_i \in \{1, 0\}$

• **COSNET:** an energy-based model

\[
Y^* = \text{arg max } E(Y, X)
\]

[1] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. KDD’15.
Local vs. Global consistency

• Given three networks,
Local vs. Global consistency

- Local matching: matching users by profiles

Pairwise similarity features
- Username similarity and uniqueness
- Profile content similarity
- Ego network similarity
- Social status

Energy function

\[ E_l(Y, X) = \sum_i w_i^T g_l(x_i, y_i) \]
Local vs. Global consistency

• Network matching: matching users’ ego networks

Encourage “neighborhood-preserving matching”
Network Matching

• Network matching: matching users’ ego networks

Input networks

$G_1$

$v^1_1$

$v^2_1$

$v^3_1$

$v^4_1$

$G_2$

$v^1_2$

$v^2_2$

$v^3_2$

$v^4_2$

Matching graph

Matching graph

$C$

$\begin{align*}
E_c(Y, X) &= \sum_{\langle x_i, x_j \rangle \in E_{MG}} w^T_e f_c(y_i, y_j) \\
&= \begin{cases}
(1, 0, 0)^T & \text{if } y_i = y_j = 0 \\
(0, 1, 0)^T & \text{if } y_i + y_j = 1 \\
(0, 0, 1)^T & \text{if } y_i = y_j = 1
\end{cases}
\end{align*}$
Global vs. Global consistency

• Global consistency: matching users by avoiding global inconsistency

**Definition 2.** (Global Inconsistency). Given a set of social networks $G$, a set of user pairs $X$ and the corresponding labels $Y$, if there exists a sequence of user pairs $(x_{i_1}, x_{i_2}, \ldots , x_{i_n})$, such that

$$\forall i = i_1, i_2, \ldots , i_n, y_i = 1$$

and

$$\forall k = 1, 2, \ldots , n - 1, \ V_{i_k}^2 = V_{i_{k+1}}^1$$

and

For the pair $(V_{i_n}^2, V_{i_1}^1)$, if the corresponding label $y_j = 0$

then we say that the assigned labels $Y$ causes global inconsistency given $G$ and $X$.

Avoid “global inconsistency”
Avoid global inconsistency

Energy function

\[ E_t(Y, X) = \sum_{c \in T_M G} w^t f_t(Y_c) \]

\[ f_t(y_i, y_j) = \begin{cases} 
(1, 0, 0, 0)^T & \text{if } |Y_c| = 0 \\
(0, 1, 0, 0)^T & \text{if } |Y_c| = 1 \\
(0, 0, 1, 0)^T & \text{if } |Y_c| = 2 \\
(0, 0, 0, 1)^T & \text{if } |Y_c| = 3 
\end{cases} \]
Model Construction

Objective function by combining all the energy functions

\[
E(Y, X) = \sum_{x_i \in V_{MG}} w_i^g g_i(x_i, y_i) + \sum_{(x_i, x_j) \in E_{MG}} w_i^e e(x_i, x_j) + \sum_{c \in T_{MG}} w_c^t t_c(Y_c)
\]  

(2)
Model Learning

- Max-margin learning

$$\min_{W} \frac{1}{2} \|W\|^2 + \mu \xi$$

subject to

$$E(\hat{Y}, X; W) \leq E(Y, X; W) - \Delta(Y, \hat{Y}) + \xi$$

- As the original problem is intractable, we use Lagrangian relaxation to decompose the original objective function into a set of easy-to-solve subproblems

$$E(Y, X; W) = \sum_{f \in F} E_f(Y_f, X_f; W)$$

$$= \sum_{f \in F} \sum_{x_i \in X_f} \left( \frac{1}{|F_i|} w_i^T g_i(x_i, y_i^f) + w_f^T f(Y_f) \right)$$

subject to

$$y_i^f = y_i, \ \forall f, y_i \in Y_f$$
Model Learning (cont.)

- Dual decomposition

\[
L(Y, X, \lambda; W) = \min_W \sum_{f \in \mathcal{F}} \left( \sum_{y_i \in Y_f} \frac{1}{|\mathcal{F}_i|} w_i^T g_i(x_i, y_i^f) + w_f^T f(Y_f) \right) \\
+ \sum_{f \in \mathcal{F}} \sum_{y_i \in Y_f} \lambda_i^f (y_i - y_i^f)
\]

This provides a lower bound to the original function.

\[
\min_{W, \lambda} \frac{1}{2} \|W\|^2 + \mu (E(\hat{Y}, X; W) - \max_\lambda L(Y, X, \lambda; W))
\]

s.t. \( \sum_{y_i \in Y_i} \lambda_i^f = 0, \forall f \in \mathcal{F} \)

The resulting objective function is convex and non-differentiable, and can be solved by projected sub-gradient method.
Results
Connecting AMiner with …

- LinkedIn and VideoLectures

<table>
<thead>
<tr>
<th>Name-match: match name only;</th>
<th>SVM: use classifier to identify the same user;</th>
<th>SiGMA: local propagation;</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNA: an optimization method;</td>
<td>COSNET: our method;</td>
<td>COSNET-: w/o global consistency.</td>
</tr>
</tbody>
</table>
Person Search

LinkedIn

VideoLectures

USPTO
AMiner 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>2008/11</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,</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, open API</td>
</tr>
<tr>
<td>2011/7</td>
<td>V8.0</td>
<td>AMiner, location search, conference analysis</td>
</tr>
<tr>
<td>2012/3</td>
<td>V9.0</td>
<td>New UI, cross-domain collaboration</td>
</tr>
<tr>
<td>2013/5</td>
<td>VII</td>
<td>Knowledge graph, new architecture</td>
</tr>
<tr>
<td>2014/10</td>
<td>VII 2.0</td>
<td>Organization ranking, conference ranking</td>
</tr>
<tr>
<td>2015/4</td>
<td>VII 3.0</td>
<td>Network integration, deep learning</td>
</tr>
</tbody>
</table>
The largest publisher: Elsevier

Conferences:
- KDD 2010
- KDD 2011
- KDD 2012
- WSDM 2011
- ICDM 2011
- ICDM 2012
- SocInfo 2011
- ICMLA 2011
- WAIM 2011
- etc.
AMiner as a platform…

**AMiner**

**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…
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

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

• 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, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. KDD’09. (Top 4 cited papers among KDD 2009’s papers)
• 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, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. KDD’12 (Full Presentation & Best Poster Award)
• Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. KDD’15.
Thanks!

AMiner.org

Jie Tang, KEG, Tsinghua U,
Download data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://aminer.org/download