Cross-domain Collaboration Recommendation

Jie Tang\textsuperscript{1}, Sen Wu\textsuperscript{1}, Jimeng Sun\textsuperscript{2}, Hang Su\textsuperscript{1}

\textsuperscript{1}Tsinghua University
\textsuperscript{2}IBM TJ Watson Research Center
Cross-domain Collaboration

• Interdisciplinary collaborations have generated huge impact, for example,
  – 51 (>1/3) of the KDD 2012 papers are result of cross-domain collaborations between graph theory, visualization, economics, medical inf., DB, NLP, IR
  – Research field evolution
Cross-domain Collaboration (cont.)

- Increasing trend of cross-domain collaborations

Data Mining (DM), Medical Informatics (MI), Theory (TH), Visualization (VIS)
Challenges

1. Sparse Connection: <1%

2. Complementary expertise

3. Topic skewness: 9%

Data Mining

- Large graph
- Heterogeneous network
- Social network

Theory

- Automata theory
- Graph theory
- Complexity theory
Related Work - Collaboration recommendation

- Collaborative topic modeling for recommending papers
  - C. Wang and D.M. Blei. [2011]

- On social networks and collaborative recommendation
  - I. Konstas, V. Stathopoulos, and J. M. Jose. [2009]

- CollabSeer: a search engine for collaboration discovery

- Referral web: Combining social networks and collaborative filtering

- Fab: content-based, collaborative recommendation
  - M. Balabanovi and Y. Shoham. [1997]
Related Work - Expert finding and matching

• Topic level expertise search over heterogeneous networks

• Formal models for expert finding in enterprise corpora
  – K. Balog, L. Azzopardi, and M.de Rijke. [2006]

• Expertise modeling for matching papers with reviewers
  – D. Mimno and A. McCallum. [2007]

• On optimization of expertise matching with various constraints
  – W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li. [2012]
Approach Framework
—Cross-domain Topic Learning
Author Matching

Data Mining

\[ G^S \]

Author

\[ v_1 \]
\[ v_2 \]
\[ \ldots \]
\[ v_N \]

Coauthorships

Medical Informatics

\[ G^T \]

Cross-domain coauthorship

\[ v'_2 \]
\[ \ldots \]
\[ v'_{N'} \]

Query user

\[ r^{(t+1)} = (1 - \tau)S \cdot r^{(t)} + \tau q \]
Topic Matching

Topics Extraction

Data Mining

Medical Informatics

Topics correlations

2. Complementary Expertise!
3. Topic skewness!
Cross-domain Topic Learning

Identify “cross-domain” Topics

Data Mining

Topics

Medical Informatics

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

$G^T$

$v'_1$

$v'_2$

$\ldots$

$v'_{N'}$

$z_1$

$z_2$

$z_3$

$\ldots$

$z_K$
Collaboration Topics Extraction

Step 1:
- Initialize an ACT model in $G^S$ by learning from documents written by authors only from $G^S$.
- Similarly, initialize an ACT model for target domain $G^T$.

Step 2:
- For each collaborated document $d$:
  - Toss a coin $s_{dt}$ according to $\text{bernoulli}(s_{dt}) \sim \text{beta}(\gamma_t, \gamma)$, where $\text{beta}(.)$ is a Beta distribution, and $\gamma_t$ and $\gamma$ are two parameters.
  - If $s_{dt} = 0$ then:
    - Randomly select a pair $(v, v')$ from $d$'s authors, where $v$ is an author from $G^S$ and $v'$ from $G^T$.
    - Draw a topic $z_{dt} \sim \text{multi}(\theta_{vv'})$ from the topic mixture $\theta_{vv'}$ specific to $(v, v')$.
  - If $s_{dt} = 1$ then:
    - Randomly select a user $v$.
    - Draw a topic $z_{dt} \sim \text{multi}(\theta_v)$ from the topic model of user $v$.
- Draw a word $x_{dt} \sim \text{multi}(\phi_{z_{dt}})$ from $z_{dt}$-specific word distribution.
Intuitive explanation of Step 2 in CTL
Experiments
Data Set and Baselines

• Arnetminer (available at http://arnetminer.org/collaboration)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Authors</th>
<th>Relationships</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining</td>
<td>6,282</td>
<td>22,862</td>
<td>KDD, SDM, ICDM, WSDM, PKDD</td>
</tr>
<tr>
<td>Medical Informatics</td>
<td>9,150</td>
<td>31,851</td>
<td>JAMIA, JBI, AIM, TMI, TITB</td>
</tr>
<tr>
<td>Theory</td>
<td>5,449</td>
<td>27,712</td>
<td>STOC, FOCS, SODA</td>
</tr>
<tr>
<td>Visualization</td>
<td>5,268</td>
<td>19,261</td>
<td>CVPR, ICCV, VAST, TVCG, IV</td>
</tr>
<tr>
<td>Database</td>
<td>7,590</td>
<td>37,592</td>
<td>SIGMOD, VLDB, ICDE</td>
</tr>
</tbody>
</table>

• Baselines
  – Content Similarity(Content)
  – Collaborative Filtering(CF)
  – Hybrid
  – Katz
  – Author Matching(Author), Topic Matching(Topic)
# Performance Analysis

**Training:** collaboration before 2001  
**Validation:** 2001-2005

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining(S) to Theory(T)</td>
<td>Content</td>
<td>10.3</td>
<td>10.2</td>
<td>10.9</td>
<td>31.4</td>
<td>4.9</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>15.6</td>
<td>13.3</td>
<td>23.1</td>
<td>26.2</td>
<td>4.9</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>17.4</td>
<td>19.1</td>
<td>20.0</td>
<td>29.5</td>
<td>5.0</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>27.2</td>
<td>22.3</td>
<td>25.7</td>
<td>32.4</td>
<td>10.1</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>28.0</td>
<td>26.0</td>
<td>32.4</td>
<td>33.5</td>
<td>13.4</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>Katz</td>
<td>30.4</td>
<td>29.8</td>
<td>21.6</td>
<td>27.4</td>
<td>11.2</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>CTL</td>
<td>37.7</td>
<td>36.4</td>
<td>40.6</td>
<td>35.6</td>
<td>14.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>

**Content Similarity (Content):** based on similarity between authors’ publications  
**Collaborative Filtering (CF):** based on existing collaborations  
**Hybrid:** a linear combination of the scores obtained by the Content and the CF methods.  
**Katz:** the best link predictor in link-prediction problem for social networks  
**Author Matching (Author):** based on the random walk with restart on the collaboration graph  
**Topic Matching (Topic):** combining the extracted topics into the random walking algorithm
CTL can still maintain about 0.3 in terms of MAP which is significantly higher than baselines.
Parameter Analysis

(a) varying the number of topics $T$
(b) varying $\alpha$ parameter
(c) varying the restart parameter $\tau$ in the random walk
(d) Convergence analysis
Prototype System

http://arnetminer.org/collaborator

Treemap: representing subtopic in the target domain

Recommend Collaborators & Their relevant publications
Conclusion

• Study the problem of cross-domain collaboration recommendation

• Propose the cross-domain topic model for recommending collaborators

• Experimental results in a coauthor network demonstrate the effectiveness and efficiency of the proposed approach
Future work

• Connect cross-domain collaborative relationships with social theories (e.g. social balance, social status, structural hole)

• Apply the proposed method to other networks
Thanks!

System:  http://arnetminer.org/collaborator
Code&Data:  http://arnetminer.org/collaboration
Challenge always be side with opportunity!

• Sparse connection:
  – cross-domain collaborations are rare;

• Complementary expertise:
  – cross-domain collaborators often have different expertise and interest;

• Topic skewness:
  – cross-domain collaboration topics are focused on a subset of topics.

How to handling these patterns?
## Performance Analysis

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Info. (S) to Database (T)</td>
<td>Content</td>
<td>10.1</td>
<td>10.9</td>
<td>12.5</td>
<td>45.9</td>
<td>3.6</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>18.3</td>
<td>20.2</td>
<td>21.4</td>
<td>47.6</td>
<td>5.3</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>25.0</td>
<td>26.5</td>
<td>28.4</td>
<td>59.1</td>
<td>6.4</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>26.2</td>
<td>29.6</td>
<td>32.2</td>
<td>54.8</td>
<td>10.5</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>29.4</td>
<td>26.3</td>
<td>34.7</td>
<td>59.3</td>
<td>11.5</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>Katz</td>
<td>27.5</td>
<td>28.3</td>
<td>30.7</td>
<td>57.2</td>
<td>10.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>CTL</td>
<td>32.5</td>
<td>30.0</td>
<td>36.9</td>
<td>59.8</td>
<td>11.4</td>
<td>5.4</td>
</tr>
</tbody>
</table>

- **Content Similarity (Content)**: based on similarity between authors’ publications
- **Collaborative Filtering (CF)**: based on existing collaborations
- **Hybrid**: a linear combination of the scores obtained by the Content and the CF methods.
- **Katz**: the best link predictor in link-prediction problem for social networks
- **Author Matching (Author)**: based on the random walk with restart on the collaboration graph
- **Topic Matching (Topic)**: combining the extracted topics into the random walking algorithm
## Performance Analysis

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Info.(S) to Data Mining(T)</td>
<td>Content</td>
<td>5.8</td>
<td>5.7</td>
<td>9.5</td>
<td>19.8</td>
<td>1.9</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>13.7</td>
<td>17.8</td>
<td>18.9</td>
<td>34.3</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>18.0</td>
<td>19.0</td>
<td>19.8</td>
<td>36.7</td>
<td>3.4</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>20.1</td>
<td>23.8</td>
<td>29.3</td>
<td>64.4</td>
<td>5.3</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>26.0</td>
<td>25.0</td>
<td>33.9</td>
<td>48.1</td>
<td>10.7</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>Katz</td>
<td>21.2</td>
<td>23.8</td>
<td>32.4</td>
<td>48.1</td>
<td>10.2</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>CTL</td>
<td>30.0</td>
<td>24.0</td>
<td>35.6</td>
<td>49.6</td>
<td>12.2</td>
<td>6.0</td>
</tr>
</tbody>
</table>

- **Content Similarity (Content)**: based on similarity between authors’ publications
- **Collaborative Filtering (CF)**: based on existing collaborations
- **Hybrid**: a linear combination of the scores obtained by the Content and the CF methods.
- **Katz**: the best link predictor in link-prediction problem for social networks
- **Author Matching (Author)**: based on the random walk with restart on the collaboration graph
- **Topic Matching (Topic)**: combining the extracted topics into the random walking algorithm
## Performance Analysis

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual.(S) to Data Mining(T)</td>
<td>Content</td>
<td>9.6</td>
<td>11.8</td>
<td>13.2</td>
<td>18.9</td>
<td>3.1</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>14.0</td>
<td>20.8</td>
<td>26.4</td>
<td>29.4</td>
<td>6.9</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>16.0</td>
<td>20.0</td>
<td>27.6</td>
<td>30.1</td>
<td>6.3</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>22.0</td>
<td>25.2</td>
<td>27.7</td>
<td>31.1</td>
<td>11.9</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>26.3</td>
<td>25.0</td>
<td>32.3</td>
<td>31.4</td>
<td>13.2</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Katz</td>
<td>23.0</td>
<td>25.1</td>
<td>29.3</td>
<td>30.2</td>
<td>10.4</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>CTL</td>
<td>28.3</td>
<td>26.0</td>
<td>32.8</td>
<td>36.3</td>
<td>14.0</td>
<td>9.1</td>
</tr>
</tbody>
</table>

- **Content Similarity (Content)**: based on similarity between authors’ publications
- **Collaborative Filtering (CF)**: based on existing collaborations
- **Hybrid**: a linear combination of the scores obtained by the Content and the CF methods.
- **Katz**: the best link predictor in link-prediction problem for social networks
- **Author Matching (Author)**: based on the random walk with restart on the collaboration graph
- **Topic Matching (Topic)**: combining the extracted topics into the random walking algorithm