Cross-domain Collaboration Recommendation

Jie Tang

Tsinghua University
Networked World

- **1.3 billion** users
- **700 billion** minutes/month

- **555 million** users
- **.5 billion** tweets/day

- **280 million** users
- **80% of users** are 80-90’s

- **560 million** users
- **influencing** our daily life

- **79 million** users per month
- **>10 billion** items/year

- **800 million** users
- **~50% revenue** from network life

- **500 million** users
- **57 billion** on 11/11
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

Collaborative Development

• It is impossible to work alone to create almost any piece of a software, in particular for a large software

• Collaborative software development model began widespread adoption with the Linux kernel in 1991
Cross-domain Collaboration (cont.)

- Increasing trend of cross-domain collaborations

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

Data Mining

1. Sparse Connection: <1%

2. Complementary expertise

3. Topic skewness: 9%

Theory

Automata theory
Graph theory
Complexity theory

Large graph
heterogeneous network
Social network
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

Author

Coauthorships

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

Medical Informatics

$G^T$

$v'_2$

$\ldots$

$v'_{N'}$

Cross-domain coauthorship

Query user

$r(t+1) = (1 - \tau)S \cdot r(t) + \tau q$

Sparse connection!
Topic Matching

Data Mining

Topics Extraction

Topics

Medical Informatics

Topics correlations

Complementary Expertise!

Topic skewness!
Topic Matching

\[ S_{z_i z'_j} = \frac{1}{\kappa} \sum_{(v,v') \in EST} [P(z_i|v) + P(z'_j|v')] \]

\[ r^{(t+1)} = (1 - \tau)S \cdot r^{(t)} + \tau q \]
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 $G^T$.

foreach collaborated document $d$ do

foreach word $x_{di} \in d$ do

Toss a coin $s_{di}$ according to $\text{bernoulli}(s_{di}) \sim \text{beta}(\gamma_t, \gamma)$, where $\text{beta}(.)$ is a Beta distribution, and $\gamma_t$ and $\gamma$ are two parameters:

if $s_{di} = 0$ 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 $z_{di} \sim \text{multi}({\theta_{vv'}})$ from the topic mixture $\theta_{vv'}$ specific to $(v, v')$;

end

if $s_{di} = 1$ then

Randomly select a user $v$;

Draw a topic $z_{di} \sim \text{multi}(\theta_v)$ from the topic model of user $v$;

end

Draw a word $x_{di} \sim \text{multi}(\phi_{z_{di}})$ from $z_{di}$-specific word distribution;

end

Step 2:

Intuitive explanation of Step 2 in CTL

Source domain

Topics

Target domain

Source domain

Topics

Target domain

Collaboration topics

$P(z_1 | s=0) < \epsilon$

$P(z_3' | s=0) < \epsilon$
Model Learning

- Model learning with Gibbs sampling. We sample $z$ and $s$ and then use the sampled $z$ and $s$ to infer the unknown distributions.

$$P(z_{di}|z_{-di}, x, \cdot) = \frac{n_{vz_{di}}^{-d_i} + \alpha}{\sum_z (n_{vz}^{-d_i} + \alpha)} \times \frac{m_{z_{di}x_{di}}^{-d_i} + \beta}{\sum_x (m_{z_{di}x}^{-d_i} + \beta)}$$
Model Learning (cont.)

• We sample $s$ to determine whether a word is generated by a collaboration or by oneself.

$$P(s_{di} = 0 | s_{-di}, z, t) = \frac{n_{ds_0} - di + \gamma t}{n_{ds_0} + n_{ds_1} + \gamma t + \gamma} \times \frac{n_{v^v z_{di}} + (n_{vz_{di}} + n_{v^v z_{di}}) + \alpha}{\sum_z (n_{v^v z} + (n_{vz_{di}} + n_{v^v z_{di}}) + \alpha)}$$
Model Learning (cont.)

- If \( s = 0 \), then we sample a pair of collaborators \((v, v')\) and construct a new topic distribution for the two collaborators, then sample the topic from the new distribution.

\[
P(z_{di} | s_{di} = 0, x, z_{-di}, \cdot) = \frac{m_{z_{di}x_{di}} + m_{z_{di}x_{di}} + m'_{z_{di}x_{di}} + \beta}{\sum_x (m_{z_{di}x} + m_{z_{di}x} + m'_{z_{di}x} + \beta)} 
\times \frac{n_{v'z_{di}} + (n_{vz_{di}} + n_{v'z_{di}}) + \alpha}{\sum_z (n_{v'z_{di}} + (n_{vz} + n_{v'z}) + \alpha)}
\]
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.
Performance on New Collaboration Prediction

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
From Peer Collaboration to Team Collaboration
Motivation

Task-Collaborator Assignment

Constraints:
1. A Task should be collaborated by k members
2. Work load balance
3. Authoritative Balance/Expertise Balance
   - at least one senior expert
4. Topic Coverage
5. Conflict-of-Interest(COI) avoidance
6. etc.

Challenge:
How to find optimal assignment under various constraints?

Constraint-based Optimization

• Objective
  – Maximize the relevance between experts and tasks
  – Satisfy the given constraints

• Definitions
  – $V(q_j)$: the set of experts who are able to do task $q_j$
  – $Q(v_i)$: the set of tasks assigned to expert $v_i$
  – $R_{ij}$: matching score between $q_j$ and $v_i$

• Basic Objective
  $$\text{Max} \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} R_{ij}$$
Various Constraints

1. Each task should be assigned to $m$ experts

$$\text{ST1: } \forall q_j \in Q, |V(q_j)| = m$$

2. Load Balance

$$\text{ST2 (strict): } \forall v_i \in V, n_1 \leq |Q(v_i)| \leq n_2$$

soft penalty: \[
\text{Min } \sum_{v_i \in V} |Q(v_i)|^2
\]

3. Authoritative balance

we divide all experts into $K$ levels, i.e., $V^1 \cup V^2 \cup \cdots \cup V^K = V$

strict: \[
|V^1 \cap V(q_j)| \geq 1
\]

soft: \[
\text{Min } \sum_{k=1}^{K} \sum_{j=1}^{N} |V^k \cap V(q_j)|^2
\]
Various Constraints (con’t)

4. **Topic Coverage**

\[
\text{Max} \sum_{z=1}^{T} \sum_{v_i \in V(q_j)} \mathbb{I}(\theta_{q_i z} > \tau_1) \mathbb{I}(\theta_{v_i z} > \tau_2)
\]

5. **COI avoidance**

Employ a binary $M \times N$ matrix $U$.

---

**Optimization Framework:**

Relevance & COI

\[
\text{Max} \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{i j} R_{i j}
\]

Load Balance

s.t.

\[
\forall q_j \in Q, |V(q_j)| = m
\]

\[
\forall v_i \in V, n_1 \leq |Q(v_i)| \leq n_2
\]

Authoritative Balance

\[
- \beta \sum_{v_i \in V} |Q(v_i)|^2
\]

\[
+ \lambda \sum_{q_j \in Q} \sum_{z=1}^{T} \sum_{v_i \in V(q_j)} \mathbb{I}(\theta_{q_j z} > \tau_1) \mathbb{I}(\theta_{v_i z} > \tau_2)
\]

Topic Coverage

\[
(\text{load balance}) \cdot (\text{authoritative balance}) \cdot (\text{topic coverage})
\]

Workflow

- **Modeling Multiple Topics**
  Associate each experts and queries with topic distribution

- **Generating Pairwise Matching Score**

- **Constraint-based Optimization framework**
  Combine various constraints

- **Optimization Solving**

**Still problems**
- How to define topic distributions?
- How to calculate the pairwise matching score $R_{ij}$?
- How to optimize the framework?
**Idea:**
Transform the problem to a convex cost network flow problem.

**min-cost max-flow || optimal matching**
Online Matching

- **User feedbacks**
  1. Pointing out a mistake match
  2. Specifying a new match

<table>
<thead>
<tr>
<th>Algorithm 2: Online matching algorithm.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> A minimum cost network flow $f$ on $G$ corresponding to the current assignment; an inappropriate match $(v_i, q_j)$ to be removed.</td>
</tr>
<tr>
<td><strong>Output:</strong> A new assignment.</td>
</tr>
<tr>
<td>2.1 $k = \text{expert level of } v_i;$</td>
</tr>
<tr>
<td>2.2 if $f(Q_{jk}, V_i) = 1$ then</td>
</tr>
<tr>
<td>2.3 Construct the residual network $G(f);$</td>
</tr>
<tr>
<td>2.4 Compute the shortest path $P_{\text{back}}$ from $T$ to $S$ on $G(f)$ which contains backward arc $(V_i, Q_{jk});$</td>
</tr>
<tr>
<td>2.5 Cancel(roll back) 1 unit of flow along $P_{\text{back}}$ and update $G(f);$</td>
</tr>
<tr>
<td>2.6 Remove arc $(Q_{jk}, V_i)$ from $G$ and update $G(f);$</td>
</tr>
<tr>
<td>2.7 Compute shortest augmenting path path $P_{\text{aug}}$ from $S$ to $T;$</td>
</tr>
<tr>
<td>2.8 Augment 1 unit of flow along $P_{\text{aug}};$</td>
</tr>
<tr>
<td>2.9 end</td>
</tr>
</tbody>
</table>
Experimental Setting

• **Paper-reviewer data set**
  - 338 papers (KDD’08, KDD’09, ICDM’09)
  - 354 reviewers (PC members of KDD’09)
  - COI matrix: coauthor relationship in the last five yrs.

• **Course-teacher data set**
  - 609 graduate courses (CMU, UIUC, Stanford, MIT)
  - Intuition: teachers’ graduate course often match his/her research interest.
Experiment Setting (con’t)

• **Evaluation measures**
  – **Matching Score (MS):**
    \[ MS = \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{ij} R_{ij} \]
  – **Load Variance (LV):**
    \[ LV = \sum_{i=1}^{M} \left( |Q(v_i)| - \frac{\sum_{i=1}^{M} |Q(v_i)|}{M} \right)^2 \]
  – **Expertise Variance (EV):**
    \[ EV = \sum_{j=1}^{N} \left( |V(q_j) \cap V^1| - \frac{\sum_{j=1}^{N} |V(q_j) \cap V^1|}{N} \right)^2 \]
  – **Precision** (in course-teacher assignment expr.)

• **Baseline:** Greedy Algorithm
Figure 2. Figure (a) and (b) illustrate how soft penalty function influences the matching score (MS) and load variance with different $\beta$ respectively. Figure (c) gives a comparison between soft penalty function and strict constraint methods towards load balance.

$\beta$ : weight of load balance  
$\mu=0$ : weight of authoritative balance
Paper-reviewer Experiment

Figure 3. Matching score (MS) and expertise variance (EV) with $\mu_1$ varied.

$\beta=0$ : weight of load balance
$\mu$ : weight of authoritative balance
## Paper-reviewer Case Study

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Assigned papers</th>
</tr>
</thead>
</table>
| Lise Getoor | - Evaluating Statistical Tests for Within-Network Classifiers of ...  
              - Discovering Organizational Structure in Dynamic Social Network  
              - Connections between the lines: augmenting social networks with text  
              - MetaFac: community discovery via relational hypergraph factorization  
              - Relational learning via latent social dimensions  
              - Influence and Correlation in Social Networks |
| Wei Fan    | - Mining Data Streams with Labeled and Unlabeled Training Examples  
              - Vague One-Class Learning for Data Streams  
              - Active Selection of Sensor Sites in Remote Sensing Applications  
              - Name-ethnicity classification from open sources  
              - Consensus group stable feature selection  
              - Categorizing and mining concept drifting data streams |
| Jie Tang   | - Co-evolution of social and affiliation networks  
              - Influence and Correlation in Social Networks  
              - Feedback Effects between Similarity and Social Influence ...  
              - Mobile call graphs: beyond power-law and lognormal distributions  
              - Audience selection for on-line brand advertising: privacy-friendly ... |

<table>
<thead>
<tr>
<th>Paper</th>
<th>Assigned reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience selection for on-line brand advertising: privacy-friendly social network targeting</td>
<td>C. Lee Giles, Jie Tang, Matthew Richardson, Hady Wirawan Lauw, Elena Zheleva</td>
</tr>
<tr>
<td>Partitioned Logistic Regression for Spam Filtering</td>
<td>Rong Jin, Chengxiang Zhai, Saharon Rosset, Masashi Sugiyama, Annalisa Appice</td>
</tr>
<tr>
<td>Structured Learning for Non-Smooth Ranking Losses</td>
<td>Xian-sheng Hua, Tie-yan Liu, Hang Li, Yunbo Cao, Lorenza Saitta</td>
</tr>
<tr>
<td>Unsupervised deduplication using cross-field dependencies</td>
<td>Chengxiang Zhai, Deepak Agarwal, Max Welling, Donald Metzler, Oren Kurland</td>
</tr>
<tr>
<td>The structure of information pathways in a social communication network</td>
<td>C. Lee Giles, Wolfgang Nejdl, Melanie Gnas, Michalis Faloutsos, Cameron Marlow</td>
</tr>
</tbody>
</table>
Course-Teacher Experiment

(a) Course assignment results

(b) Precision vs. \( \beta \) on UIUC data

Figure 5. Course-Teacher Assignment performance(%)
Online System

- http://review.arnetminer.org
Conclusion

• Study the problem of cross-domain collaboration recommendation and team collaboration
• Propose the cross-domain topic model for recommending collaborators
• Transformed the team collaboration problem as a optimization problem with convex-cost network flow problem
• 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!

Collaborators: Sen Wu (Stanford)
Jimeng Sun, Hang Su (Gatech)
Wenbin Tang (Face++), Chenhao Tan (Cornell)
Tao Lei (MIT), Bo Gao (THU)

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 the 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>Medical Info.(S) to Data Mining(T)</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>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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></td>
</tr>
<tr>
<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>
<td></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