Toward Robust Recommendation Systems for Scholarly Papers and Mobile Apps

Kazunari Sugiyama

National University of Singapore
Today’s talk

Outline of Singapore

Scholarly Paper Recommendation

Mobile App Recommendation

Popularity Prediction for Web 2.0 Items
Outline of Singapore

- **Population:** 5.39 million (in March 2014)
  - Chinese 74%, Malay 13%, Indian 9%, Others 4%
  

- **Area**
  - NUS: National University of Singapore
  - NTU: Nanyang Technological University
  - SMU: Singapore Management University
  - SUTD: Singapore University of Technology and Design

- **Language**
  Malay, Mandarin, English, Tamil

- **Area**
  716 km²

- **Area**
  94,321 km²

131.7 times larger than Singapore
Research Topics in Web IR and NLP Group (WING) by A/P Min-Yen Kan

Graduate Students

Jun Ping Ng: Temporal Relationship Identification

Aobo Wang: Informal Chinese Language Processing

Jovian Lin: Recommendation Systems for Mobile Applications

Tao Chen: Topics in Weibo

Xiangnan He: Topics in Web 2.0

Research Staffs

Kazunari Sugiyama: Recommender Systems in Digital Libraries

Dongyuan Lu: Social Media Mining

Muthu Kumar: NUS Co-author analysis

http://wing.comp.nus.edu.sg/
Scholarly Paper Recommendation

Kazunari Sugiyama and Min-Yen Kan:
- “Scholarly Paper Recommendation via User's Recent Research Interests” (JCDL’10)

- “Exploiting Potential Citation Papers in Scholarly Paper Recommendation” (JCDL’13, “Vannevar Bush Best Paper Award”)
Introduction

How many papers are published in 2012?

Introduction

How many papers are published in 2012?

Introduction

Recommendation System
(Especially, content-based system)

Researcher ➔ User profile $U$ ➔ Candidate papers to recommend $F_1, F_2, \ldots, F_P$

 Recommended papers
Introduction
Feature vector construction for candidate papers

Exploiting Potential Citation Papers in Scholarly Paper Recommendation

Kazunari Sugiyama
National University of Singapore
Computing 1, 13 Computing Drive,
Singapore 117417
sugiyama@comp.nus.edu.sg

Min-Yen Kan
NUS Interactive and Digital Media Institute
Computing 1, 13 Computing Drive,
Singapore 117417
kanny@comp.nus.edu.sg

ABSTRACT
To help generate relevant suggestions for researchers, recommendation systems have started to leverage the latest interests in the publication profiles of the researchers themselves. While using such a publication signature network has been shown to enhance performance, the network is often sparse, making recommendations inefficient. To alleviate this sparsity, we identify “potential citation papers” through the use of collaborative filtering. Also, at different levels of a paper, different sections have different significance, as a secondary consequence, we envisage which sections of papers can be leveraged to represent papers effectively.

On a scholarly paper recommendation dataset, we show that recommendation accuracy significantly outperforms state-of-the-art recommendation baselines as measured by uDCG and MBR, when we discover potential citation papers using improved variations of collaborative filtering and augment candidate papers using both the full text and assigning more weight to the conclusion sections.
Introduction

Feature vector construction for candidate papers

Citation and Reference Papers
Introduction

Feature vector construction for candidate papers

Citation and Reference Papers

Citation papers:
- Endorsement of the target paper
- Full text
- Fragments
  - "abstract,"
  - "introduction,"
  - "conclusion,"
  ...

Target paper

References
Introduction

Feature vector construction for candidate papers

Citation and Reference Papers

Citation papers:
- Endorsement of the target paper
- Full text
- Fragments
  - “abstract,” “introduction,” “conclusion,” …

Authors of citation papers
- May not cite relevant papers due to space limit
- Unaware of the relevant papers
Introduction

Feature vector construction for candidate papers

Citation and Reference Papers

Citation papers:
- Endorsement of the target paper
- Full text
- Fragments
  - "abstract,"
  - "introduction,"
  - "conclusion,"
  - ...

Authors of citation papers
- May not cite relevant papers due to space limit
- Unaware of the relevant papers

Potential citation papers
Our Goal

• To find potential citation papers to model candidate papers to recommend much better for better recommendation

• To refine the use of citation papers in characterizing candidate papers to recommend using fragments
Outline of Baseline System

[Sugiyama and Kan, JCDL’10]

1. Construct user profile from each researcher’s past papers
2. Compute similarity between user profile and \( P_{user} \) and \( F_{precj} \) (\( j = 1, \ldots, t \))
3. Recommend papers with high similarity
Forgetting factor

Weighting scheme
Cosine similarity
Outline of Baseline System

[Sugiyama and Kan, JCDL’10]

(1) Construct user profile from each researcher’s past papers

(2) Compute similarity between $P_{user}$ and $F^{p_{rec_j}} (j = 1, \ldots, t)$

(3) Recommend papers with high similarity

Candidate papers to recommend $F^{p_{rec_1}}$ to $F^{p_{rec_t}}$
[Sugiyama and Kan, JCDL’10]

Proposed Method [Sugiyama and Kan, JCDL’13]

Citation (cit) and potential citation (pc) papers

\[ W_{p_{cit1}} \rightarrow p \]

\[ W_{p_{cit2}} \rightarrow p \]

\[ W_{p_{citk}} \rightarrow p \]

\[ W_{p_{cit1}} \rightarrow p \]

\[ W_{p_{cit2}} \rightarrow p \]

\[ W_{p_{citk}} \rightarrow p \]

\[ W_{p_{pc1}} \rightarrow p \]

\[ W_{p_{pc2}} \rightarrow p \]

\[ W_{p_{pcj}} \rightarrow p \]

\[ p_{cit1} \]

\[ p_{cit2} \]

\[ p_{citk} \]

\[ p_{pc1} \]

\[ p_{pc2} \]

\[ p_{pcj} \]

\[ P_{ref1} \]

\[ P_{ref2} \]

\[ P_{refi} \]

\[ P_{ref1} \]

\[ P_{ref2} \]

\[ P_{refi} \]

\[ P_{cit1} \]

\[ P_{cit2} \]

\[ P_{citk} \]

\[ P_{pc1} \]

\[ P_{pc2} \]

\[ P_{pcj} \]

\[ : \text{Citation (cit) paper} \]

\[ : \text{Potential citation (pc) paper} \]
Proposed Method

(1) Leveraging Potential Citation Papers

(2) Leveraging Fragments in Potential Citation Papers
Proposed Method

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?
**Proposed Method**

**1. Leveraging Potential Citation Papers**

How are potential citation papers discovered?

<table>
<thead>
<tr>
<th>( p_{\text{cit}_1} )</th>
<th>( p_{\text{cit}_2} )</th>
<th>( p_{\text{cit}_3} )</th>
<th>( \ldots )</th>
<th>( p_{\text{cit}_j} )</th>
<th>( \ldots )</th>
<th>( p_{\text{cit}_{N-2}} )</th>
<th>( p_{\text{cit}_{N-1}} )</th>
<th>( p_{\text{cit}_N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>0.212</td>
<td></td>
<td>0.735</td>
<td></td>
<td>0.687</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.656</td>
<td></td>
<td>0.328</td>
<td></td>
<td>0.436</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_3 )</td>
<td>0.764</td>
<td></td>
<td></td>
<td>0.527</td>
<td>0.385</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{\text{tgt}} )</td>
<td>0.581</td>
<td></td>
<td></td>
<td></td>
<td>0.330</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-1} )</td>
<td>0.383</td>
<td></td>
<td>0.248</td>
<td></td>
<td>0.176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-2} )</td>
<td>0.654</td>
<td></td>
<td></td>
<td></td>
<td>0.525</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_N )</td>
<td>0.265</td>
<td></td>
<td></td>
<td></td>
<td>0.430</td>
<td></td>
<td></td>
<td>0.226</td>
</tr>
</tbody>
</table>

\( p_i \) \( (i = 1, 2, \ldots, N) \): All papers in dataset

\( p_{\text{cit}_j} \) \( (j = 1, 2, \ldots, N) \): Papers as citation papers in dataset
Proposed Method

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?

\[ p_{cit_1} \quad p_{cit_2} \quad p_{cit_3} \quad \ldots \quad p_{cit_j} \quad \ldots \quad p_{cit_{N-2}} \quad p_{cit_{N-1}} \quad p_{cit_N} \]

| \( p_1 \) | 0.212 | \( \ldots \) | 0.735 | \( \ldots \) | 0.687 |
| \( p_2 \) | 0.656 | \( \ldots \) | 0.328 | \( \ldots \) | 0.436 |
| \( p_3 \) | 0.764 | \( \ldots \) | \( \ldots \) | 0.527 | 0.385 |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( p_{tgt} \) | 0.581 | \( \ldots \) | \( \ldots \) | 0.330 | \( \ldots \) |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( p_{N-1} \) | 0.383 | \( \ldots \) | 0.248 | \( \ldots \) | 0.176 |
| \( p_{N-2} \) | 0.654 | \( \ldots \) | \( \ldots \) | 0.525 | \( \ldots \) |
| \( p_N \) | 0.265 | \( \ldots \) | \( \ldots \) | 0.430 | 0.226 |

\( p_i \) (\( i = 1, 2, \ldots, N \)) : All papers in dataset

\( p_{cit_j} \) (\( j = 1, 2, \ldots, N \)) : Papers as citation papers in dataset

sim: cosine similarity
**Proposed Method**

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?

<table>
<thead>
<tr>
<th>$p_{cit_1}$</th>
<th>$p_{cit_2}$</th>
<th>$p_{cit_3}$</th>
<th>(\ldots)</th>
<th>$p_{cit_j}$</th>
<th>(\ldots)</th>
<th>$p_{cit_{N-2}}$</th>
<th>$p_{cit_{N-1}}$</th>
<th>$p_{cit_N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0.212</td>
<td>0.735</td>
<td>(\ldots)</td>
<td>0.687</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.656</td>
<td>0.328</td>
<td>0.436</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.764</td>
<td>0.527</td>
<td>0.385</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_{tgt}$</td>
<td>0.581</td>
<td>(\ldots)</td>
<td>0.330</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_{N-1}$</td>
<td>0.383</td>
<td>0.248</td>
<td>0.176</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_{N-2}$</td>
<td>0.654</td>
<td>(\ldots)</td>
<td>0.525</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>$p_N$</td>
<td>0.265</td>
<td>(\ldots)</td>
<td>0.430</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>

**Pearson correlation**

- 0.538
- 0.216
- 0.475
- 0.304
- 0.513
- 0.487

$p_i (i = 1, 2, \ldots, N)$: All papers in dataset

$p_{cit_j} (j = 1, 2, \ldots, N)$: Papers as citation papers in dataset

Neighborhood of the target paper (e.g., set to 4)
Proposed Method

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?

<table>
<thead>
<tr>
<th>( p_{cit_1} )</th>
<th>( p_{cit_2} )</th>
<th>( p_{cit_3} )</th>
<th>( \ldots )</th>
<th>( p_{cit_j} )</th>
<th>( \ldots )</th>
<th>( p_{cit_{N-2}} )</th>
<th>( p_{cit_{N-1}} )</th>
<th>( p_{cit_N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>0.212</td>
<td>0.735</td>
<td>0.687</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.656</td>
<td>0.328</td>
<td>0.436</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_3 )</td>
<td>0.764</td>
<td>0.152</td>
<td>0.527</td>
<td>0.385</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{tgt} )</td>
<td>0.581</td>
<td></td>
<td></td>
<td></td>
<td>0.330</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-1} )</td>
<td>0.383</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.176</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-2} )</td>
<td>0.654</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.525</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_N )</td>
<td>0.265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.430</td>
<td>0.226</td>
<td></td>
</tr>
</tbody>
</table>

Pearson correlation

\[ \text{Pearson correlation} \]

\[ 0.538 \]

\[ 0.216 \]

\[ 0.475 \]

\[ --- \]

\[ 0.304 \]

\[ 0.513 \]

\[ 0.487 \]

\( p_i \) (\( i = 1, 2, \ldots, N \)): All papers in dataset

\( p_{cit_j} \) (\( j = 1, 2, \ldots, N \)): Papers as citation papers in dataset

Neighborhood of the target paper (e.g., set to 4)
### Proposed Method

(1) Leveraging Potential Citation Papers

**How are potential citation papers discovered?**

<table>
<thead>
<tr>
<th>( p_{cit_1} )</th>
<th>( p_{cit_2} )</th>
<th>( p_{cit_3} )</th>
<th>( \ldots )</th>
<th>( p_{cit_j} )</th>
<th>( \ldots )</th>
<th>( p_{cit_{N-2}} )</th>
<th>( p_{cit_{N-1}} )</th>
<th>( p_{cit_N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>0.212</td>
<td></td>
<td>0.735</td>
<td></td>
<td>0.687</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.656</td>
<td></td>
<td>0.328</td>
<td></td>
<td>0.436</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_3 )</td>
<td></td>
<td>0.764</td>
<td>0.152</td>
<td></td>
<td>0.527</td>
<td></td>
<td>0.385</td>
<td></td>
</tr>
<tr>
<td>( p_{tgt} )</td>
<td>0.435</td>
<td>0.581</td>
<td>0.536</td>
<td>( \ldots )</td>
<td>0.211</td>
<td>( \ldots )</td>
<td>0.330</td>
<td>0.472</td>
</tr>
<tr>
<td>( p_{N-1} )</td>
<td></td>
<td></td>
<td>0.248</td>
<td></td>
<td>0.176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-2} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.654</td>
<td></td>
<td>0.525</td>
<td></td>
</tr>
<tr>
<td>( p_N )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.430</td>
<td>0.226</td>
</tr>
</tbody>
</table>

---

**Pearson correlation**

\[ 0.538 \]
[0.216]

| \( p_i \) (\( i = 1, 2, \ldots, N \)) : All papers in dataset |
| \( p_{cit_j} \) (\( j = 1, 2, \ldots, N \)) : Papers as citation papers in dataset |

---

(‘potential citation papers’ (e.g., set to 3)
Proposed Method

Identified Potential Citation Papers
Proposed Method

(1) Leveraging Potential Citation Papers

How is the sparsity of matrix solved?
Proposed Method

(1) Leveraging Potential Citation Papers

How is the sparsity of matrix solved?

**Original matrix**

<table>
<thead>
<tr>
<th>$p_{cit_1}$</th>
<th>$p_{cit_2}$</th>
<th>$p_{cit_3}$</th>
<th>$p_{cit_4}$</th>
<th>$p_{cit_5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0.233</td>
<td></td>
<td></td>
<td>0.628</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.233</td>
<td>0.147</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.147</td>
<td>0.265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td></td>
<td>0.265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.628</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Imputed matrix**

<table>
<thead>
<tr>
<th>$p_{cit_1}$</th>
<th>$p_{cit_2}$</th>
<th>$p_{cit_3}$</th>
<th>$p_{cit_4}$</th>
<th>$p_{cit_5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>1.000</td>
<td>0.233</td>
<td>0.723</td>
<td>0.538</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.233</td>
<td>1.000</td>
<td>0.147</td>
<td>0.476</td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.723</td>
<td>0.147</td>
<td>1.000</td>
<td>0.265</td>
</tr>
<tr>
<td>$p_4$</td>
<td>0.538</td>
<td>0.476</td>
<td>0.265</td>
<td>1.000</td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.628</td>
<td>0.156</td>
<td>0.521</td>
<td>0.268</td>
</tr>
</tbody>
</table>

The values in the cell:
Cosine similarity between papers

Imputation
Proposed Method

(1) Leveraging Potential Citation Papers

How is the sparsity of matrix solved?

Target paper ($p_1$) and corresponding
Imputed similarities of neighborhood
($p_2$, $p_4$, and $p_5$) from “Imputed matrix”
Proposed Method

(1) Leveraging Potential Citation Papers

Feature Vector Construction for Target Papers

\[
\mathbf{F}^P = \mathbf{f}^P + \sum_{x=1}^{j} \mathbf{W}^P_{p_{c_x}} \rightarrow p \mathbf{f}^P_{p_{c_x}} + \sum_{y=1}^{k} \mathbf{W}^P_{p_{c_y}} \rightarrow p \mathbf{f}^P_{p_{c_y}} + \sum_{z=1}^{l} \mathbf{W}^P_{p} \rightarrow p_{r_f} \mathbf{f}^P_{p_{r_f}}
\]
Proposed Method

(1) Leveraging Potential Citation Papers

Feature Vector Construction for Target Papers

\[ F^p = f^p + \sum_{x=1}^{j} W^{p_{pc_x} \rightarrow p} f^{p_{pc_x}} \]
\[ + \sum_{y=1}^{k} W^{p_{city} \rightarrow p} f^{p_{city}} \]
\[ + \sum_{z=1}^{l} W^{p \rightarrow p_{ref_z}} f^{p_{ref_z}} \]

\[ \text{cosine similarity} \]
Proposed Method

(2) Leveraging Fragments in Potential Citation Papers

- \([\textit{frg}}-	ext{SIM}]:\) Fragments with cosine similarity weighting

- \([\textit{frg}}-	ext{TW}]:\) \([\textit{frg}}-	ext{SIM}]\) with tunable weight
Proposed Method

(2) Leveraging Fragments in Potential Citation Papers

\[ \text{[frg-SIM]: Fragments with cosine similarity weighting} \]

\[
F_P = \sum_{x=1}^{j} W_{p_{pcx} \rightarrow p} f_{p_{pcx}} + \sum_{y=1}^{k} W_{p_{city} \rightarrow p} f_{p_{city}} + f_P + \sum_{x=1}^{j} W_{p_{pcx} \rightarrow p} f_{p_{pcx}} + \sum_{y=1}^{k} W_{p_{city} \rightarrow p} f_{p_{city}} + \sum_{z=1}^{l} W_{p \rightarrow p_{refz}} f_{p_{refz}}
\]

Fragments
(“abstract,” “introduction,” “conclusion,” etc.)

Full text
Proposed Method

(2) Leveraging Fragments in Potential Citation Papers

\[ [frg-TW]: [frg-SIM] \text{ with tunable weight} \]

Fragments

("abstract," "introduction," "conclusion," etc.)

\[
F^p = \alpha \left( \sum_{x=1}^{i} W^{p_{pc_x} \rightarrow p} f^{p_{pc_x}} + \sum_{y=1}^{k} W^{p_{city} \rightarrow p} f^{p_{city}} \right) p_{cit_1} \\
+ (1 - \alpha) \left( f^p + \sum_{x=1}^{i} W^{p_{pc_x} \rightarrow p} f^{p_{pc_x}} \\
+ \sum_{y=1}^{k} W^{p_{city} \rightarrow p} f^{p_{city}} \\
+ \sum_{z=1}^{l} W^{p \rightarrow p_{ref_z}} f^{p_{ref_z}} \right)
\]

References

\[
W^{p_{cit_1} \rightarrow p} p_{cit_1} \\
W^{p_{pc_1} \rightarrow p} p_{pc_1} \\
W^{p_{pc_2} \rightarrow p} p_{pc_2} \\
W^{p_{cit_2} \rightarrow p} p_{cit_2} \\
W^{p_{cit_k} \rightarrow p} p_{cit_k} \\
W^{p \rightarrow p_{ref_1}} p_{ref_1} \\
W^{p \rightarrow p_{ref_2}} p_{ref_2} \\
\]

Full text
Experiments

Experimental Data

(Basic dataset has been released from http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html)

(a) Researchers (they have publication lists in DBLP)

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of researchers</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Average number of DBLP papers</td>
<td>10.4</td>
<td>9.6</td>
</tr>
<tr>
<td>Average number of relevant papers in our dataset</td>
<td>76.3</td>
<td>74.5</td>
</tr>
<tr>
<td>Average number of citations</td>
<td>15.3 (max. 169)</td>
<td>14.4 (max. 145)</td>
</tr>
<tr>
<td>Average number of references</td>
<td>15.8 (max. 47)</td>
<td>14.2 (max. 58)</td>
</tr>
</tbody>
</table>

(b) Candidate papers to recommend (constructed from ACM Digital Library)

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of papers</td>
<td>50,176</td>
<td>50,175</td>
</tr>
<tr>
<td>Average number of citations</td>
<td>19.4 (max. 175)</td>
<td>16.5 (max. 158)</td>
</tr>
<tr>
<td>Average number of references</td>
<td>15.7 (max. 45)</td>
<td>15.4 (max. 53)</td>
</tr>
</tbody>
</table>
Experiments

Evaluation Measure

• **NDCG@5, 10 [Järvelin and Kekäläinen, SIGIR’00]**
  - Gives more weight to highly ranked items
  - Incorporates different relevance levels through different gain values
    - 1: Relevant search results
    - 0: Irrelevant search results

• **MRR [Voorhees, TREC-8, ’99]**
  - Provides insight in the ability to return a relevant item at the top of the ranking
Experiments

Experimental Results

(1) Leveraging potential citation papers*
   [Tune: \textit{pc}] Parameter tuning to discover potential citation papers

(2) Leveraging fragments in potential citation papers*
   [Tune: \textit{frg-SIM}] Fragments with cosine similarity weighting
   [Tune: \textit{frg-TW}] [\textit{frg-SIM}] with tunable weight

(3) Applying optimized parameters to test set

* Please refer to the following paper about the detailed optimization process:
K. Sugiyama and M.-Y. Kan: “Exploiting Potential Citation Papers in Scholarly Paper Recommendation” (JCDL’13)
Baseline
[Nascimento et al., JCDL’11]
Baseline

[Wang and Blei., KDD’11]: Collaborative topic regression
Combines ideas from collaborative filtering and content analysis based on probabilistic topic modeling

\[
\begin{array}{cccccc}
 & p_1 & p_2 & \cdots & p_j & \cdots & p_N \\
\hline
u_1 & & & & 0 & & \\
\hline
u_2 & 1 & 1 & 1 & & \\
\vdots & & & & & \\
\hline
u_i & 1 & 1 & r_{ij} & 0 & \\
\vdots & & & & & \\
\hline
u_U & 0 & 1 & & & \\
\hline
\end{array}
\]

\[r_{ij} \in \{0,1\}\] whether user \(u_i\) includes paper \(p_j\) in the user’s preference

Title and abstract

\[
\begin{array}{ccc}
\alpha & \theta & \beta \\
\lambda_v & v & \lambda_u \\
r & z & w \\
u & l & \end{array}
\]
### (3) Applying Optimized Parameters to Test Set

<table>
<thead>
<tr>
<th>Method</th>
<th>nDCG@5</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pc-IMP</strong> ( n=4, Npc=6 )</td>
<td>0.572</td>
<td>0.787</td>
</tr>
<tr>
<td>( frg)-SIM (Full text + Conclusion)</td>
<td>0.579</td>
<td>0.793</td>
</tr>
<tr>
<td>( frg)-TW ( \alpha=0.4, \text{Full text + Conclusion} )</td>
<td>0.579</td>
<td>0.793</td>
</tr>
<tr>
<td><strong>Baseline system</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Sugiyama and Kan, JCDL’10] ( \text{Weight “SIM,” Th=0.4, } \gamma=0.23, d=3 )</td>
<td>0.525</td>
<td>0.751</td>
</tr>
<tr>
<td>[Nascimento et al., JCDL’11] ( \text{“Frequency of bi-gram” obtained from title and abstract} )</td>
<td>0.336</td>
<td>0.438</td>
</tr>
<tr>
<td>[Wang and Blei., KDD’11] ( \text{“In-matrix prediction” in collaborative topic regression} )</td>
<td>0.393</td>
<td>0.495</td>
</tr>
</tbody>
</table>
Microscopic Analysis

• 1st Relevant Result in Recommendation List for a “Mobile Computing” Researcher
  [Sugiyama and Kan, JCDL’10]: 52nd
  [Sugiyama and Kan, JCDL’13]: 1st

• Example of Identified Potential Citation Papers

  “Biomechanics”
  “Computer-based music conducting systems”
  “Machine learning”
  “Human Computer Interaction” papers

Target Paper: “Real world Gesture analysis”
Limitations

“Understanding mobile user’s behavior”
- Mobile technology
- User search behavior
- Clustering

Interdisciplinary paper

Identified Potential Citation Papers

Target paper: “Understanding mobile user’s behavior”
Mobile App Recommendation

Jovian Lin, Kazunari Sugiyama, Min-Yen Kan and Tat-Seng Chua:
- “Addressing Cold-Start in App Recommendation: Latent User Models Constructed from Twitter Followers”
  (SIGIR ’13)

- “New and Improved: Modeling Versions to Improve App Recommendation”
  (SIGIR ’14, to appear)
INFORMATION OVERLOAD
Two Important Observations in Apps

1. Apps contain references to their Twitter accounts.
2. Early signals about apps can be present in social networks, even before ratings are received.

By May 2012, Evernote’s Twitter account already had 120,000 followers and 1,300 tweets.
Estimate the probability that “a target user $u$ will like an app $a$.”

$$p( + | a, u ) = \sum_{t \in T(a)} p( + | t, u ) \cdot p( t | a)$$

- **“like” app user**
  - Probability that the presence of Twitter-follower $t$ indicates that it is “liked” by user $u$.
  - Derived from Pseudo-Documents and Pseudo-Words.

- **Uniform distribution over the various Twitter-followers ($t$) following app $a$.**
Pseudo-Document and Pseudo-Words

User u

- Disliked
  - App a
    - Followed by:
      - twitterID\textsubscript{10}
      - twitterID\textsubscript{12}

- Liked
  - App b
    - Followed by:
      - twitterID\textsubscript{10}
      - twitterID\textsubscript{12}
      - twitterID\textsubscript{29}

- Liked
  - App c
    - Followed by:
      - twitterID\textsubscript{29}
      - twitterID\textsubscript{31}

Twitter-follower ID

- (twitterID\textsubscript{10}, DISLIKED)
- (twitterID\textsubscript{12}, DISLIKED)
- (twitterID\textsubscript{10}, LIKED)
- (twitterID\textsubscript{12}, LIKED)
- (twitterID\textsubscript{29}, LIKED)
- (twitterID\textsubscript{31}, LIKED)

Pseudo-document u
Constructing Latent Groups

Pseudo-documents

\[ p( + | t, u) = \sum_{z \in Z} p( +, t | z) p( z | u) \]

Probability that the presence of Twitter-follower \( t \) indicates that it is “liked” by user \( u \).
Dataset

We collected data from the Apple iTunes Store and Twitter during September to December 2012.

Statistics:
1,289,668 ratings
7,116 apps (with Twitter accounts)
10,133 users.

Restrictions:
Each user must give at least 10 ratings for apps.
Each Twitter ID is related to at least 5 apps.
RQ1: How does the performance of Twitter-followers feature compare with other features?

![Graph showing the performance comparison of different features]

- **All** = All features
- **T** = Twitter-followers
- **G** = Genres
- **D** = Developers
- **W** = Words
RQ2: How does our method compare with other techniques?
RQ3: Do the latent groups make any sense? What can we learn from them?

<table>
<thead>
<tr>
<th>Latent Group 1</th>
<th>Latent Group 2</th>
<th>Latent Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top genres:</strong></td>
<td><strong>Top genres:</strong></td>
<td><strong>Top genres:</strong></td>
</tr>
<tr>
<td>Books (45%)</td>
<td>Music (92%)</td>
<td>Games (77%)</td>
</tr>
<tr>
<td>Education (33%)</td>
<td>AmpRit+ (Music)</td>
<td>Photo &amp; Video (12%)</td>
</tr>
<tr>
<td>Games (13%)</td>
<td>GuitarStudio (Music)</td>
<td>Example apps:</td>
</tr>
<tr>
<td>Example apps:</td>
<td>Everyday Looper (Music)</td>
<td>- Another World 20th Anniversary (Games)</td>
</tr>
<tr>
<td>The Cat in the Hat (Books)</td>
<td>Mixr DJ (Music)</td>
<td>- Paper Monsters (Games)</td>
</tr>
<tr>
<td>Christmas Cutie (Books)</td>
<td>KORG iELECTRIBE (Music)</td>
<td>- Stickman Cliff Diving (Games)</td>
</tr>
<tr>
<td>Happy Earth Day, Dear Planet (Books)</td>
<td>KORG iPolysix (Music)</td>
<td>- Lili (Games)</td>
</tr>
<tr>
<td>Friendly Shapes (Education)</td>
<td>Pro Metronome (Music)</td>
<td>- Snoopy’s Street Fair (Games)</td>
</tr>
<tr>
<td>There’s No Place Like Space! (Education)</td>
<td>Chord Detector (Music)</td>
<td>- Gizmonauts (Games)</td>
</tr>
<tr>
<td>Pasta Crazy Chef (Games)</td>
<td></td>
<td>- InstaBooth+ (Photo &amp; Video)</td>
</tr>
<tr>
<td>Gingerbread Dress (Games)</td>
<td></td>
<td>- ArtStudio for iPad (Photo &amp; Video)</td>
</tr>
</tbody>
</table>

Nosy Crow Apps (twitter.com/nosycrowapps)
Nosy Crow creates children’s books and apps. You may know our 3-D Fairytale apps, The Three Little Pigs & Cinderella.

The iMums (twitter.com/TheiMums)
Four mums dedicated to reviewing apps and technology products for children to help educate their parents about the variety available. Loads of giveaways too!

Mums with Apps (twitter.com/momswithapps)
Supporting family-friendly developers seeking to promote quality apps for kids and families.

Charly James (twitter.com/CharlyJames2)
Div. Mom of 2 w/varying SN & medical d/x. dandelion moms; A4 Free Apps @CharlyJames4; Ellie’s Games; Fernandez Design.

Next is Great (twitter.com/nextisgreat)
We create and develop brain teasing educational iOS apps for kids and teenagers. Check out Pirate Trio Academy and Geek Kids.

Derek Jones (twitter.com/MusicInclusive)
Indie music publishing label, studio & brand. Blues&Rock, Progressive&Funk, Jazz&Fusion, Alternate&Christian, Classical, Education & a lot in-between too!

Chip Boaz (twitter.com/iosmusicandyou)
I'm a musician based in the San Francisco Bay Area with an interest in using my iPad, iPhone, & iPod to make music. Follow my iOS adventures @ iOS Music And You

Dave Gibson (twitter.com/MicroTrackdB)

Ashley Elsdon (twitter.com/IamAshleyElsdon)
Everything from Mobile Music Creation, geekery, tech, art and Doctor Who! http://www.ashleyelsdon.com

Andrew Wardell (twitter.com/andrewwardell)
Nostalgic futurist, amateur photographer, sax-playing synthesist, musical mountain-biking metacolic. More than just a bag of salty water...

Sarah Thomson (twitter.com/SarahLuvsVGames)
Video games warrior, lover of life, eternal student of the universe, drinker of Kombucha, Baroness of PlayStation Mobile.

JasonLeeNester (twitter.com/JasonNester)
I am a Multimedia developer working at Kent State University! I also do art services for the game industry as well as run a small indie game company, True Media.

Agalag iOS Games (twitter.com/AgalagGames)
Agalag Games is an independent iOS game studio. Our aim is to create fun innovative and casual iPhone games which we really like and want to play. Publisher.

Samadhi Games (twitter.com/SamadhiGames)
Hi! Samadhi Games LLC is an Indie Developer of iOS, Android and Desktop Apps. Arizona - http://www.samadhigames.com

Finger Arts: App Dev (twitter.com/fingerartsgames)
We develop cool & innovative iPhone, iPod Touch & iPad Games. Rocking the charts in iTunes: Sudoku 2, Hangman RSS, 4 in a Row & now Solitaire :)
Version Sensitive Recommendation

Relationship between version of apps and users

Legend
- Red: A version of the app.
- Green: An ID of a topic.

App X

------- | ------- | ------- | ------- | ------------
1.0     | 1.1     | 1.2     | 2.0     | 3.0
1       | 1       | 1       | 2       | 5
2       |         | 3       | 3       |      
4       |         |         | 4       |      

Users
- Alex
- Bob
- Clark

Topics
- 1
- 2
- 3
- 4
- 5
Example of Changelog and Genres in App

Changelog

Version 2.0 (major update)
A total rewrite including:
• A beautifully simple new interface for managing multiple blogs.
• Improvements to Dashboard browsing.
• Improvements to posting, including landscape editing.
• Read and reply to Messages.
• Find followers via your address book.
• Sign up on your iPhone.

Version 1.2 (minor update)
• Native reblogging.
• German localization.
• Photo from URL and click-through URL support on photo posts.
• 'Send to Twitter' switch in advanced options now respects your per-blog settings.
• iOS 4.0 compatibility fixes.
• Bug fixes and optimizations.

Version 1.1 (minor update)
• Post geotagging.
• Built-in web browser.
• Fixed a bug where Photo posts can cause crashes.
• Fixed memory leaks.
• .........

• From Tumblr app

Genres in App


• From Apple’s iOS app store (as of January 2014)
Our Approach to Version Sensitive Recommendation for Mobile Apps

1. Generate latent topics from version features using semi-supervised topics models (labeled LDA, LLDA) to characterize each version.

2. Discriminate the topics based on genre metadata and identify important topics based on a customized popularity score.

3. Compute a personalized score with respect to an app and its version to recommend relevant mobile apps to each user.
Experiments

Experimental Data

Crawled from iTunes App Store and App Annie

• App metadata (6,524 apps)
• Version information (109,338 versions)
• Ratings (1,000,809 ratings)
• Users (9,797 users)

Evaluation Measure

• Recall
Description
Movies by Flixster. The #1 app for movie reviews, trailers, and showtimes.
- The most downloaded movies app of all time
- Featured in the App Store Essentials Hall of Fame by Apple
- Best showtimes app for iPhone - Lifehacker.com
1. Browse the top box office movies and movies opening soon
2. Look up showtimes at your favorite theater and buy tickets (from participating theaters)
3. Get critic reviews from Rotten Tomatoes
4. Watch high quality trailers
5. Stream and download full-length movies.
6. Create your own "Want to See" list, rate & review movies
7. View and manage your Netflix queue

What's New
Version 8.2 (May 9, 2013)
1) We've totally revamped the DVD tab so that you can now use filters to discover movies you want to watch at home.
2) The DVD tab has a new addition - Netflix Streaming. Browse through Netflix's streaming catalog and add them to your queue.
3) You now have the option of adding a movie to your Netflix queue or launching the Netflix app to stream it.
4) Fix for a license expired bug when trying to play downloaded movies.
5) You can now see all critic reviews for a movie on iPad
6) Fixed errors while redeeming gift movies
7) Performance improvements all around

Average Ratings (United States)
- Current Version - 3.5 - 581 ratings
- All Versions - 3.5 - 470,649 ratings

Version Information
- 6.2 May 9, 2013 Current release
- 6.1 Apr 22, 2013
- 6.0 Mar 22, 2013

Rating Information
- 5 stars
- 4 stars
- 3 stars
- 2 stars
- 1 star

Version: 8.1 (Apr 22, 2013)
1. Swipe to the right/left on a movie (in Box Office or DVD screen) to access quick controls - "Want to See" to add the movie to your want to see list, "Add Rating" to add your rating, "Trailer" & "Showtimes" to launch trailer & showtimes for the movie.
2. Movies in your want to see list are now grouped into 3 sections - upcoming movies, movies currently playing in theaters and all other movies available on dvd/streaming.
3. Improvements to movie pages on iPad
4. Sort movies in your UltraViolet collection
5. Switching between tabs is now faster
Experimental Results

Recall of version sensitive recommendation (VSR) against other individual recommendation techniques

Recall of various combinations of recommendation techniques
Experimental Results

Three most important topics

<table>
<thead>
<tr>
<th>Top Terms</th>
<th>Latent Topic #385</th>
<th>Latent Topic #47</th>
<th>Observed Topic “Medical Genre”</th>
</tr>
</thead>
<tbody>
<tr>
<td>retina: 0.065947</td>
<td>map: 0.052109</td>
<td>#medical_genre: 0.072711</td>
<td></td>
</tr>
<tr>
<td>display: 0.048744</td>
<td>#navigation_genre: 0.043457</td>
<td>#health &amp; fitness_genre: 0.056700</td>
<td></td>
</tr>
<tr>
<td>support: 0.046697</td>
<td>traffic: 0.030958</td>
<td>pain: 0.032384</td>
<td></td>
</tr>
<tr>
<td>graphic: 0.0341819</td>
<td>rout: 0.023651</td>
<td>report: 0.026726</td>
<td></td>
</tr>
<tr>
<td>resolut: 0.029084</td>
<td>improv: 0.023267</td>
<td>medic: 0.026605</td>
<td></td>
</tr>
<tr>
<td>full: 0.026627</td>
<td>locat: 0.023011</td>
<td>pregnancy: 0.022512</td>
<td></td>
</tr>
<tr>
<td>#minor_update: 0.024579</td>
<td>tri: 0.019485</td>
<td>graph: 0.018781</td>
<td></td>
</tr>
<tr>
<td>touch: 0.024579</td>
<td>#travel_genre: 0.019421</td>
<td>period: 0.015290</td>
<td></td>
</tr>
<tr>
<td>fix: 0.020074</td>
<td>road: 0.016857</td>
<td>health: 0.015169</td>
<td></td>
</tr>
<tr>
<td>ipad: 0.020074</td>
<td>crash: 0.014935</td>
<td>inform: 0.014326</td>
<td></td>
</tr>
<tr>
<td>#utilities_genre: 0.019066</td>
<td>address: 0.014294</td>
<td>track: 0.012521</td>
<td></td>
</tr>
<tr>
<td>#productivity_genre: 0.018435</td>
<td>poi: 0.013204</td>
<td>diary: 0.012400</td>
<td></td>
</tr>
<tr>
<td>high: 0.017207</td>
<td>fuel: 0.011936</td>
<td>chart: 0.011798</td>
<td></td>
</tr>
<tr>
<td>optm: 0.016387</td>
<td>auto: 0.011794</td>
<td>drug: 0.011076</td>
<td></td>
</tr>
<tr>
<td>auto: 0.015159</td>
<td>alert: 0.011602</td>
<td>calcul: 0.010595</td>
<td></td>
</tr>
</tbody>
</table>

#xxx: version category
#xxx: genre category
Popularity Prediction for Web 2.0 Items

Xiangnan He, Ming Gao, Min-Yen Kan, Yiqun Liu, and Kazunari Sugiyama:
-“Predicting the Popularity of Web 2.0 Items Based on User Comments”
(SIGIR ’14, to appear)
What is the Characteristics of Web 2.0 Items?

- Weblog
- Twitter
- YouTube
- Flicker

Users can join by posting contents, commenting, voting, ...

Important to take these signals into account to rank Web 2.0 items
Problem of Current Search Engine

Query: “The Voice of China” (on July 24th, 2013)

- Chinese reality talent show
  - Started in 2012
    [The top 3 results in YouTube.com domain in Google]

  - Less than 10,000 views
    - Ranked 16th, but more than 100,000 views

- Started 2nd season on July 2013
Proposed Approach

Bipartite User-Item Ranking (BUIR)

Regularization:
• Map each vertex in the graph to a real number so that the value reflects
  - the vertex’s popularity (for items)
  or
  - influence (for users)

Incorporate the following three hypotheses into our model
• Temporal factor
• Social Influence factor
• Current popularity factor
Experiments
Evaluate predicted ranking in 3 days after a target day with ground truth ranking

Experimental Data
• YouTube (21,653 videos)
• Flicker (26,815 images)
• Last.fm (16,824 artists)
(10%: Parameter tuning in regularization, 90%: Test)

Evaluation Measure
• Spearman’s rank correlation coefficient
• Normalized discounted cumulative gain (nDCG)
Experimental Results

Spearman coefficient (%) of overall evaluation

<table>
<thead>
<tr>
<th></th>
<th>YouTube</th>
<th>Flickr</th>
<th>Last.fm</th>
</tr>
</thead>
<tbody>
<tr>
<td>View Count</td>
<td>73.39</td>
<td>58.42</td>
<td>67.31</td>
</tr>
<tr>
<td>Comment count in the past</td>
<td>83.35</td>
<td>59.43</td>
<td>67.21</td>
</tr>
<tr>
<td>Comment count in the future</td>
<td>84.53</td>
<td>59.41</td>
<td>67.20</td>
</tr>
<tr>
<td>Multivariate linear Model</td>
<td>78.24</td>
<td>58.00</td>
<td>38.09</td>
</tr>
<tr>
<td>PageRank</td>
<td>80.72</td>
<td>28.15</td>
<td>10.24</td>
</tr>
<tr>
<td>BUIR</td>
<td>87.72</td>
<td>64.60</td>
<td>70.43</td>
</tr>
</tbody>
</table>

Improvement in Spearman coefficient between BUIR and the best baselines of query-specific evaluation
Summary of Today’s Talk

- Scholarly Paper Recommendation
  - Identify “potential citation papers”
  - Leverage fragments of the paper

- Mobile App Recommendation
  - Apply LDA by taking Twitter followers and version updates into account

- Popularity Prediction for Web 2.0 Items
  - Apply regularization based on user’s comments

Thank you very much!