DL, IR, and NLP Research at Web Information Retrieval and Natural Language Processing Group (WING) in National University of Singapore

Kazunari Sugiyama

National University of Singapore
Today’s talk
- Outline of Singapore
- Scholarly Paper Recommendation
- Mobile App Recommendation
- Multi-Document Abstractive Summarization
- Recent Work on Health Informatics
Outline of Singapore

- **Population:** 5.61 million (in June 2016)
  - Chinese 76.1%, Malay 15.0%, Indian 7.4%, Others 1.5%
  

- **Area**
  
  ![Map of Singapore](image)

  - **716 km²**

- **Language**

  Malay, Mandarin, English, Tamil

- **Tokyo 23 wards**

  ![Map of Tokyo 23 wards](image)

  - **622 km²**
Outline of Singapore
Six Universities in Singapore

NUS: National University of Singapore
NTU: Nanyang Technological University
SMU: Singapore Management University
SUTD: Singapore University of Technology and Design
SIT: Singapore Institute of Technology
SUSS: Singapore University of Social Sciences
Research Topics in Web IR and NLP Group (WING) by A/P Min-Yen Kan

**Students**
- Muthu Kumar: MOOC
- Kishaloy Halder: Health Forum Recommendation
- Wenqiang Lei: Discourse Parsing
- Animesh Prasad: Information Extraction
- Yuanxin Xiang: Verb Duration Determination
- Sumei Su: Serendipitous Mobile App Recommendation
- You Zhou: User Behavior Analysis across Multiple Social Networks
- Yichi Zhang: MOOC
- Van Hoang Nguyen: Credibility Analysis in Health Communities

**Research Staffs**
- Kazunari Sugiyama: Recommender Systems in Digital Libraries
- Shenhao Jiang: Topic evolution in scholarly papers
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)
Introduction

Feature vector construction for candidate papers

Fragments

Full text

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
kannya@comp.nus.edu.sg

ABSTRACT
To help generate relevant suggestions for researchers, recommendation systems have started to leverage the latent interests in the publication profiles of the researchers themselves. While using such a publication-based network has been shown to enhance performance, the network is often sparse, making recommendation difficult. To alleviate this sparsity, we identify potential citation papers through the use of collaborative filtering. Also, as different logical sections of a paper have different significance, as a secondary contribution, we invent a new scheme in which sections of papers can be rearranged to represent papers effectively.

On a scholarly paper recommendation dataset, we show that recommendation accuracy significantly improves using our technique, compared to previous recommendation baselines as measured by ADCC and MRR. When we discover potential citation papers using support estimation via collaborative filtering and reorganize candidate papers using both the full text and assigning more weight to the conclusion sections.

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
kannya@comp.nus.edu.sg

ABSTRACT
To help generate relevant suggestions for researchers, recommendation systems have started to leverage the latent interests in the publication profiles of the researchers themselves. While using such a publication-based network has been shown to enhance performance, the network is often sparse, making recommendation difficult. To alleviate this sparsity, we identify potential citation papers through the use of collaborative filtering. Also, as different logical sections of a paper have different significance, as a secondary contribution, we invent a new scheme in which sections of papers can be rearranged to represent papers effectively.

On a scholarly paper recommendation dataset, we show that recommendation accuracy significantly improves using our technique, compared to previous recommendation baselines as measured by ADCC and MRR. When we discover potential citation papers using support estimation via collaborative filtering and reorganize 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,” …
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 $P_{user}$ and $F^{pre_j}_{rec}$ $(j = 1, \ldots, t)$

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

\[ \text{P}_{\text{user}} \]

(2) Compute similarity between \( \text{P}_{\text{user}} \) and \( F^{\text{prec}_j} \) \((j = 1, \ldots, t)\)

Candidate papers to recommend

\[ F^{\text{prec}_1} \text{ to } F^{\text{prec}_t} \]

(3) Recommend papers with high similarity
[Sugiyama and Kan, JCDL’10]

Citation (cit) papers

\( p_{cit_1} \) --- \( p_{cit_2} \) --- \( p_{cit_k} \)

\( W^{p_{cit_1} \rightarrow p} \)
\( W^{p_{cit_2} \rightarrow p} \)
\( W^{p_{cit_k} \rightarrow p} \)

\( p \)

References

\( p_{ref_1} \) --- \( p_{ref_2} \) --- \( p_{ref_i} \)

\( W^{p \rightarrow p_{ref_1}} \)
\( W^{p \rightarrow p_{ref_2}} \)
\( W^{p \rightarrow p_{ref_i}} \)

Reference (ref) papers

Proposed Method [Sugiyama and Kan, JCDL’13]

Citation (cit) and potential citation (pc) papers

\( p_{pc_2} \)
\( p_{cit_2} \)
\( p_{pc_j} \)

\( W^{p_{pc_2} \rightarrow p} \)
\( W^{p_{cit_2} \rightarrow p} \)
\( W^{p_{pc_j} \rightarrow p} \)

\( p_{cit_1} \)
\( p_{pc_1} \)
\( p_{cit_k} \)

\( W^{p_{cit_1} \rightarrow p} \)
\( W^{p_{pc_1} \rightarrow p} \)
\( W^{p_{cit_k} \rightarrow p} \)

\( p \)

References

\( p_{ref_1} \) --- \( p_{ref_2} \) --- \( p_{ref_i} \)

\( W^{p \rightarrow p_{ref_1}} \)
\( W^{p \rightarrow p_{ref_2}} \)
\( W^{p \rightarrow p_{ref_i}} \)

Reference (ref) papers

: Citation (cit) paper

: 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_{cit_1} )</th>
<th>( p_{cit_2} )</th>
<th>( p_{cit_3} )</th>
<th>( \cdots )</th>
<th>( p_{cit_j} )</th>
<th>( \cdots )</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></td>
<td>0.735</td>
<td></td>
<td>0.687</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_2 )</td>
<td>0.656</td>
<td></td>
<td></td>
<td>0.328</td>
<td></td>
<td>0.436</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_3 )</td>
<td>0.764</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.527</td>
<td>0.385</td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td></td>
<td>0.330</td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_{N-1} )</td>
<td>0.383</td>
<td></td>
<td></td>
<td>0.248</td>
<td></td>
<td></td>
<td>0.176</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>

\( p_i \ (i = 1, 2, \ldots, N) \) : All papers in dataset
\( p_{cit_j} \ (j = 1, 2, \ldots, N) \) : Papers as citation papers in dataset
**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>(\ldots)</td>
<td>0.735</td>
<td>(\ldots)</td>
<td>0.687</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_2)</td>
<td>0.656</td>
<td></td>
<td>(\ldots)</td>
<td>0.328</td>
<td>(\ldots)</td>
<td>0.436</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_3)</td>
<td>0.764</td>
<td></td>
<td>(\ldots)</td>
<td>0.527</td>
<td>(\ldots)</td>
<td>0.385</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_{tgt})</td>
<td></td>
<td>0.581</td>
<td>(\ldots)</td>
<td></td>
<td>(\ldots)</td>
<td>0.330</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_{N-1})</td>
<td>0.383</td>
<td></td>
<td>0.248</td>
<td></td>
<td>(\ldots)</td>
<td>0.176</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_{N-2})</td>
<td>0.654</td>
<td></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></td>
<td>0.430</td>
<td>(\ldots)</td>
<td>0.226</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>

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

\[\text{sim}: \text{cosine similarity}\]
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>------</th>
<th>$p_{cit_j}$</th>
<th>------</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>0.687</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.656</td>
<td></td>
<td>0.328</td>
<td>0.436</td>
<td></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>$p_{tgt}$</td>
<td>0.581</td>
<td></td>
<td></td>
<td>0.330</td>
<td></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>0.176</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{N-2}$</td>
<td>0.654</td>
<td></td>
<td></td>
<td>0.525</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_N$</td>
<td>0.265</td>
<td></td>
<td>0.430</td>
<td>0.226</td>
<td></td>
<td></td>
<td></td>
<td></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>......</th>
<th>$p_{cit_j}$</th>
<th>......</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></td>
<td>0.527</td>
<td></td>
<td>0.385</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{tgt}$</td>
<td>0.581</td>
<td></td>
<td></td>
<td>0.330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{N-1}$</td>
<td>0.383</td>
<td>0.248</td>
<td></td>
<td>0.176</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{N-2}$</td>
<td>0.654</td>
<td></td>
<td></td>
<td>0.525</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_N$</td>
<td>0.265</td>
<td></td>
<td></td>
<td>0.430</td>
<td></td>
<td>0.226</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pearson correlation

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

$p_i$ (i = 1, 2, ..., N): All papers in dataset
$p_{cit_j}$ (j = 1, 2, ..., 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>$p_{cit_j}$</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>0.687</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.656</td>
<td></td>
<td>0.328</td>
<td>0.436</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.764</td>
<td>0.152</td>
<td></td>
<td>0.527</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>0.211</td>
<td>0.330</td>
<td>0.472</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>
</tr>
<tr>
<td>$p_{N-2}$</td>
<td>0.654</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>0.430</td>
<td>0.226</td>
<td></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

“potential citation papers” (e.g., set to 3)
Proposed Method

Identified Potential Citation Papers

\[ p_{cit_1} \rightarrow p_{cit_2} \rightarrow p_{cit_3} \rightarrow p_{cit_{N-2}} \rightarrow p_{cit_{N-1}} \rightarrow p_{tgt} \]

Potential citation paper
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></th>
<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></td>
<td>0.628</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.233</td>
<td>0.147</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.147</td>
<td></td>
<td>0.265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td></td>
<td></td>
<td></td>
<td>0.265</td>
<td></td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.628</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Imputed matrix

<table>
<thead>
<tr>
<th></th>
<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>
<td>0.628</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.233</td>
<td>1.000</td>
<td>0.147</td>
<td>0.476</td>
<td>0.156</td>
</tr>
<tr>
<td>$p_3$</td>
<td>0.723</td>
<td>0.147</td>
<td>1.000</td>
<td>0.265</td>
<td>0.521</td>
</tr>
<tr>
<td>$p_4$</td>
<td>0.538</td>
<td>0.476</td>
<td>0.265</td>
<td>1.000</td>
<td>0.268</td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.628</td>
<td>0.156</td>
<td>0.521</td>
<td>0.268</td>
<td>1.000</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?

<table>
<thead>
<tr>
<th></th>
<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>?</td>
<td>?</td>
<td>0.628</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.233</td>
<td>1.000</td>
<td>0.147</td>
<td>0.476</td>
<td>0.156</td>
</tr>
<tr>
<td>$p_4$</td>
<td>0.538</td>
<td>0.476</td>
<td>0.265</td>
<td>1.000</td>
<td>0.268</td>
</tr>
<tr>
<td>$p_5$</td>
<td>0.628</td>
<td>0.156</td>
<td>0.521</td>
<td>0.268</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

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

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

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

(2) Leveraging Fragments in Potential Citation Papers

[frg-SIM]: Fragments with cosine similarity weighting

\[
F^p = \sum_{x=1}^{j} W \cdot P_{pcx} \cdot f_{pcx} + \sum_{y=1}^{k} W \cdot P_{city} \cdot f_{city} + \sum_{x=1}^{j} W \cdot P_{pcx} \cdot f_{pcx} + \sum_{y=1}^{k} W \cdot P_{city} \cdot f_{city} + \sum_{z=1}^{l} W \cdot P_{refz} \cdot f_{refz}
\]

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

(2) Leveraging Fragments in Potential Citation Papers

[frg-TW]: [frg-SIM] with tunable weight

Frac = \sum_{x=1}^{i} W^{p_{pcx}} f^{p_{pcx}} + \sum_{y=1}^{k} W^{p_{city}} f^{p_{city}} + (1 - \alpha) (f^p + \sum_{x=1}^{i} W^{p_{pcx}} f^{p_{pcx}} + \sum_{y=1}^{k} W^{p_{city}} f^{p_{city}} + \sum_{z=1}^{l} W^{p_{refz}} f^{p_{refz}})

Full text

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

Experimental Data
(Basic dataset has been released from

(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:pc] Parameter tuning to discover potential citation papers

(2) Leveraging fragments in potential citation papers*
   [Tune:frg-SIM] Fragments with cosine similarity weighting
   [Tune:frg-TW] [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 (CTR)
Combines ideas from collaborative filtering and content analysis based on probabilistic topic modeling

<table>
<thead>
<tr>
<th></th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>( \ldots )</th>
<th>( p_j )</th>
<th>( \ldots )</th>
<th>( p_N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>( \lambda v )</td>
<td>( \theta )</td>
<td>( \lambda u )</td>
<td>( v )</td>
<td>( r )</td>
<td>( u )</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \lambda v )</td>
<td>( \theta )</td>
<td>( \lambda u )</td>
<td>( v )</td>
<td>( r )</td>
<td>( u )</td>
</tr>
<tr>
<td>( u_i )</td>
<td>1</td>
<td>1</td>
<td>( r_{ij} )</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \lambda v )</td>
<td>( \theta )</td>
<td>( \lambda u )</td>
<td>( v )</td>
<td>( r )</td>
<td>( u )</td>
</tr>
<tr>
<td>( u_U )</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( r_{ij} \in \{0,1\} \) whether user \( u_i \) includes paper \( p_j \) in the user’s preference
(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></td>
<td></td>
</tr>
<tr>
<td>frg-SIM (Full text + Conclusion)</td>
<td>0.572</td>
<td>0.787</td>
</tr>
<tr>
<td>frg-TW ((\alpha=0.4, \text{Full text + Conclusion}))</td>
<td><strong>0.579</strong></td>
<td><strong>0.793</strong></td>
</tr>
<tr>
<td>Baseline system</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Sugiyama and Kan, JCDL’10]</td>
<td>0.525</td>
<td>0.751</td>
</tr>
<tr>
<td>(Weight “SIM,” Th=0.4,(\gamma=0.23, d=3))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Nascimento et al., JCDL’11]</td>
<td>0.336</td>
<td>0.438</td>
</tr>
<tr>
<td>(“Frequency of bi-gram” obtained from title and abstract)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Wang and Blei., KDD’11]</td>
<td>0.393</td>
<td>0.495</td>
</tr>
<tr>
<td>(“In-matrix prediction” in collaborative topic regression)</td>
<td></td>
<td></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

Identified Potential Citation Papers

Target paper:
“Understanding mobile user’s behavior”

Interdisciplinary paper

“Mobile Technology”

“Mobile Technology”

“Mobile Technology”
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)
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 \( a \)."

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

"like" app user

Probability that user \( u \) likes app \( a \) given that app \( a \) has Twitter-follower \( t \).

Derived from Pseudo-Documents and Pseudo-Words.

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

Followed by:
- twitterID_{10}
- twitterID_{12}

Followed by:
- twitterID_{10}
- twitterID_{12}
- twitterID_{29}

Followed by:
- twitterID_{29}
- twitterID_{31}

Followed by:
- twitterID_{10}, DISLIKED
- twitterID_{12}, DISLIKED
- twitterID_{10}, LIKED
- twitterID_{12}, LIKED
- twitterID_{29}, LIKED
- twitterID_{29}, LIKED
- twitterID_{29}, LIKED
- twitterID_{31}, LIKED

Pseudo-document u

Twitter-follower ID
Preference indicator
Example of Apple’s iTunes App Store

**App developer**
By Apple

**App description**
Clips is a new app for making fun videos to share with friends, family, and the world. With a few taps you can create and send a video message or tell a quick story with animated text, graphics and emoji, music, and more.

**App genres**
Photo & Video

**Free**
Category: Photo & Video
Updated: Jul 20, 2017
Version: 1.1
Size: 55.3 MB
Languages: English, Catalan, Chinese (Hong Kong), Croatian, Czech, Danish, Dutch, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Malay, ...
Constructing Latent Groups

Pseudo-documents
- Twitter followers (T)
- App genres (G)
- App developers (D)
- Words in app description (W)

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

Probability that user \( u \) likes app \( a \) given that app \( a \) has Twitter-follower \( t \)

Per-topic word distribution

Per-document topic distribution

LDA

Per-topic word distribution

Per-document topic distribution
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 performance comparison]

- All = All features
- T = Twitter-followers
- G = Genres
- D = Developers
- W = Words
RQ2: How does our method compare with other techniques?

![Graph showing comparison of different techniques](image-url)
**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>AmpKit+ (Music)</td>
<td>Photo &amp; Video (12%)</td>
</tr>
<tr>
<td>Games (12%)</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/iomusicandyou)
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 metacomic. 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 :)
Multi-Document Abstractive Summarization

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama:
- “Multi-Document Abstractive Summarization Using ILP based Multi-Sentence Compression” (IJCAI ’15)
Approaches to Summarization

Extractive summarization
• Selects few sentences
• Loses some information
• Does not resemble human-written summaries

Abstractive summarization
• Uses information from multiple sentences
• Can contain more complete and all necessary information
• Resembles human-written summaries

Challenges in abstractive summarization
• How do we generate abstractive summaries from multiple documents?
• How do we maximize informativeness and linguistic quality of the summaries?
Proposed Approach

Two steps:

Step1: Sentence clustering
- Identify the most important document
- Generate cluster of similar sentences

Step2: Summary generation
- Construct $K$ shortest path from word graph
- Select the paths that maximize informativeness and linguistic quality using integer linear programming (ILP)
Step 1: Sentence Clustering

(1) Identify the most important document $D_{imp}$ in document set $D$

Comparative approaches
- LexRank [Erkan and Radev, JAIR’04]
- Maximum pairwise cosine similarity
- Overall document collection similarity
Step1: Sentence Clustering

Document set $D$

(2) Initialize each cluster by assigning each sentence in $D_{imp}$
Step 1: Sentence Clustering

(3) Align sentences from other documents based on cosine similarity
Step 1: Sentence Clustering

(4) Remove and reorder clusters
- Remove clusters that have less than \(|D|/2\) sentences
- Compare the followings to reorder the clusters
  * Majority ordering (MO)
  * Average position ordering (APO)
Step1: Sentence Clustering

(4) Remove and reorder clusters
- Remove clusters that have less than $|D|/2$ sentences
- Compare the followings to reorder the clusters
  * Majority ordering (MO)
  * Average position ordering (APO)
Step 2: Summary Generation

Goal: Generate one sentence from each cluster

- Word-graph [Filippova, 2010]
- Example:
  Eg. 1) The American killed in the crash was 31-year-old Seth J. Foti, a diplomatic courier carrying classified information.
  Eg. 2) 31-year-old Seth Foti was carrying pouches containing classified information.
- Generate new sentences between the start and end nodes

How do we select the best sentence with acceptable informativeness and linguistic quality?
**Integer Linear Programming Formulation**

Randomly select 200 sentences from the set of generated sentences

\[ p_i^{C_j} : \text{Each path in a cluster } C_j \]

Maximize

\[
\sum_{j=1}^{m} \sum_{i=1}^{K} \frac{1}{T(p_i^{C_j})} \cdot I(p_i^{C_j}) \cdot LQ(p_i^{C_j}) \cdot p_i^{C_j}
\]

- Number of tokens in the sentence
- Informativeness: “TextRank” [Mihalcea and Tarau, EMNLP’04]

**Constraints**

- Select only one path
- Do not include similar sentences
Experimental Results

- Evaluation measures: ROUGE
- Dataset: DUC2004 and DUC2005

<table>
<thead>
<tr>
<th>Baselines</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
<th>Baselines</th>
<th>ROUGE-L</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GreedyKL</td>
<td>0.08658</td>
<td>0.13253</td>
<td>Random</td>
<td>0.26395</td>
<td>0.09066</td>
</tr>
<tr>
<td>FreqSum</td>
<td>0.08218</td>
<td>0.12448</td>
<td>Centroid</td>
<td>0.32562</td>
<td>0.11007</td>
</tr>
<tr>
<td>Centroid</td>
<td>0.08139</td>
<td>0.12642</td>
<td>LexRank</td>
<td>0.33179</td>
<td>0.12021</td>
</tr>
<tr>
<td>TsSum</td>
<td>0.08068</td>
<td>0.12209</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LexRank</td>
<td>0.07796</td>
<td>0.12484</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State-of-the-arts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPP</td>
<td>0.10079</td>
<td>0.14556</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Submodular</td>
<td>0.09602</td>
<td>0.14227</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RegSum</td>
<td>0.09712</td>
<td>0.13812</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASSY04</td>
<td>0.09168</td>
<td>0.13250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCCAMSV</td>
<td>0.09420</td>
<td>0.13105</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICSIsumm</td>
<td>0.09585</td>
<td>0.13314</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLASSY11</td>
<td>0.08912</td>
<td>0.12779</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstractive Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$MD_{LexRank}^{imp}$ + APO + MSC [Filippova, 2010]</td>
<td>0.09612</td>
<td>0.13911</td>
<td>$MD_{LexRank}^{imp}$ + APO + MSC</td>
<td>0.35589</td>
<td>0.12211</td>
</tr>
<tr>
<td>$MD_{LexRank}^{imp}$ + MO + ILP</td>
<td>0.09799</td>
<td>0.13884</td>
<td>$MD_{LexRank}^{imp}$ + MO + ILP</td>
<td>0.35281</td>
<td>0.12107</td>
</tr>
<tr>
<td>$MD_{LexRank}^{imp}$ + APO + ILP</td>
<td>0.10317</td>
<td>0.14218</td>
<td>$MD_{LexRank}^{imp}$ + APO + ILP</td>
<td>0.35342</td>
<td>0.12117</td>
</tr>
<tr>
<td>$MD_{CosStm}^{imp}$ + MO + ILP</td>
<td>0.09799</td>
<td>0.13884</td>
<td>$MD_{CosStm}^{imp}$ + MO + ILP</td>
<td>0.35661</td>
<td>0.12331</td>
</tr>
<tr>
<td>$MD_{CosStm}^{imp}$ + APO + ILP</td>
<td>0.10577</td>
<td>0.14215</td>
<td>$MD_{CosStm}^{imp}$ + APO + ILP</td>
<td>0.35577</td>
<td>0.12298</td>
</tr>
<tr>
<td>$MD_{DocSetStm}^{imp}$ + MO + ILP [ILPSumm]</td>
<td>0.11199 †</td>
<td>0.14765 †</td>
<td>$MD_{DocSetStm}^{imp}$ + MO + ILP [ILPSumm]</td>
<td>0.35772 †</td>
<td>0.12411 †</td>
</tr>
<tr>
<td>$MD_{DocSetStm}^{imp}$ + APO + ILP</td>
<td>0.11712</td>
<td>0.13567</td>
<td>$MD_{DocSetStm}^{imp}$ + APO + ILP</td>
<td>0.35679</td>
<td>0.12393</td>
</tr>
</tbody>
</table>

WING, NUS
Abstractive Summary (ILPSumm)
Lebanese Foreign Minister Kamal Kharrazi made the mediation offer Sunday, in a telephone conversation with his Syrian counterpart, Farouk al-Sharaa. Egyptian President Hosni Mubarak met here Sunday with Syrian President Hafez Assad to show Lebanon's support for Syria and Turkey. In a show of force on Friday, Turkish troops were deployed this week on the Turkish-Syrian border to eradicate Krudish rebel bases.

Extractive Summary (DPP)
Egyptian President Hosni Mubarak met here Sunday with Syrian President Hafez Assad to try to defuse growing tension between Syria and Turkey. The talks in Damascus came as Turkey has massed forces near the border with Syria after threatening to eradicate Kurdish rebel bases in the neighboring country. Egypt already has launched a mediation effort to try to prevent a military confrontation over Turkish allegations that Syria is harboring Turkish Kurdish rebels.

Hand-written Summary
Tensions between Syria and Turkey increased as Turkey sent 10,000 troops to its border with Syria. The dispute comes amid accusations by Turkey that Syria helping Kurdish rebels based in Syria. Kurdish rebels have been conducting cross border raids into Turkey in an effort to gain Kurdish autonomy in the region.
Our Recent Work on Health Informatics

Kishaloy Halder, Min-Yen Kan, Kazunari Sugiyama:
- “Health Forum Thread Recommendation Using an Interest Aware Topic Models”
  (CIKM ’17, To appear)
Health Forum Recommendation

Example of asthma forum in “patientslikeme”

Asthma Problem
Tags: Hospital

Mashawilliams on Jul 10, 2013
Nowadays Asthma has become very common, being a medical student I have studied this disease in detail, my advice would say see a specialist as soon as possible. I can provide you name of some specialist doctors from which you can take help. Details: Dr. PRAMOD...

Eddie the Poet on Jul 10, 2013
or not which your are taking. In addition, you may have allergies causing this. This has been a problem for my life-long asthmatic system. These allergies could include pollen, dust mites (which live in dust), animal pets, and some types of food. Even if...

aasha_arora on Jul 09, 2013
I am a 22 year old girl being diagnosed with asthma. Initially I was suffering with cough problems but didn't notice, then the situation got worsen with more ans more coughing, especially at night. Wheezing Shortness of breath, Chest tightness, pain, or pressure...

Go to post
Outline of Proposed Approach

Two-stage topic model approach

- Interest-aware Topic Model (IATM)
  - Capture each user’s interest

- Joint Normalized Collaborative Topic Regression (JNCTR)
  - Capture user-document interactions
Our Prediction Tasks

(a) In-matrix

(b) Out-of-matrix (for thread)

(c) Out-of-matrix (for user)
Experiments

Experimental Data
• PatientsLikeMe
• HealthBoards

Evaluation Measure
• Mean reciprocal rank (MRR)
• Normalized discounted cumulative gain (nDCG)
• Recall
# Experimental Results

## MRR and nDCG scores

<table>
<thead>
<tr>
<th>Method</th>
<th>PatientsLikeMe (PLM)</th>
<th>HealthBoards (HBD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR</td>
<td>nDCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@5</td>
</tr>
<tr>
<td>1. CF (NMF) [12]</td>
<td>0.171</td>
<td>0.071</td>
</tr>
<tr>
<td>2. AT [27]</td>
<td>0.051</td>
<td>0.019</td>
</tr>
<tr>
<td>3. CTR [32]</td>
<td>0.175</td>
<td>0.075</td>
</tr>
<tr>
<td>4. IATM</td>
<td>0.101</td>
<td>0.039</td>
</tr>
<tr>
<td>5. CAR [26]</td>
<td>0.151</td>
<td>0.065</td>
</tr>
<tr>
<td>6. AT+JNCTR</td>
<td>0.176</td>
<td>0.082</td>
</tr>
<tr>
<td>7. IATM+JNCTR</td>
<td>0.183</td>
<td>0.083</td>
</tr>
</tbody>
</table>
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 into account

• Multi-Document Abstractive Summarization
  - Apply clustering and integer linear programming to maximize information content and linguistic quality

• Health Forum Recommendation
  - Employ two-stage topic model to capture each user’s interest and user-document interactions

Thank you very much!