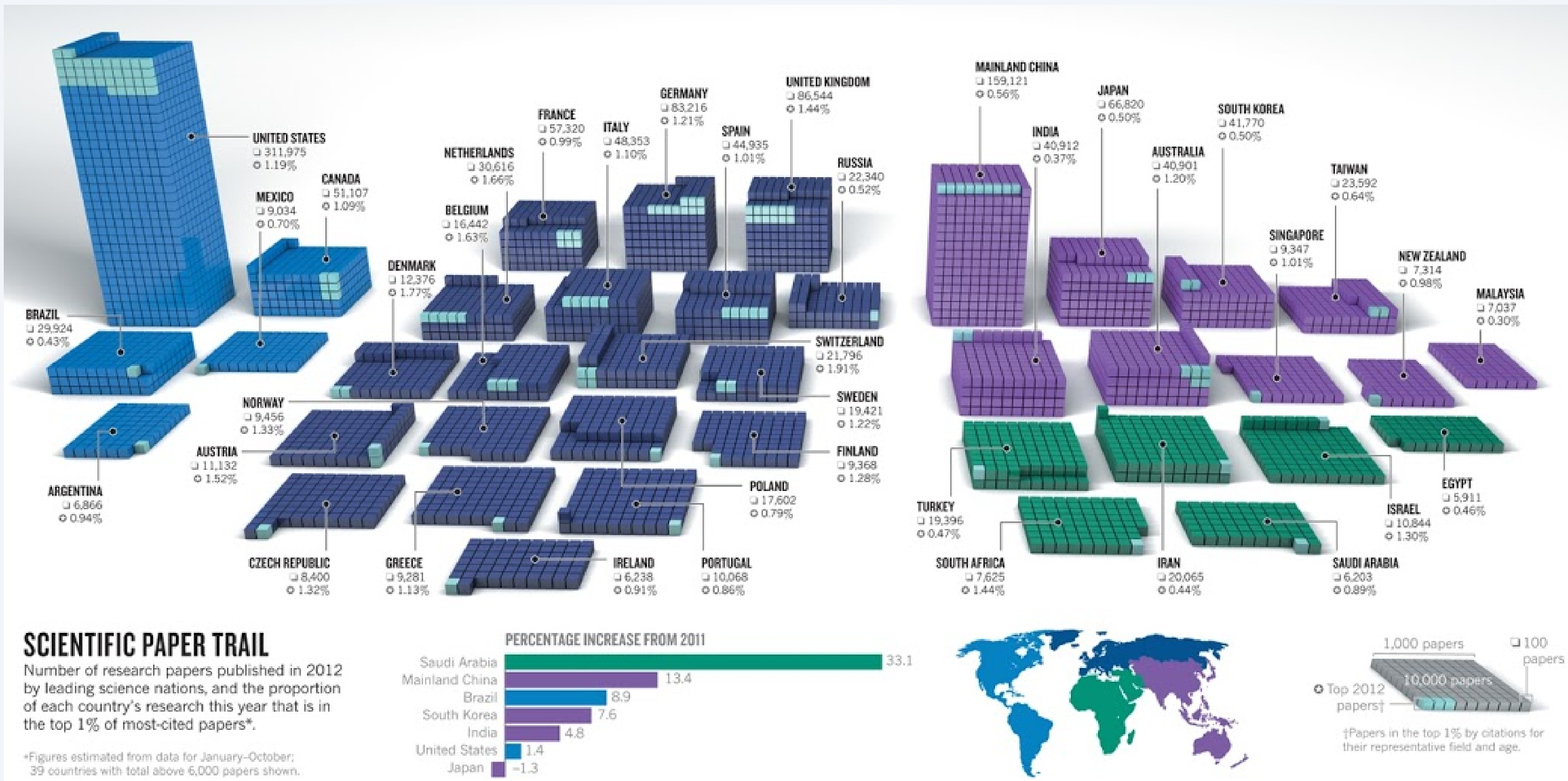
Exploiting Potential Citation Papers in Scholarly Paper Recommendation

Kazunari Sugiyama, Min-Yen Kan

National University of Singapore

Introduction

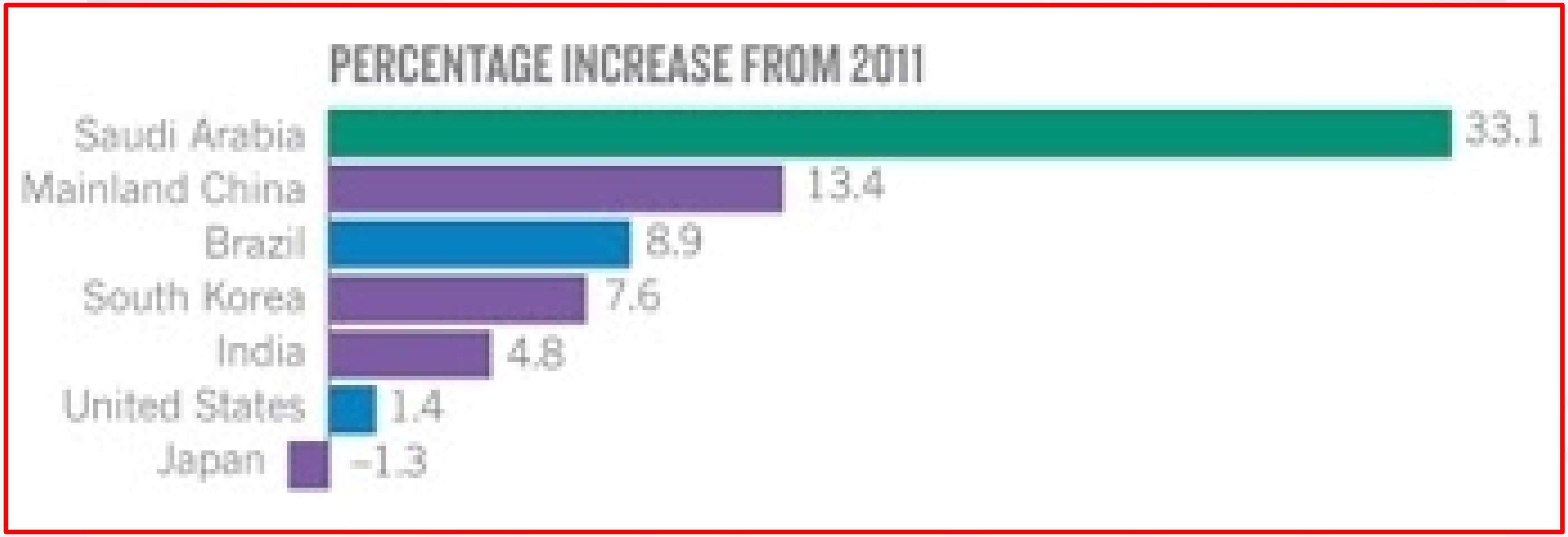
How many papers are published in 2012?



Source: Originally pinned from Nature (http://www.nature.com/news/366-days-2012-in-review-1.12042), with data from Thomson Reuters/Essential Science Indicators.

Introduction

How many papers are published in 2012?



Source: Originally pinned from Nature (http://www.nature.com/news/366-days-2012-in-review-1.12042), with data from Thomson Reuters/Essential Science Indicators.

NUS National University of Singapore

Introduction

Recommendation System

(Especially, content-based system)

Researcher

User profile $\boldsymbol{U}$

Candidate papers to recommend

$$\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_P$$

Recommended papers

Introduction

Feature vector construction for candidate papers

Fragments

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[1,2]
[1]National University of Singapore
[2]NUS Interactive and Digital Media Institute
Computing 1, 13 Computing Drive,
Singapore 117417
kanmy@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 citation 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 investigate 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 nDCG and MRR, when we discover potential citation papers using imputed similarities via collaborative filtering and represent candidate papers using both the full text and assigning more weight to the conclusion sections.

to the sum total of human knowledge. While certainly advantageous, this creates a problem of over abundance, commonly known as "information overload": where researchers find an overwhelming number of matches to their search queries, but for which the majority are largely irrelevant to their latent information needs.

Work in recommendation systems is one promising approach to address the information overload. In digital library studies, this approach has been employed to obtain and refine search results to satisfy each user's information needs [10, 18, 3, 23, 13]. However, these approaches do not fully leverage the user's context, largely relying on the idea of session-as-context. This legacy is ported from research in Web search, where session click-through data are used to form the context. To address this problem, in our previous work, we observed that the scholarly context allows us to leverage the role of the searcher-as-author [16]. We modeled a searcher's context in the form of a profile by capturing previous research interests embodied in their past publications, and showed elevated success at scholarly paper recommendation. Our approach in [16] also

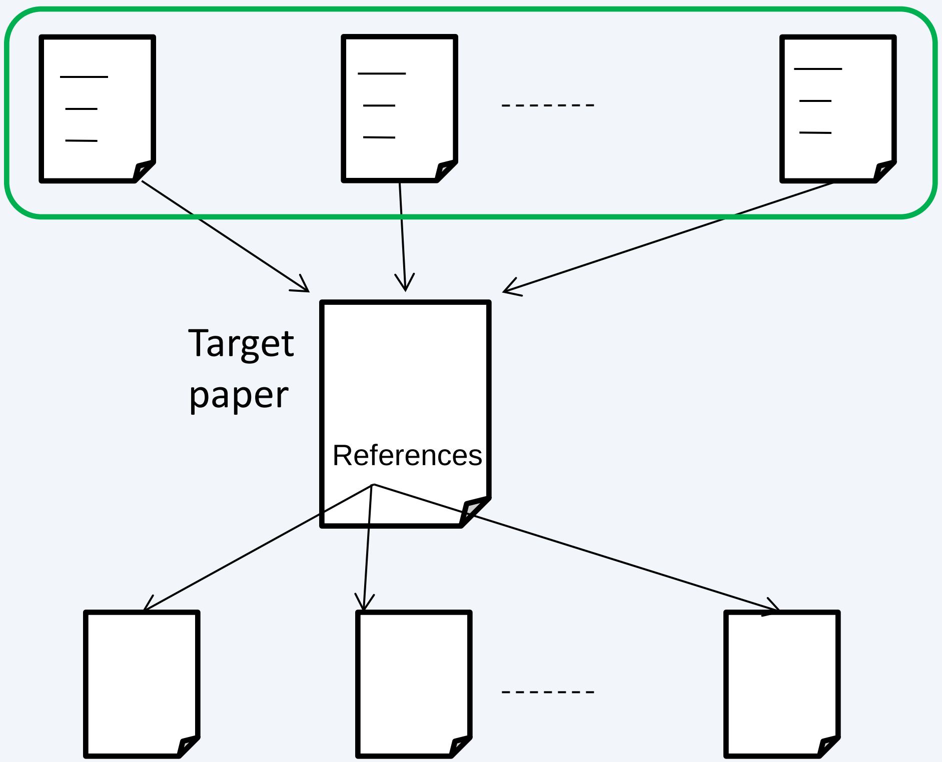
Full text

Exploiting Potential Citation Papers in Scholarly Paper Recommendation

Introduction

Feature vector construction for candidate papers

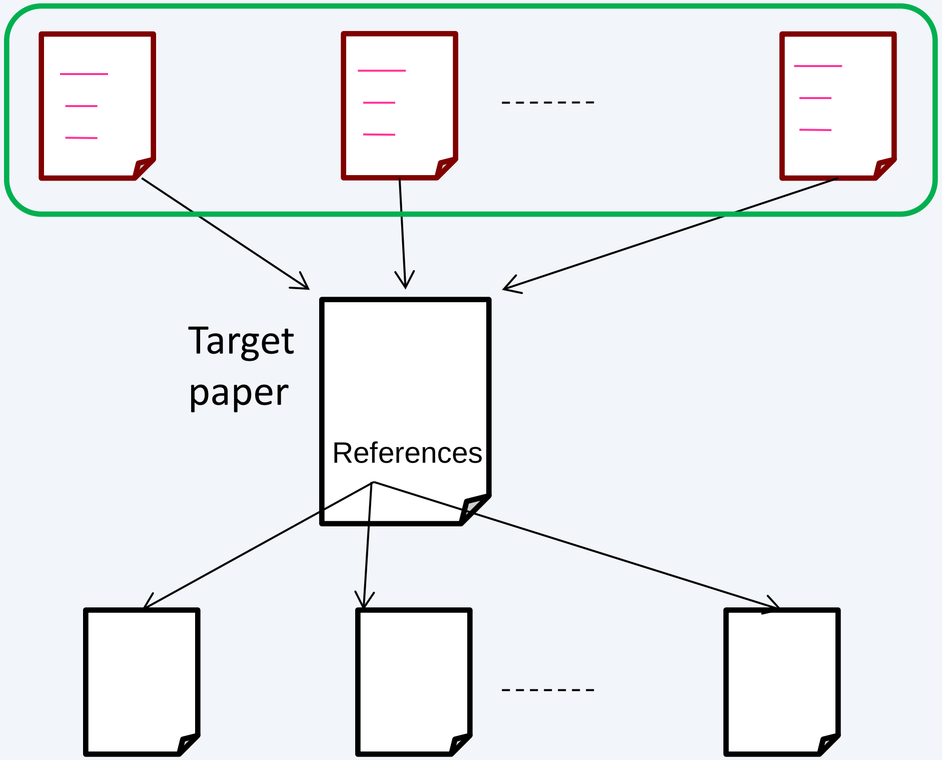
Citation and Reference Papers



Introduction

Feature vector construction for candidate papers

Citation and Reference Papers



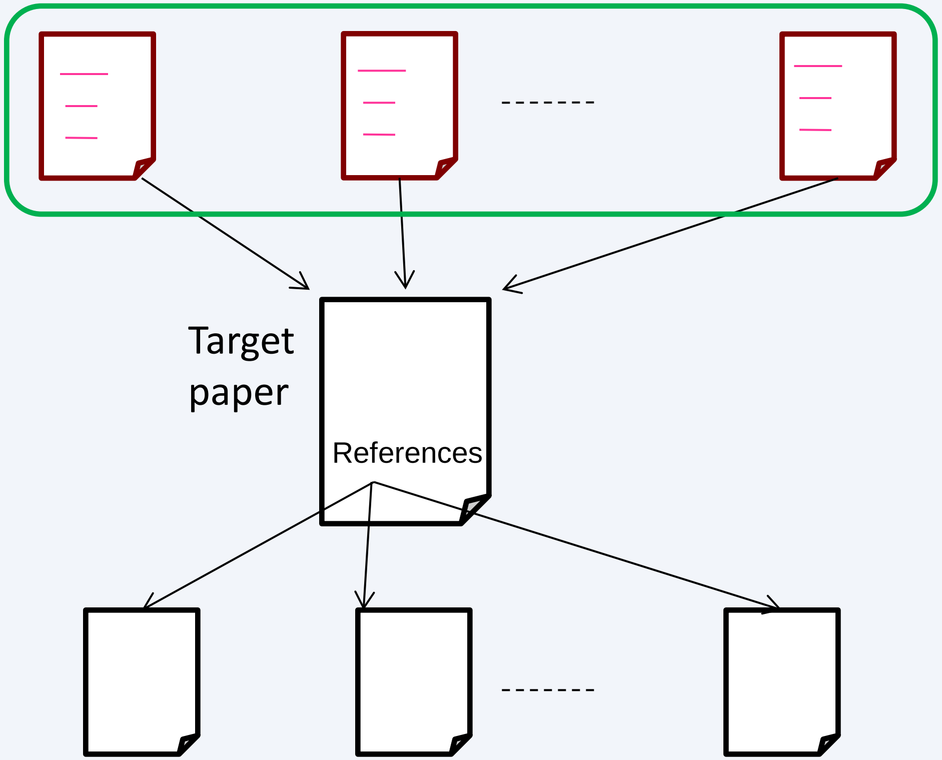
Citation papers:

- **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,” …**

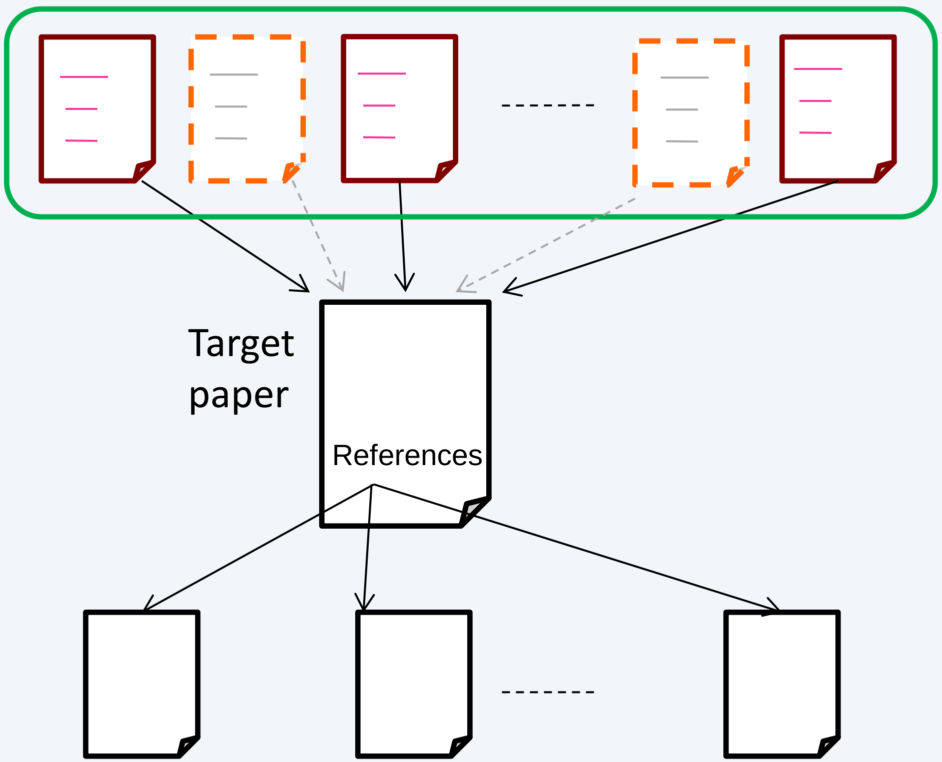
Authors of citation papers

- **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

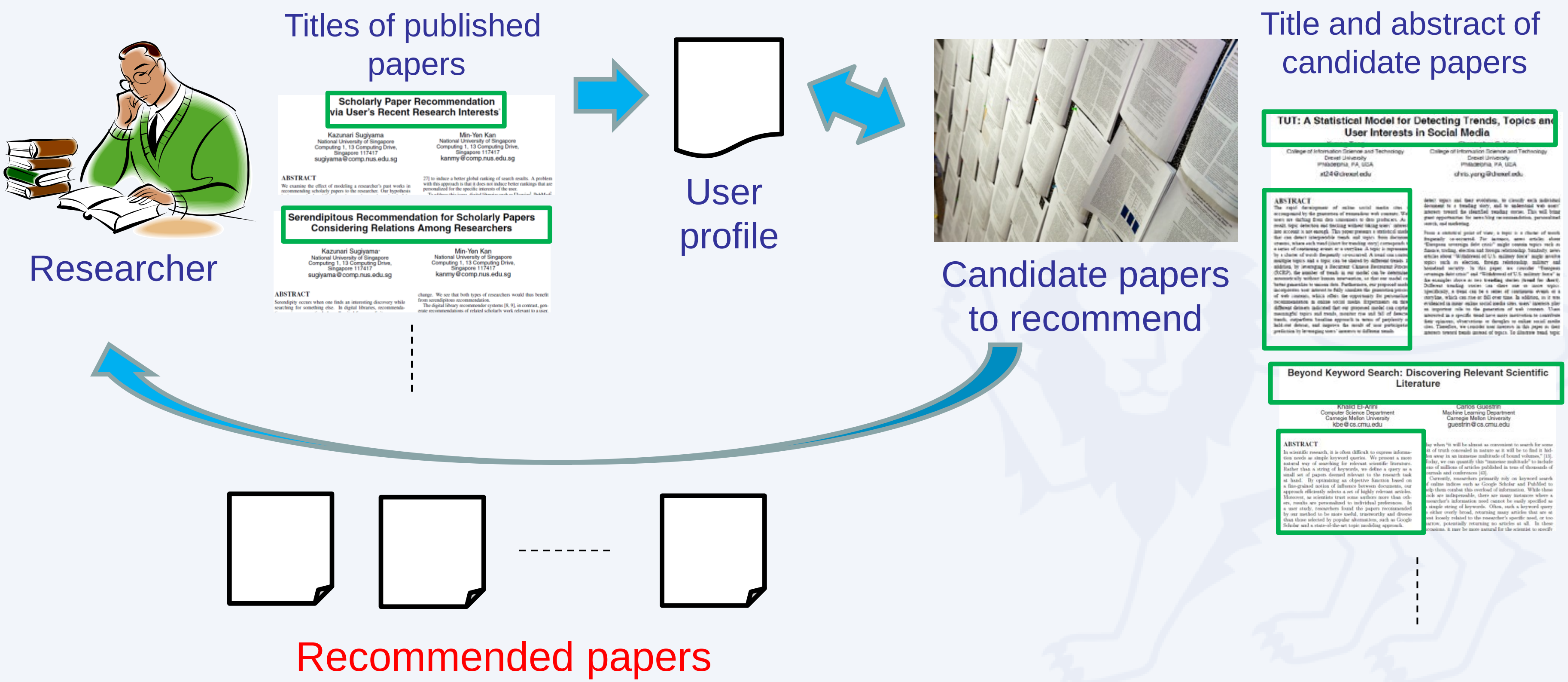
- To find potential citation papers to model candidate papers to recommend much better for better recommendation

- To refine the use of citation papers in characterizing candidate papers to recommend using fragments

Related Work

Scholarly Paper Recommendation

[Nascimento et al., JCDL’11]



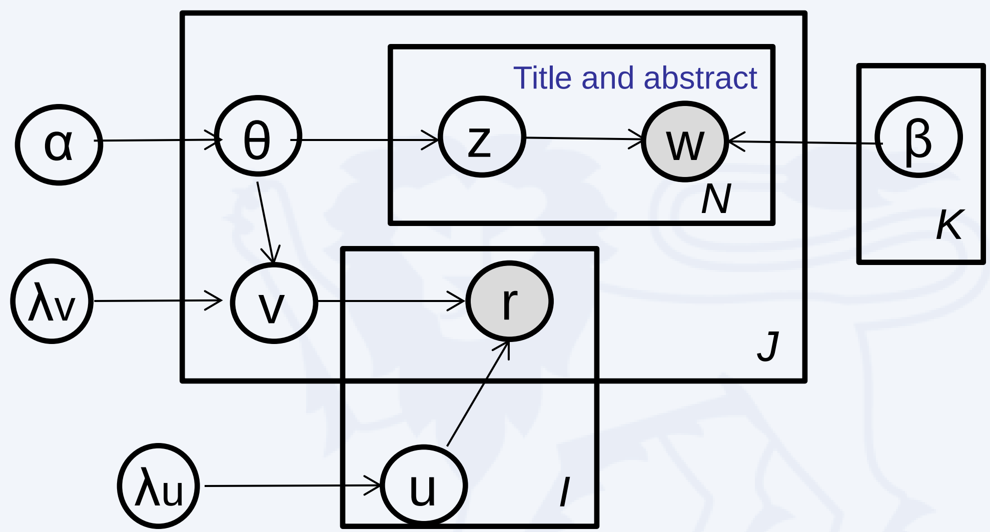
Related Work

Scholarly Paper Recommendation

[Wang and Blei., KDD'11]: Collaborative topic regression

Combines ideas from collaborative filtering and content analysis based on probabilistic topic modeling

	p_1	p_2	------	p_j	------	p_N
u_1				0		
u_2		1		1		1
$\vdots$						
u_i	1	1		r_{ij}		0
$\vdots$						
u_U		0		1		



$r_{ij} \in \{0,1\}$ whether user u_i includes paper p_j In the user's preference

Outline of Baseline System

[Sugiyama and Kan, JCDL'10]

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

$\mathbf{P}_{user}$

Researcher

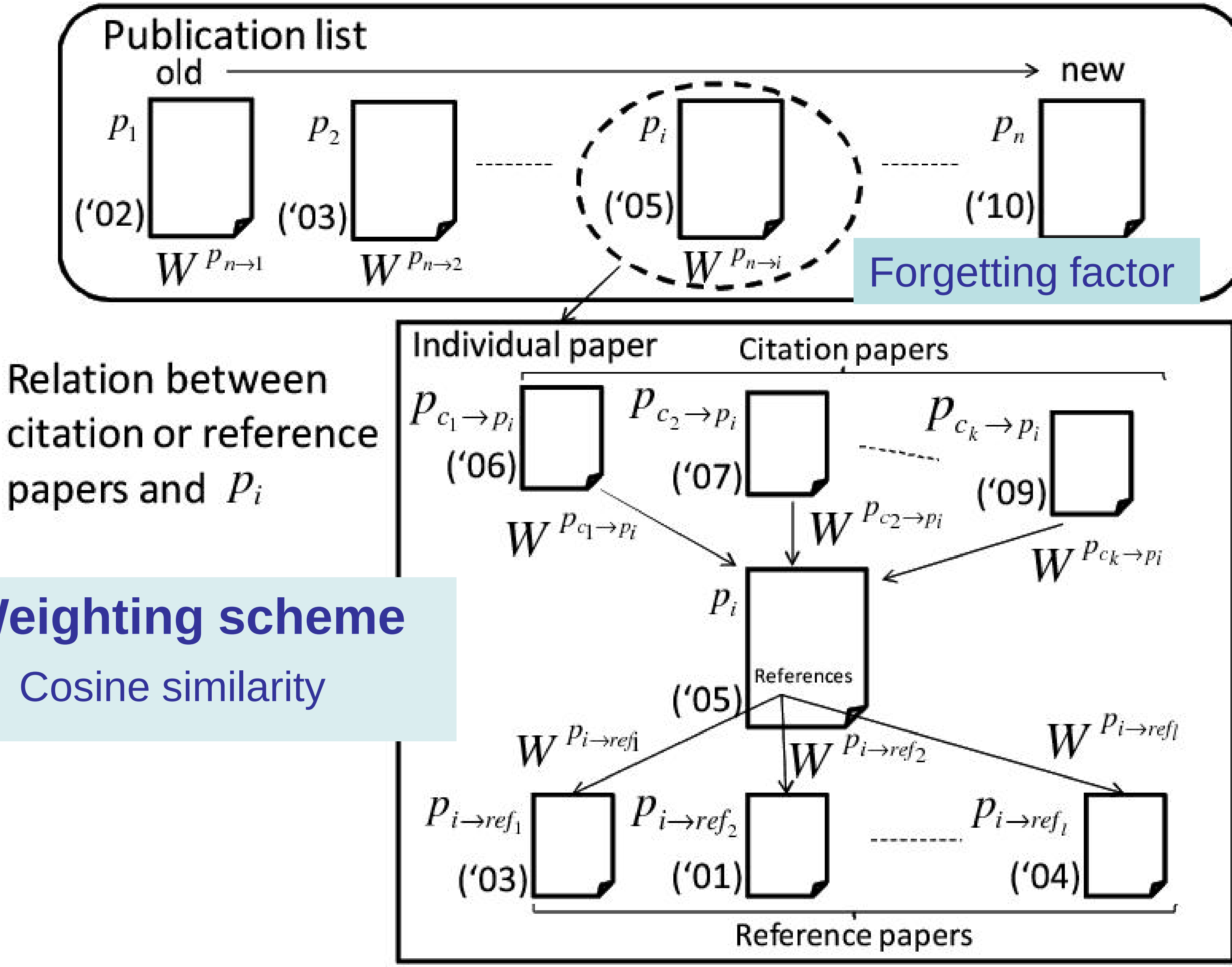
(2) Compute similarity between

$\mathbf{P}_{user}$ and $\mathbf{F}^{p_{rec_j}}$ $(j = 1, \cdots, t)$

Candidate papers to recommend

$\mathbf{F}^{p_{rec_1}}$ to $\mathbf{F}^{p_{rec_t}}$

(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

$\mathbf{P}_{user}$

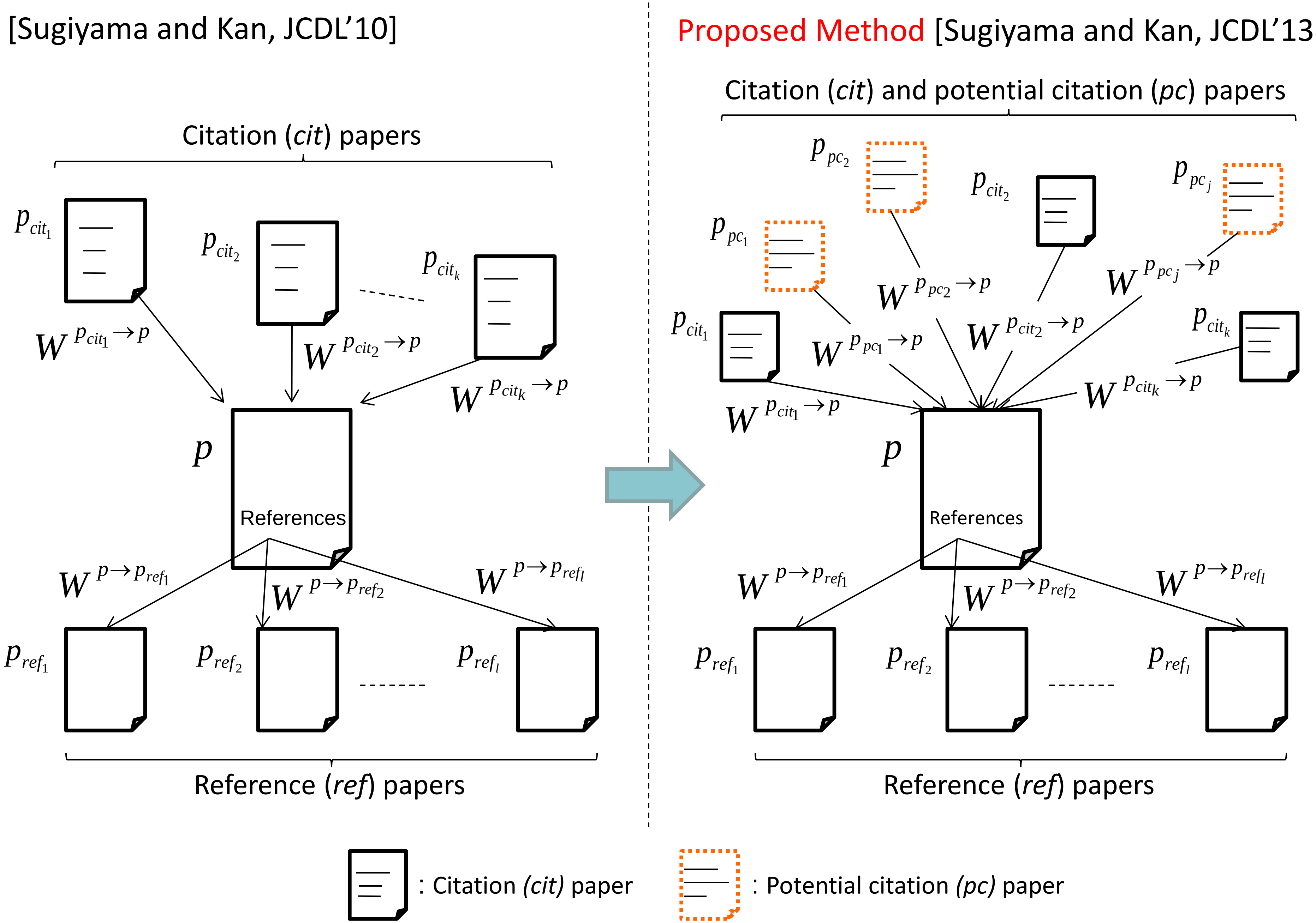
Researcher

(2) Compute similarity between $\mathbf{P}_{user}$ and $\mathbf{F}^{p_{rec_j}}$ $(j = 1, \cdots, t)$

Candidate papers to recommend

$\mathbf{F}^{p_{rec_1}}$ to $\mathbf{F}^{p_{rec_t}}$

(3) Recommend papers with high similarity


[Sugiyama and Kan, JCDL'10]
Proposed Method [Sugiyama and Kan, JCDL'13]
Citation (cit) papers
Citation (cit) and potential citation (pc) papers
p_{cit_1}
p_{cit_2}
p_{cit_k}
p_{pc_1}
p_{pc_2}
p_{pc_j}
$W^{p_{cit_1} \to p}$
$W^{p_{cit_2} \to p}$
$W^{p_{cit_k} \to p}$
$W^{p_{pc_1} \to p}$
$W^{p_{pc_2} \to p}$
$W^{p_{pc_j} \to p}$
p
References
$W^{p \to p_{ref_1}}$
$W^{p \to p_{ref_2}}$
$W^{p \to p_{ref_l}}$
p_{ref_1}
p_{ref_2}
p_{ref_l}
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?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	-------	p_{cit_j}	-------	$p_{cit_{N-2}}$	$p_{cit_{N-1}}$	p_{cit_N}
p_1		0.212			0.735		0.687		
p_2	0.656				0.328		0.436		
p_3		0.764					0.527		0.385
$\vdots$									
p_{tgt}		0.581					0.330		
$\vdots$									
p_{N-1}	0.383				0.248			0.176	
p_{N-2}		0.654					0.525		
p_N		0.265					0.430		0.226

$p_i \ (\mathrm{i} = 1,2,\cdots,N)$: All papers in dataset

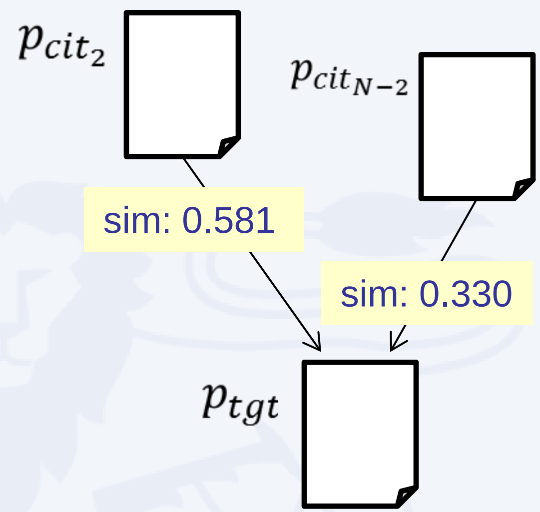
$p_{cit_j} \ (\mathrm{j} = 1,2,\cdots,N)$: Papers as citation papers in dataset

Proposed Method

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	-------	p_{cit_j}	-------	$p_{cit_{N-2}}$	$p_{cit_{N-1}}$	p_{cit_N}
p_1		0.212			0.735		0.687		
p_2	0.656				0.328		0.436		
p_3		0.764					0.527		0.385
$\vdots$									
p_{tgt}		$\boxed{0.581}$					$\boxed{0.330}$		
$\vdots$									
p_{N-1}	0.383				0.248			0.176	
p_{N-2}		0.654					0.525		
p_N		0.265					0.430		0.226



sim: cosine similarity

p_i $(\mathrm{i} = 1,2,\cdots,N)$: All papers in dataset

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

Proposed Method

(1) Leveraging Potential Citation Papers

How are potential citation papers discovered?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	-------	p_{cit_j}	-------	$p_{cit_{N-2}}$	$p_{cit_{N-1}}$	p_{cit_N}	Pearson correlation
p_1		0.212			0.735		0.687			0.538
p_2	0.656				0.328		0.436			0.216
p_3		0.764					0.527		0.385	0.475
⋮										⋮
p_{tgt}		0.581					0.330			---
⋮										⋮
p_{N-1}	0.383				0.248			0.176		0.304
p_{N-2}		0.654					0.525			0.513
p_N		0.265					0.430		0.226	0.487

p_i $(\mathrm{i} = 1,2,\cdots,N)$: All papers in dataset

p_{cit_j} $(\mathrm{j} = 1,2,\cdots,N)$: 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?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	-------	p_{cit_j}	-------	$p_{cit_{N-2}}$	$p_{cit_{N-1}}$	p_{cit_N}	Pearson correlation
p_1		0.212			0.735		0.687			0.538
p_2	0.656				0.328		0.436			0.216
p_3		0.764	0.152				0.527		0.385	0.475
⋮										⋮
p_{tgt}		0.581					0.330			---
⋮										⋮
p_{N-1}	0.383				0.248			0.176		0.304
p_{N-2}		0.654					0.525			0.513
p_N		0.265					0.430		0.226	0.487

p_i $(\mathrm{i} = 1,2,\cdots,N)$: All papers in dataset

p_{cit_j} $(\mathrm{j} = 1,2,\cdots,N)$: 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?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	-------	p_{cit_j}	-------	$p_{cit_{N-2}}$	$p_{cit_{N-1}}$	p_{cit_N}	Pearson correlation
p_1		0.212			0.735		0.687			0.538
p_2	0.656				0.328		0.436			0.216
p_3		0.764	0.152				0.527		0.385	0.475
⋮										⋮
p_{tgt}	**0.435**	0.581	**0.536**	-------	**0.211**	-------	0.330	**0.472**	**0.368**	---
⋮										⋮
p_{N-1}	0.383				0.248			0.176		0.304
p_{N-2}		0.654					0.525			0.513
p_N		0.265					0.430		0.226	0.487

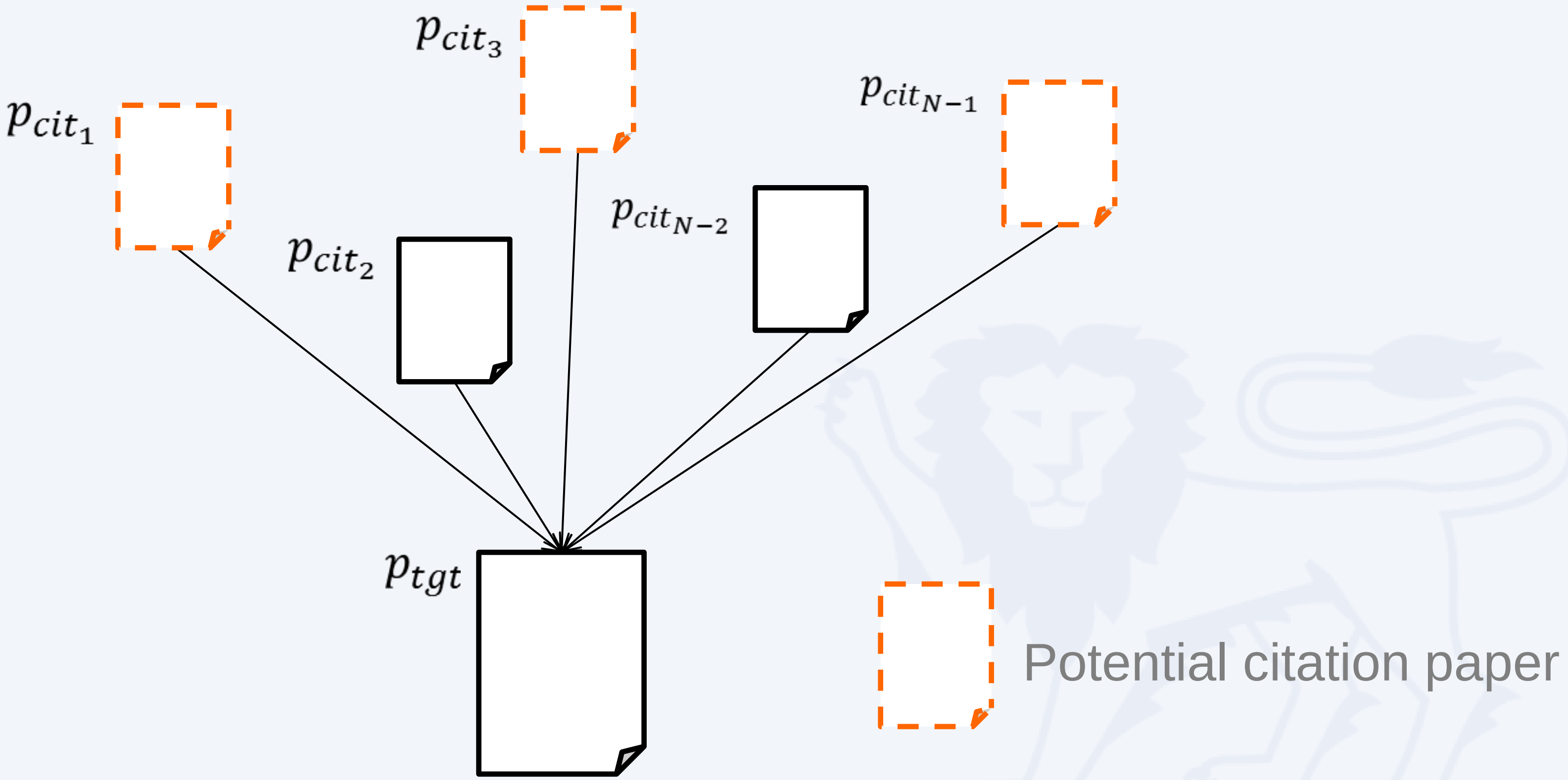
p_i $(\mathrm{i} = 1,2,\cdots,N)$: All papers in dataset

p_{cit_j} $(\mathrm{j} = 1,2,\cdots,N)$: 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 is solved?

Proposed Method

(1) Leveraging Potential Citation Papers

How is the sparsity of matrix is solved?

Original matrix

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1		0.233			0.628
p_2	0.233		0.147		
p_3		0.147		0.265	
p_4			0.265		
p_5	0.628				

The values in the cell:
Cosine similarity between papers

Imputation

Imputed matrix

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**0.723**	**0.538**	0.628
p_2	0.233	1.000	0.147	**0.476**	**0.156**
p_3	**0.723**	0.147	1.000	0.265	**0.521**
p_4	**0.538**	**0.476**	0.265	1.000	**0.268**
p_5	0.628	**0.156**	**0.521**	**0.268**	1.000

Proposed Method

(1) Leveraging Potential Citation Papers

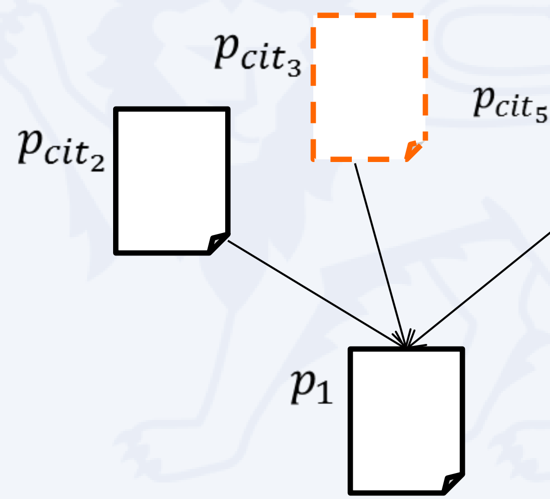
How is the sparsity of matrix is solved?

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**?**	**?**	0.628
p_2	0.233	1.000	0.147	**0.476**	**0.156**
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_4	**0.538**	**0.476**	0.265	1.000	**0.268**
p_5	0.628	**0.156**	**0.521**	**0.268**	1.000

Target paper (p_1) and corresponding Imputed similarities of neighborhood (p_2, p_4, and p_5) from "Imputed matrix"

	p_{cit_1}	p_{cit_2}	p_{cit_3}	p_{cit_4}	p_{cit_5}
p_1	1.000	0.233	**0.682**	**0.453**	0.628

"potential citation papers" (e.g., set to 1)

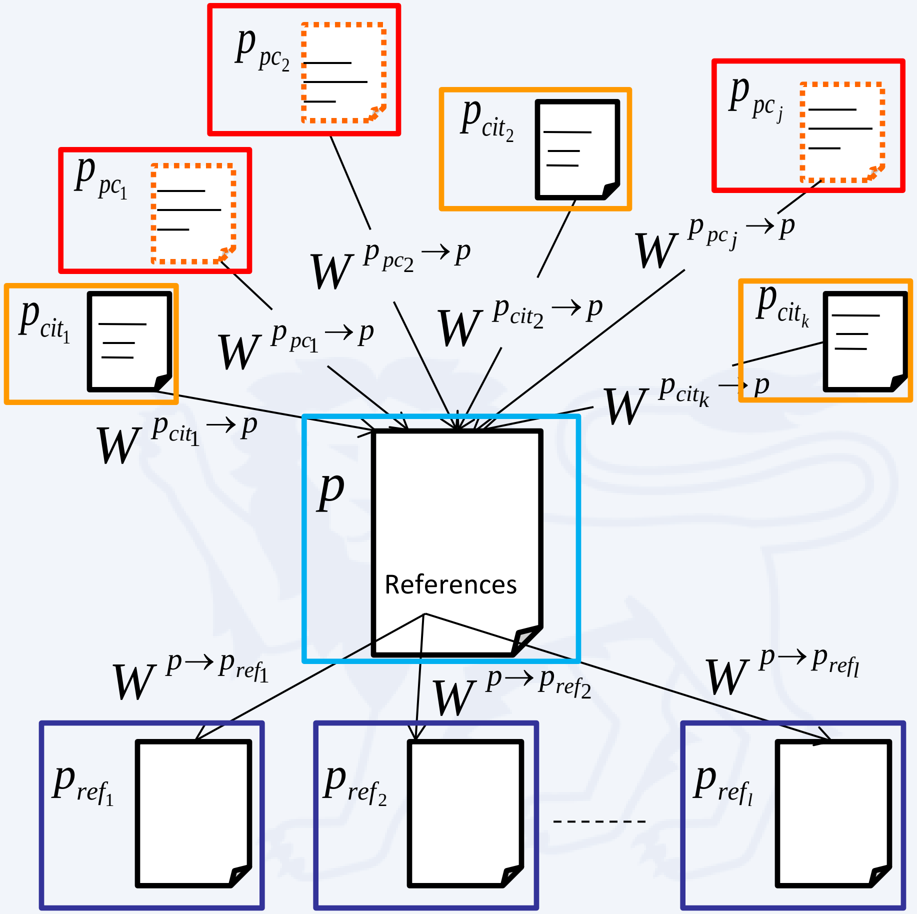


Proposed Method

(1) Leveraging Potential Citation Papers

Feature Vector Construction for Target Papers

$$\mathbf{F}^{p} = \boxed{f^{p}} + \sum_{x=1}^{j} W^{p_{pc_x} \to p} \boxed{f^{p_{pc_x}}} + \sum_{y=1}^{k} W^{p_{cit_y} \to p} \boxed{f^{p_{cit_y}}} + \sum_{z=1}^{l} W^{p \to p_{ref_z}} \boxed{f^{p_{ref_z}}}$$



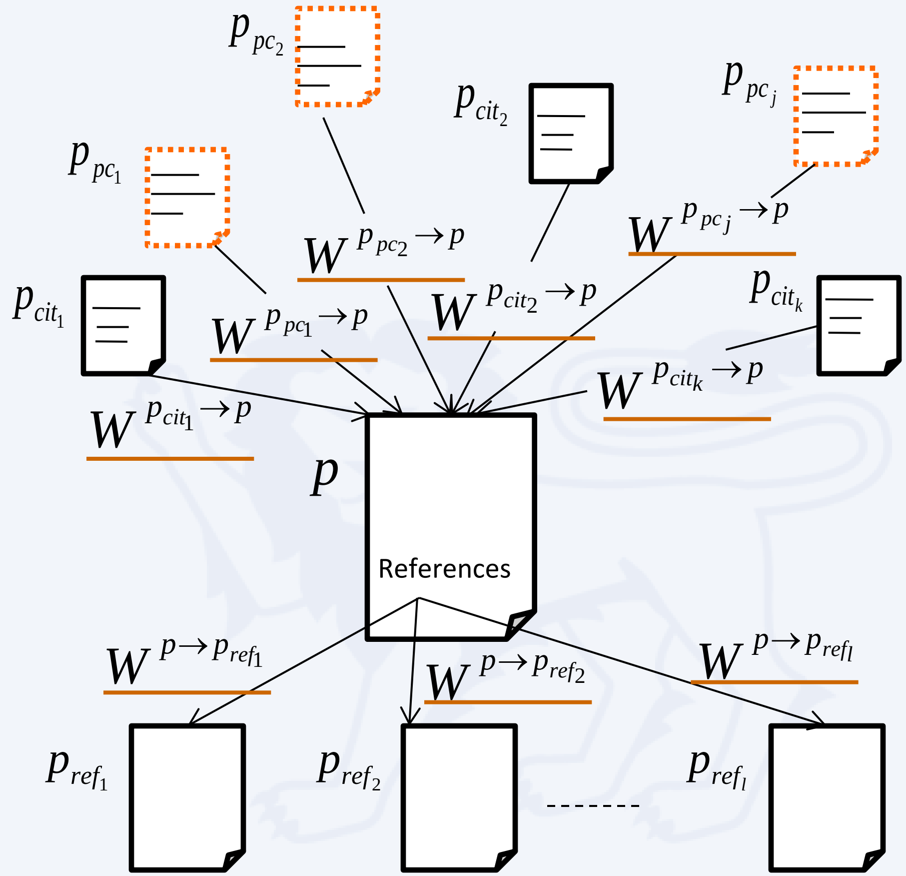
Proposed Method

(1) Leveraging Potential Citation Papers

Feature Vector Construction for Target Papers

$$\mathbf{F}^{p} = f^{p} + \sum_{x=1}^{j} \underline{W^{p_{pc_x} \rightarrow p}}\, f^{p_{pc_x}} + \sum_{y=1}^{k} \underline{W^{p_{cit_y} \rightarrow p}}\, f^{p_{cit_y}} + \sum_{z=1}^{l} \underline{W^{p \rightarrow p_{ref_z}}}\, f^{p_{ref_z}}$$

cosine similarity



Proposed Method

(2) Leveraging Fragments in Potential Citation Papers

- [*frg*-SIM]: Fragments with cosine similarity weighting

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

Proposed Method

(2) Leveraging Fragments in Potential Citation Papers

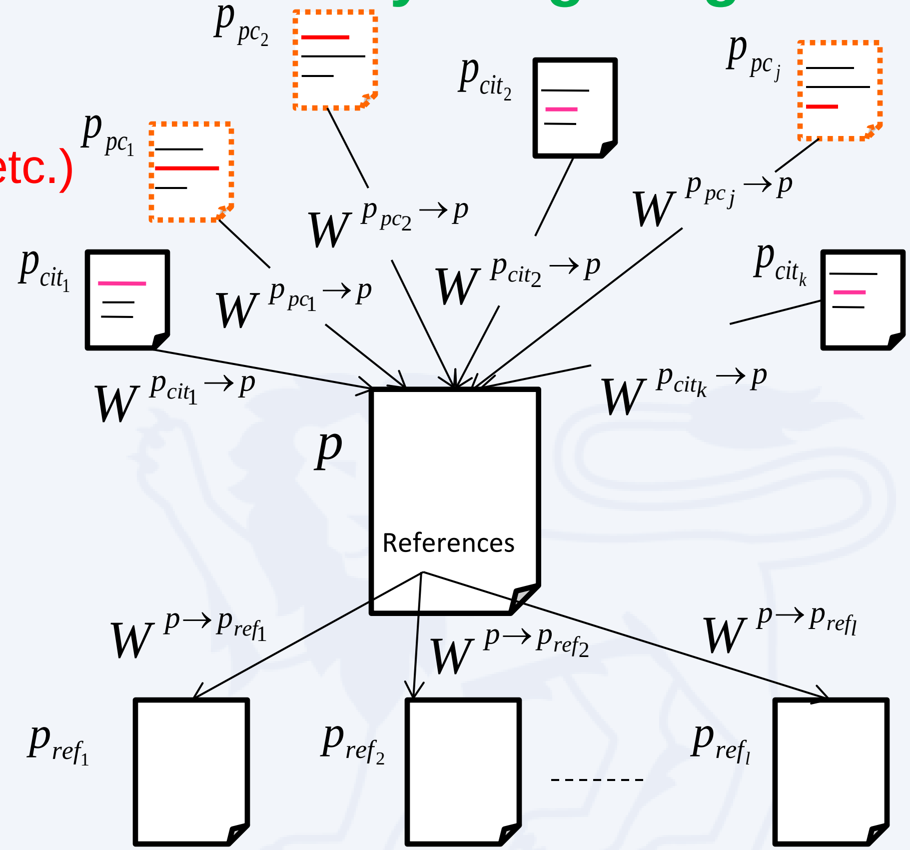
[*frg*-SIM]: Fragments with cosine similarity weighting

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

$$\mathbf{F}^{p} = \boxed{\sum_{x=1}^{j} W_{(frg)}^{p_{pc_x} \to p} \, \underline{f_{(frg)}^{p_{pc_x}}} + \sum_{y=1}^{k} W_{(frg)}^{p_{cit_y} \to p} \, \underline{f_{(frg)}^{p_{cit_y}}}}$$

$$\boxed{\begin{aligned} &+ f^{p} + \sum_{x=1}^{j} W^{p_{pc_x} \to p} f^{p_{pc_x}} \\ &+ \sum_{y=1}^{k} W^{p_{cit_y} \to p} f^{p_{cit_y}} \\ &+ \sum_{z=1}^{l} W^{p \to p_{ref_z}} f^{p_{ref_z}} \end{aligned}}$$

Full text



Proposed Method

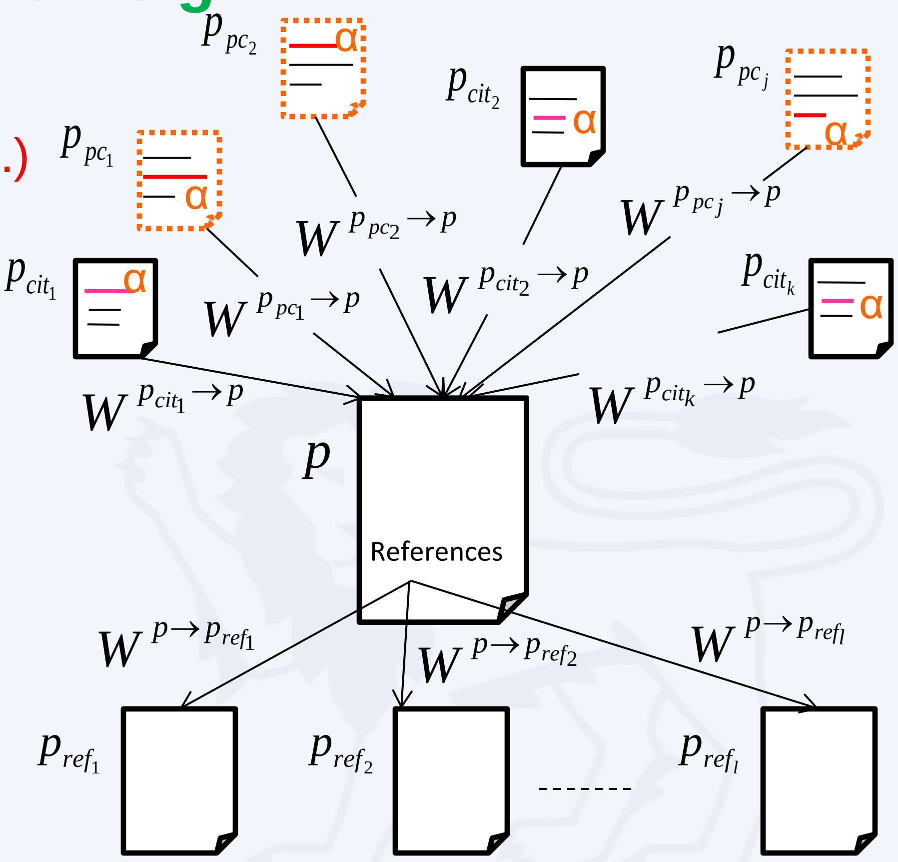
(2) Leveraging Fragments in Potential Citation Papers

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

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

$$
\begin{aligned}
\mathbf{F}^{p} = \; & \alpha \left(\sum_{x=1}^{j} W_{(frg)}^{p_{pc_x} \to p} f_{(frg)}^{p_{pc_x}} + \sum_{y=1}^{k} W_{(frg)}^{p_{cit_y} \to p} f_{(frg)}^{p_{cit_y}} \right) \\
& + (1-\alpha) \left(f^{p} + \sum_{x=1}^{j} W^{p_{pc_x} \to p} f^{p_{pc_x}} \right. \\
& \quad + \sum_{y=1}^{k} W^{p_{cit_y} \to p} f^{p_{cit_y}} \\
& \quad \left. + \sum_{z=1}^{l} W^{p \to p_{ref_z}} f^{p_{ref_z}} \right)
\end{aligned}
$$

Full text



Experiments

Experimental Data

(to be released soon from
http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html)

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

	Training set	Test set
Number of researchers	25	25
Average number of DBLP papers	10.4	9.6
Average number of relevant papers in our dataset	76.3	74.5
Average number of citations	15.3 (max. 169)	14.4 (max. 145)
Average number of references	15.8 (max. 47)	14.2 (max. 58)

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

	Training set	Test set
Number of papers	**50,176**	**50,175**
Average number of citations	19.4 (max. 175)	16.5 (max. 158)
Average number of references	15.7 (max. 45)	15.4 (max. 53)

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

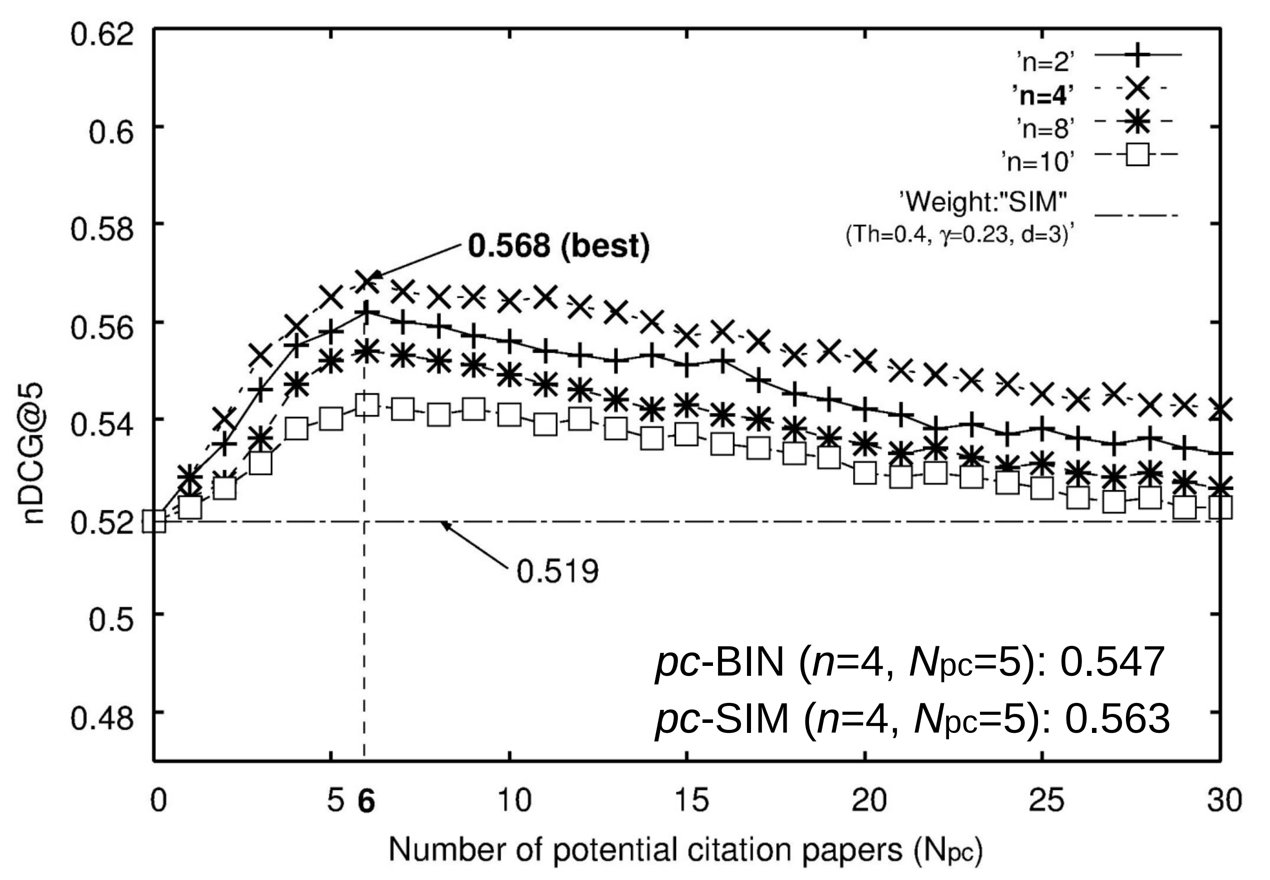
(3) Applying optimized parameters to test set

(1) Leveraging Potential Citation Papers

[Tune:*pc*]

Parameter tuning to discover potential citation papers

nDCG@5 [pc-IMP]



(2) Leveraging Fragments in Potential Citation Papers [Tune:*frg*-SIM]

Fragments with cosine similarity weighting

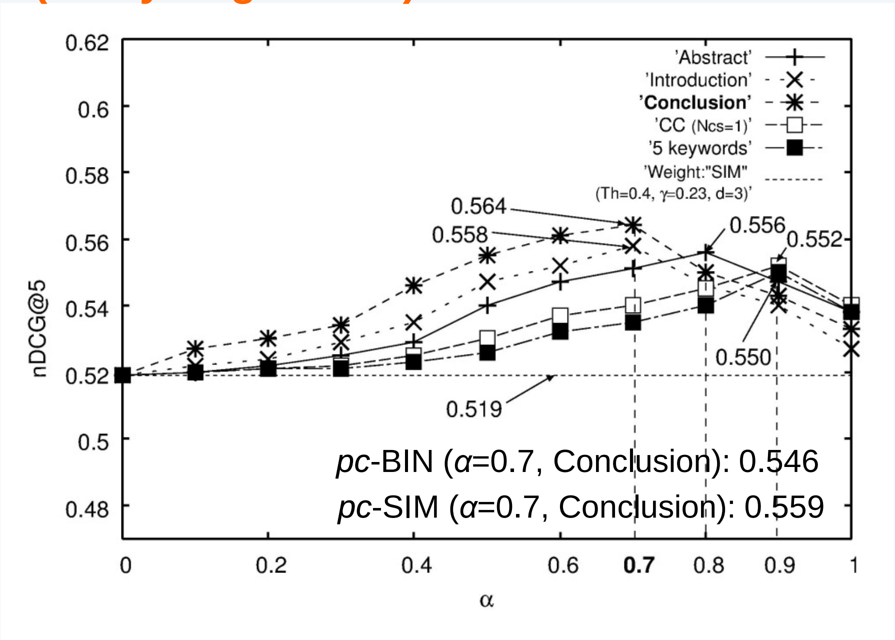
***pc*-IMP (*n*=4, *Npc*=6)**	***n*DCG@5**
Abstract	0.535
Introduction	0.538
Conclusion	0.543
Full text	0.568
Full text + Abstract	0.570
Full text + Introduction	0.569
Full text + Conclusion	**0.574**
pc-BIN (*n*=4, N_{pc}=5), Full text + Conclusion	0.558
pc-SIM (*n*=4, N_{pc}=5), Full text + Conclusion	0.569

(2) Leveraging Fragments in Potential Citation Papers

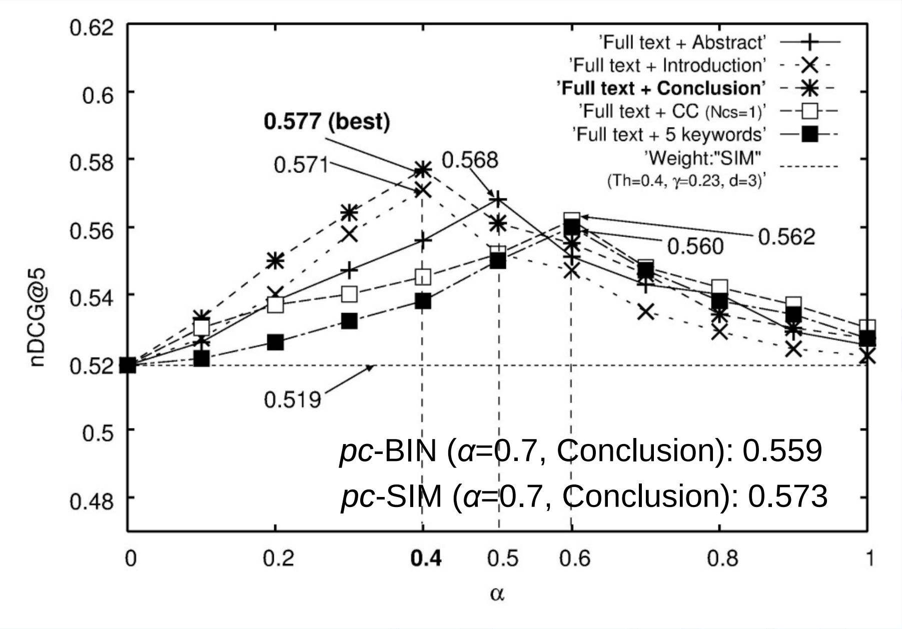
[Tune:*frg*-TW]

[*frg*-SIM] with tunable weight

nDCG@5 [*pc*-IMP]
("Only fragments")



nDCG@5 [*pc*-IMP]
("Full text and fragments")



(3) Applying Optimized Parameters to Test Set

	*n*DCG@5	MRR
***pc*-IMP** (*n*=4, *Npc*=6) *frg*-SIM (Full text + Conclusion) ***frg*-TW** (α=0.4, Full text + Conclusion)	 0.572 **0.579**	 0.787 **0.793**
Baseline system [Sugiyama and Kan, JCDL'10] (Weight "SIM," Th=0.4,γ=0.23,d=3)	0.525	0.751
[Nascimento et al., JCDL'11] ("Frequency of bi-gram" obtained from title and abstract)	0.336	0.438
[Wang and Blei., KDD'11] ("In-matrix prediction" in collaborative topic regression)	0.393	0.495

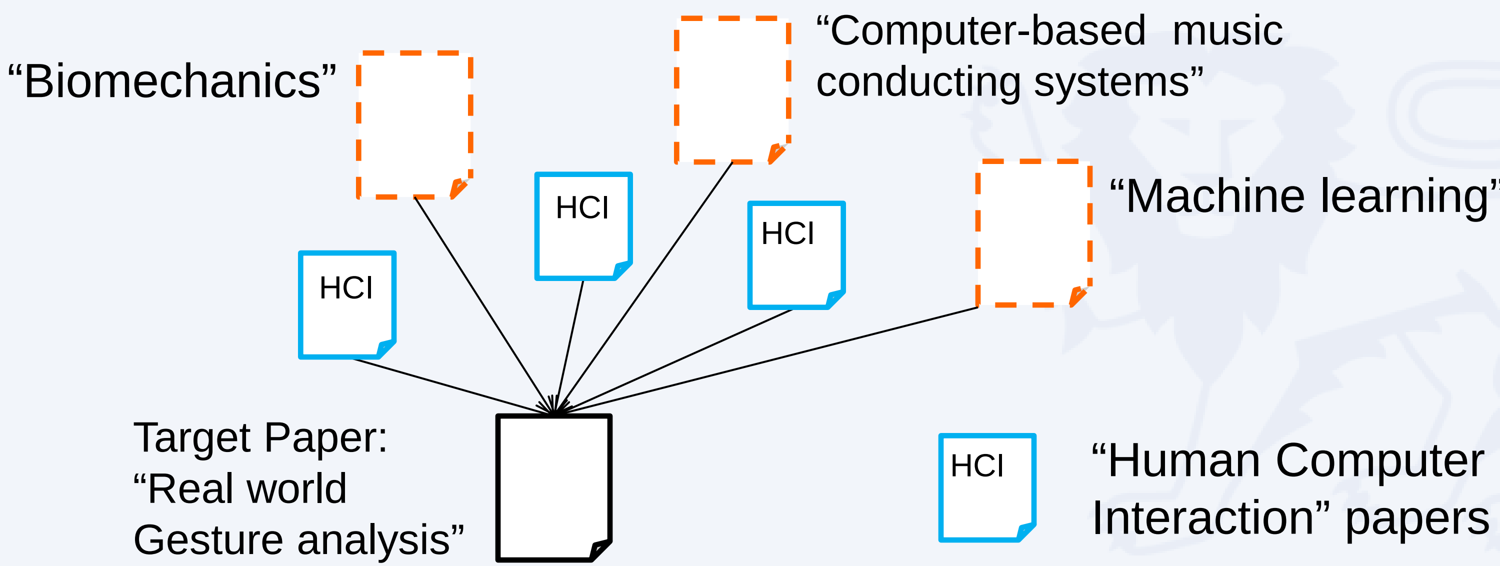
Microscopic Analysis

- **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**



Limitations

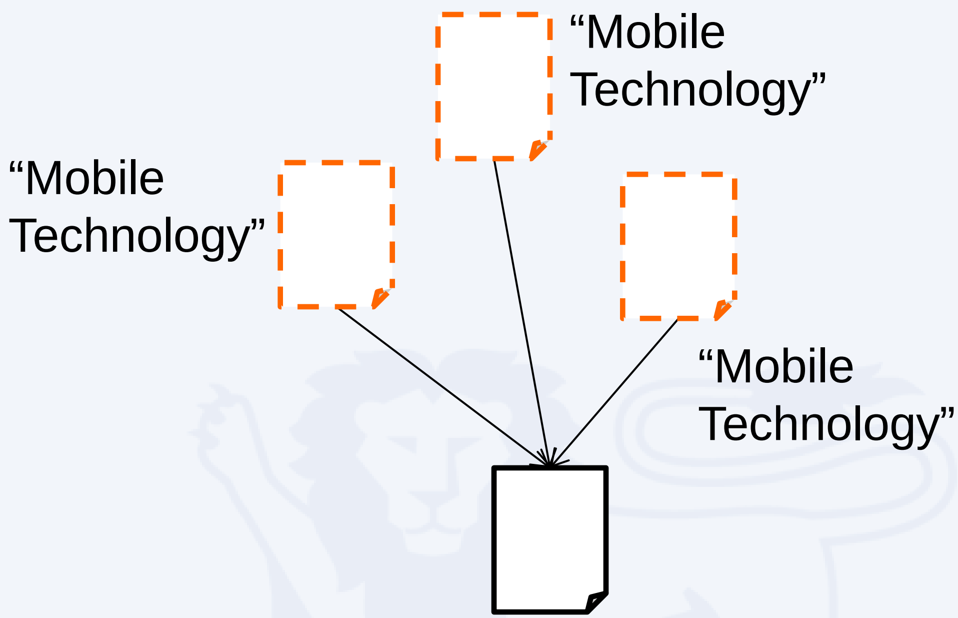
Target paper

"Understanding mobile user's behavior"

- Mobile technology
- User search behavior
- Clustering

Interdisciplinary paper

Identified Potential Citation Papers



Target paper:
"Understanding mobile user's behavior"

Conclusion

To recommend scholarly papers much more relevant to a researcher's interests, we have proposed:

(1) How to identify potential citation papers

(2) How to leverage fragments in citation papers

to characterize candidate papers to recommend.

(1) → Collaborative filtering with imputed matrix

(2) → "Full text + Conclusion" with tunable weight

Future Work

We plan to develop methods for:

- **Selecting balanced neighborhood for interdisciplinary target papers in collaborative filtering**

- **Developing a novel term weighting scheme suitable for small-sized text fragments in papers**

***Thank you very much*!**