Scholarly Paper Recommendation via User's Recent Research Interests


Kazunari Sugiyama,

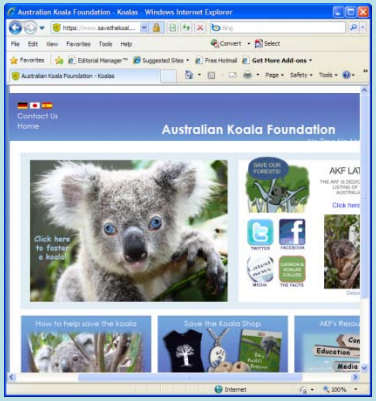
Min-Yen Kan


National University of Singapore

Introduction

Digital Contents

Documents



Scholarly Paper Recommendation via User's Recent Research Interests*

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

Min-Yen Kan
National University of Singapore
Computing 1, 13 Computing Drive,
Singapore 117417
kanmy@comp.nus.edu.sg

ABSTRACT

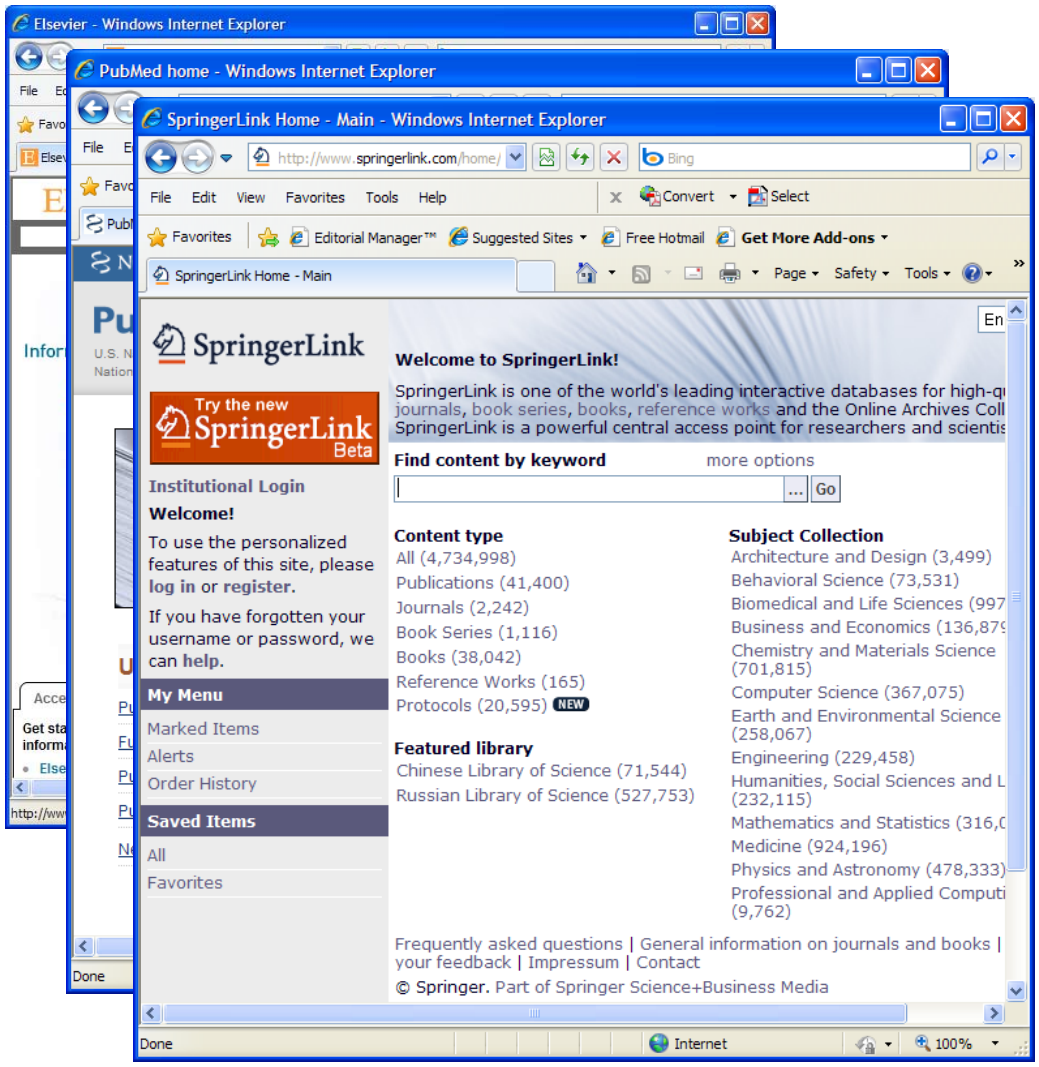
We examine the effect of modeling a researcher's past works in recommending scholarly papers to the researcher. Our hypothesis is that an author's published works constitute a clean signal of the latent interests of a researcher. A key part of our model is to enhance the profile derived directly from past works with information coming from the past works' referenced papers as well as papers that cite the work. In our experiments, we differentiate between junior researchers that have only published one paper and senior researchers that have multiple publications. We show that filtering these sources of information is advantageous – when we additionally prune noisy citations, referenced papers and publication history, we achieve statistically significant higher levels of recommendation accuracy.

To alleviate these problems, past researchers have focused their attention on finding better ranking algorithms for paper search. In particular, the PageRank algorithm [24] has been employed [34, 18, 30] to induce a better global ranking of search results. A problem with this approach is that it does not induce better rankings that are personalized for the specific interests of the user.

To address this issue, digital libraries such as Elsevier[1], PubMed[2], SpringerLink[3] all have systems that can send out email alerts or provide RSS feeds on paper recommendations that match user interests. These systems make the DL more proactive, sending out matched articles in a timely fashion. Unfortunately, these require the user to state their interests explicitly, either in terms of categories or as saved searches, and take up valuable time on the part of the user to set up.

"Information Overload"

Digital Library



- **Email alerts**
- **RSS feeds**

Users are required to inputs their interests explicitly.

Introduction

Our aim

- **To provide recommendation of papers by using latent information about each user's research interests**
  - Historical and current publication lists

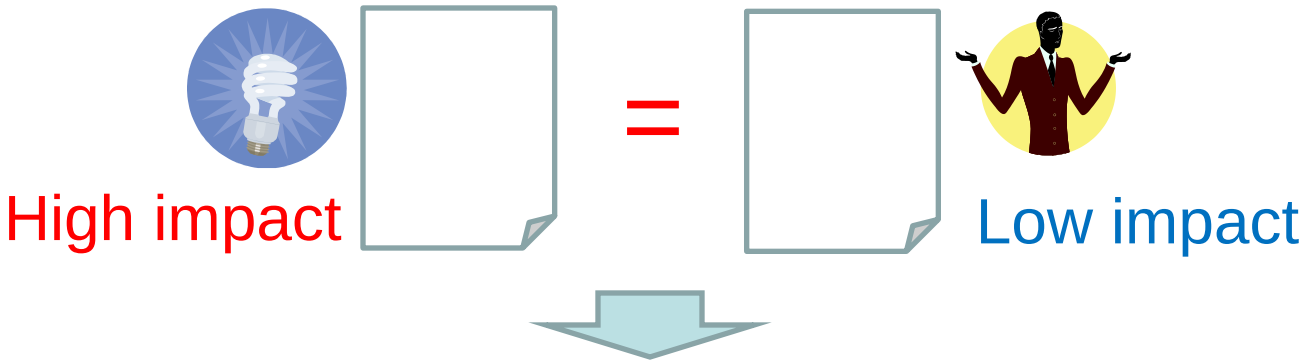
Users are not required to input their interests explicitly.

Related Work

Improving Ranking in Digital Library

- **Ranking Search Results**

  ISI impact factor [Garfield, ‘79]



Recent works introduce PageRank to weight and control for the impact of papers

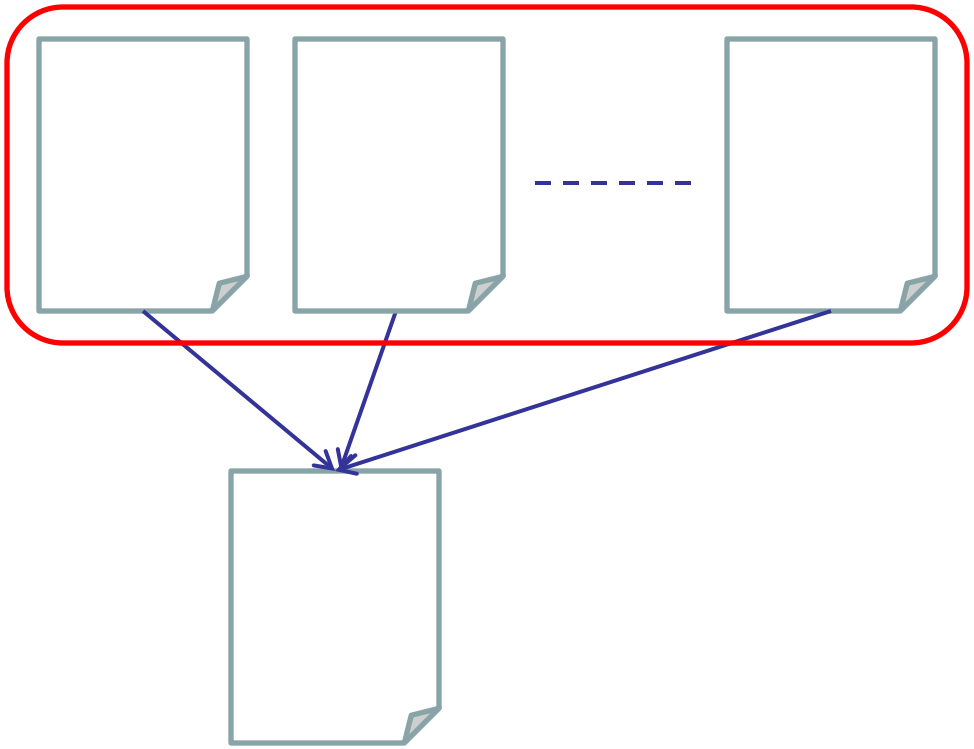
[Sun and Giles, ECIR’07]
[Krapivin and Marchese, ICADL’08],
[Sayyadi and Getoor, SIAM Data Mining, ‘09]

Related Work

Improving Ranking in Digital Library

- **Measuring the Importance of Scholarly Papers**

  ISI impact factor [Garfield, '79]



Popularity biased

PageRank also controls the popularity bias

[Bollen et al., Scientometrics'06]
[Chen et al., Informetrics'07]

Related Work

Recommendation in Scholarly Digital Libraries

- **Collaborative Filtering Approach**

  [McNee et al., CSCW'02]: Focuses on citation network of papers

  [Yang et al., JCDL'09]: Ranking-oriented collaborative filtering

- **Hybrid Approach of Collaborative Filtering and Content-based Filtering**

  [Torres et al., JCDL'04]: Many users satisfied with the recommended papers

- **PageRank-based Approach**

  [Gori and Pucci, WI'06]: Focuses on graph structure of papers

Related Work

Robust User Profile Construction in Recommendation Systems

- **Web Search Results**

  [Teevan et al., SIGIR'05]: Visited Web pages and emails history

  [White et al., SIGIR'09]: A small number of Web pages preceding the current browsing page

- **Dynamic Content such as News**

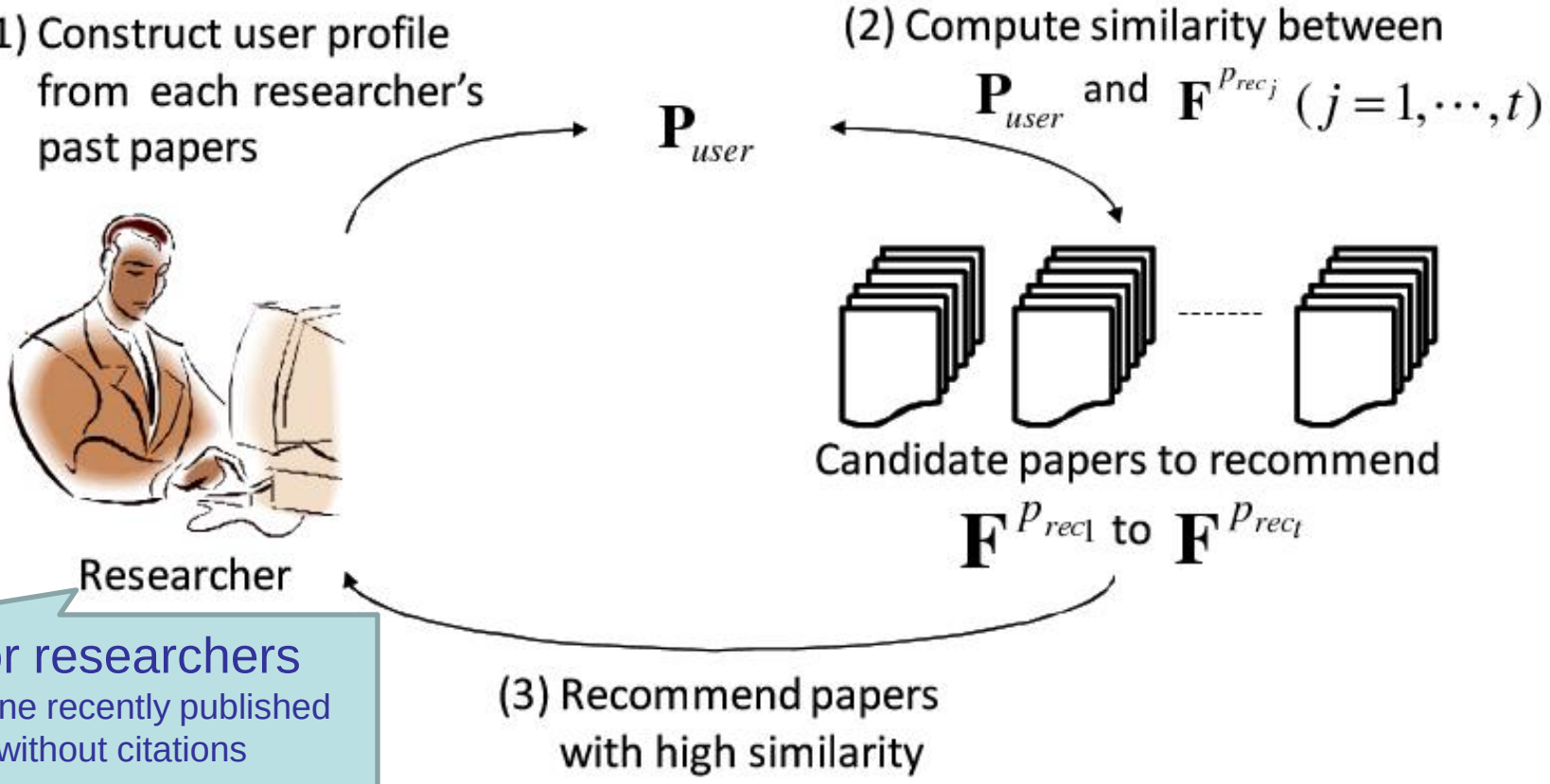
  [Shen et al., SIGIR'05]
  [Tan et al., KDD'06]
  } Kullback-Leibler divergence is used to represent a user's information need

  [Chu and Park, WWW'09]: Use demographics and interaction data

- **Abstracts of Scholarly Papers**

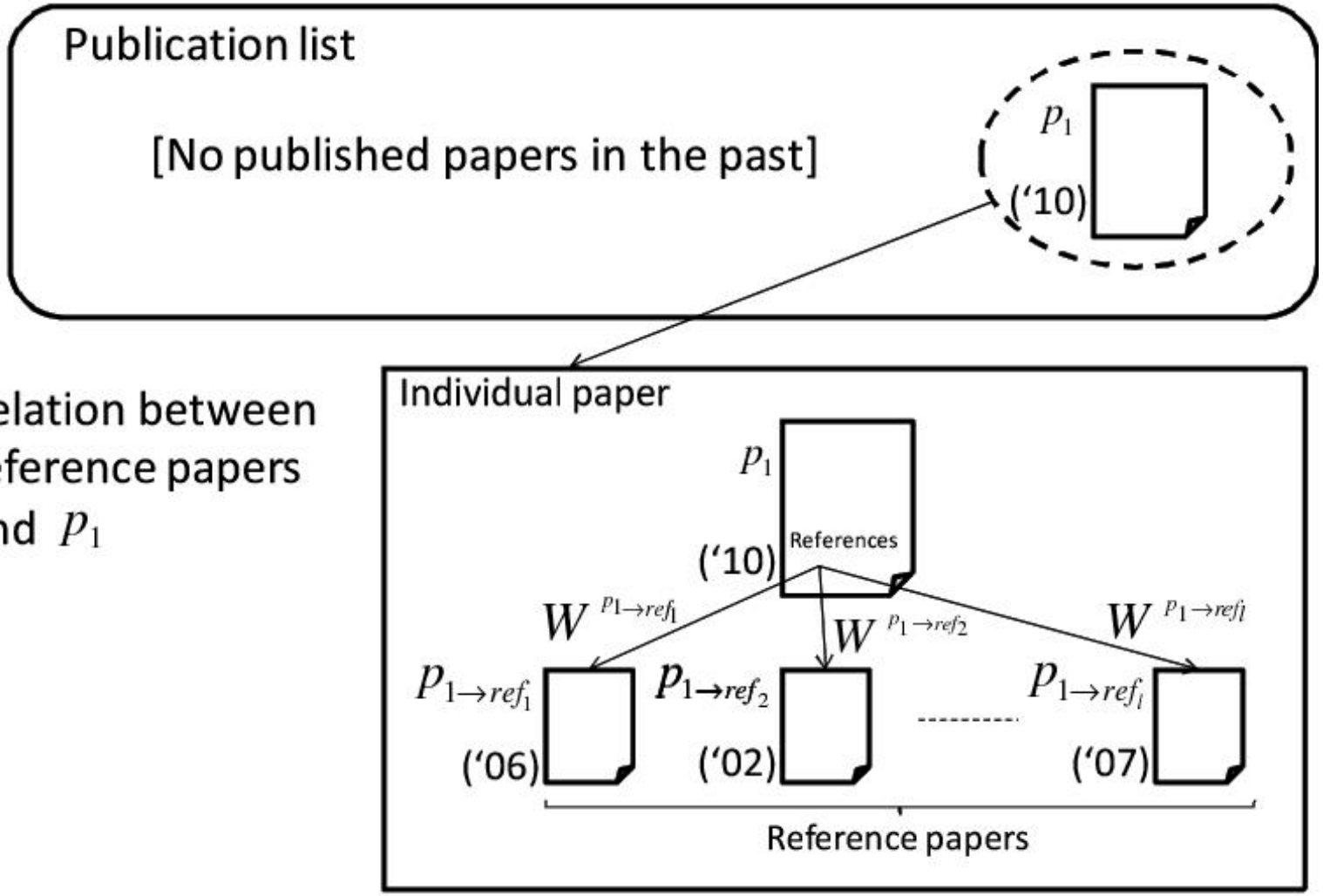
  [Kim et al., ICADL'08]: Frequent patterns from click-history and term weight

Proposed Method

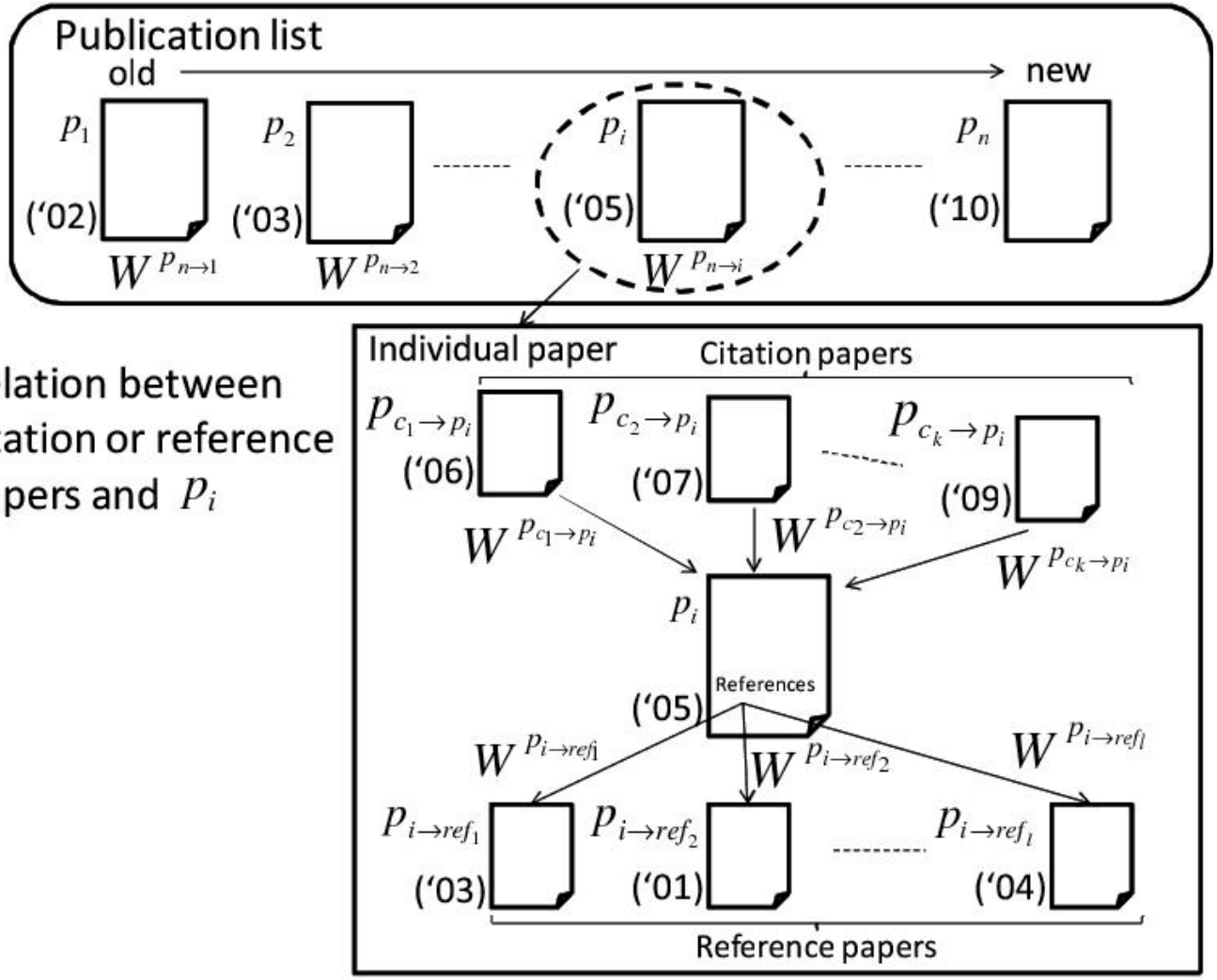


- Junior researchers
  Only one recently published paper without citations
- Senior researchers
  Multiple published papers with citation papers

User Profile Construction (Junior Researchers)



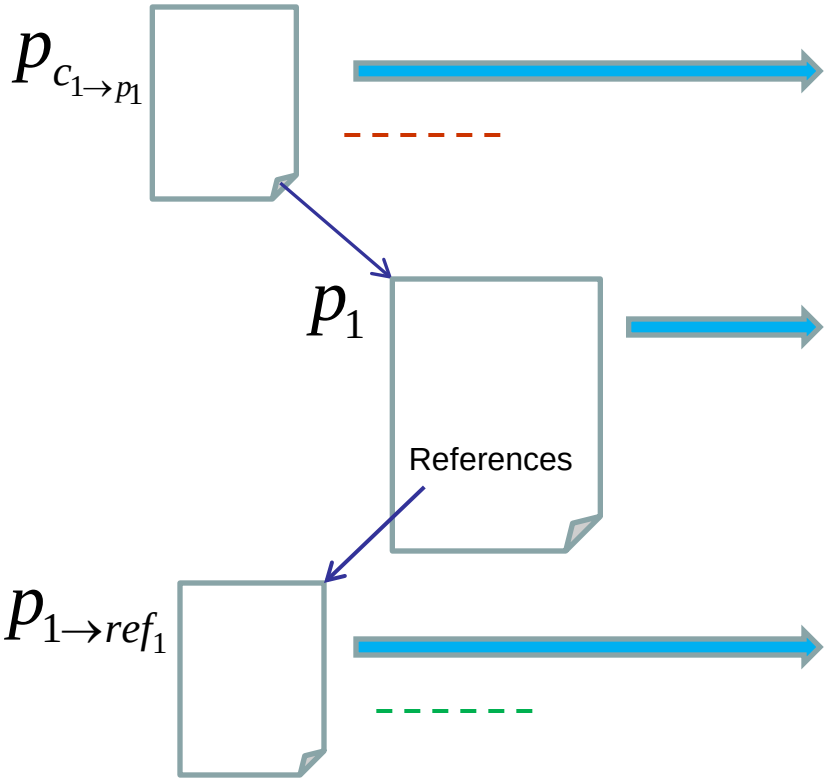
User Profile Construction (Senior Researchers)



Relation between citation or reference papers and p_i

Linear Combination

Weighting Scheme (LC)



(…, digital, library, recommendation, …)

$\mathbf{f}^{p_{c_{1\to p_1}}} = (\ldots, 0.25, 0.13, 0.47, \ldots)$

(…, digital, library, recommendation, …)

$\mathbf{f}^{p_1} = (\ldots, 0.53, 0.38, 0.62, \ldots)$

(…, digital, library, recommendation, …)

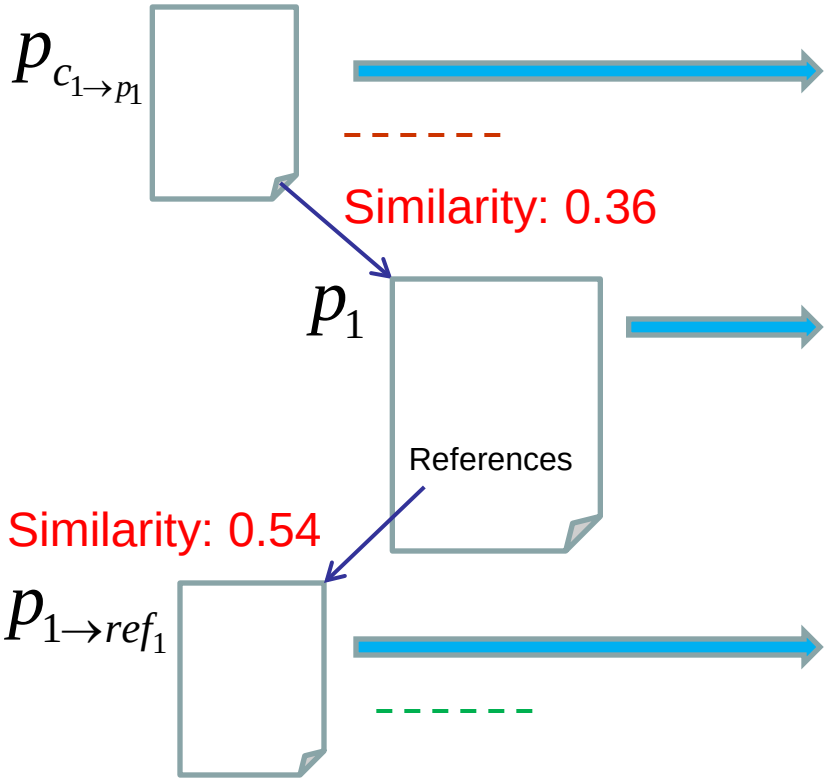
$\mathbf{f}^{p_{1\to ref_1}} = (\ldots, 0.61, 0.72, 0, \ldots)$

$$\mathbf{F}^{p_1} = \mathbf{f}^{p_1} + 1 \times \mathbf{f}^{p_{c\to p_1}} + \cdots\cdots$$
$$+ 1 \times \mathbf{f}^{p_{1\to ref_1}} + \cdots\cdots$$

$$\mathbf{P}_{user} = \mathbf{F}^{p_1} + \mathbf{F}^{p_2} + \cdots$$

Cosine Similarity

Weighting Scheme (SIM)



(…, digital, library, recommendation, …)

$\mathbf{f}^{p_{c_1 \to p_1}} = (\ldots, 0.25, 0.13, 0.47, \ldots)$

(…, digital, library, recommendation, …)

$\mathbf{f}^{p_1} = (\ldots, 0.53, 0.38, 0.62, \ldots)$

(…, digital, library, recommendation, …)

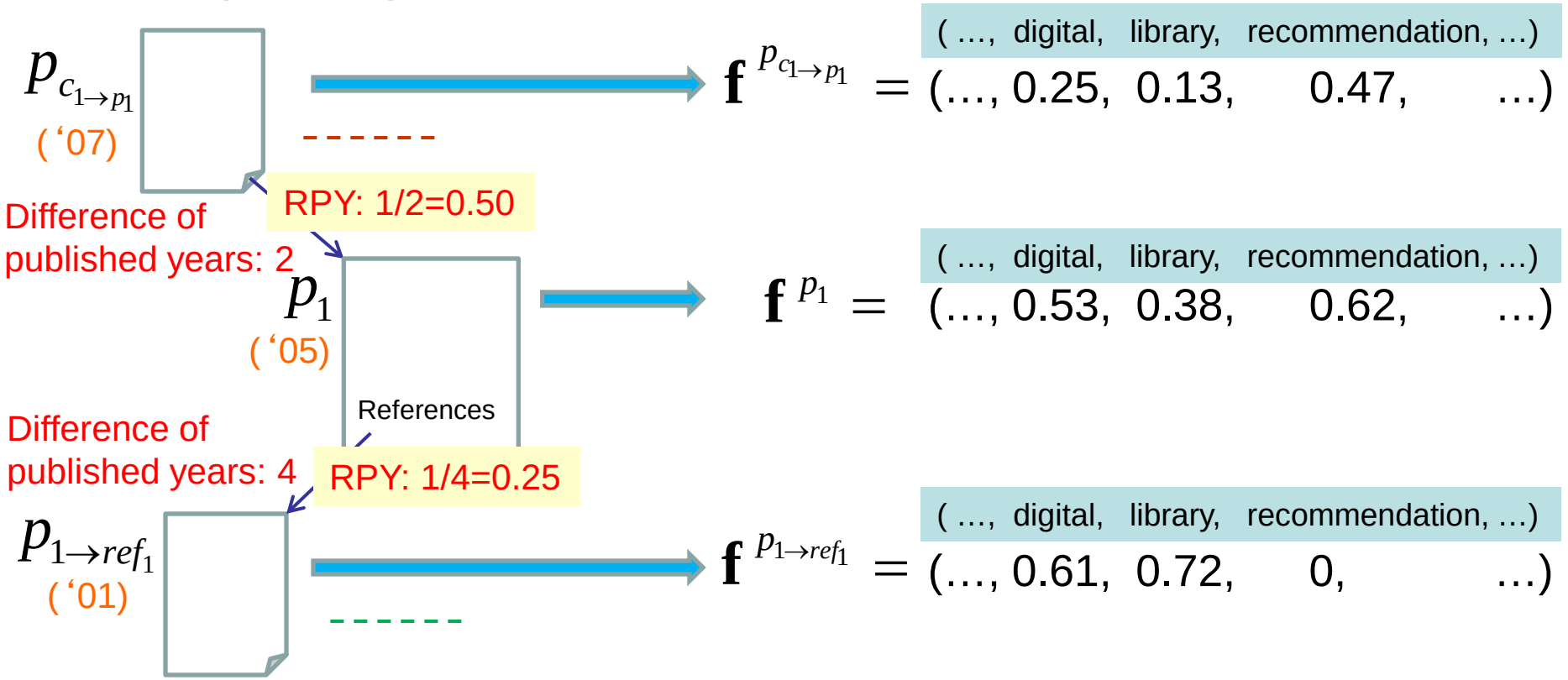
$\mathbf{f}^{p_{1 \to ref_1}} = (\ldots, 0.61, 0.72, 0, \ldots)$

$$\mathbf{F}^{p_1} = \mathbf{f}^{p_1} + 0.36 \times \mathbf{f}^{p_{c \to p_1}} + \cdots\cdots + 0.54 \times \mathbf{f}^{p_{1 \to ref_1}} + \cdots\cdots$$

$$\mathbf{P}_{user} = \mathbf{F}^{p_1} + \mathbf{F}^{p_2} + \cdots$$

Reciprocal of the Difference Between Published Years

Weighting Scheme (RPY)

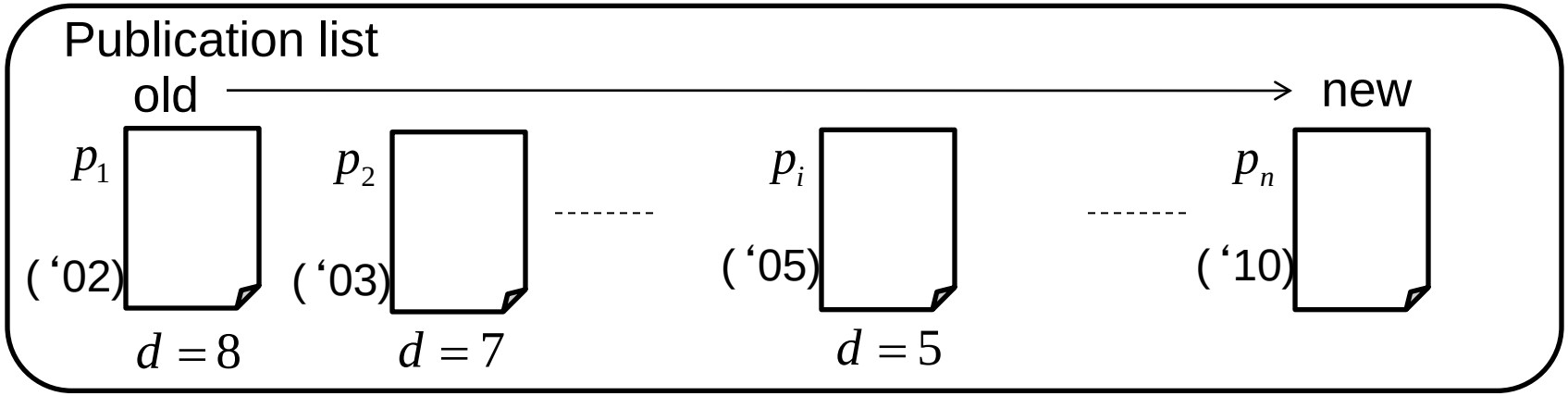


$$\mathbf{F}^{p_1} = \mathbf{f}^{p_1} + 0.50 \times \mathbf{f}^{p_{c\to p_1}} + \cdots\cdots$$
$$+ 0.25 \times \mathbf{f}^{p_{1\to ref_1}} + \cdots\cdots$$

$$\mathbf{P}_{user} = \mathbf{F}^{p_1} + \mathbf{F}^{p_2} + \cdots$$

Forgetting Factor

Weighting Scheme (FF, senior researchers only)



$$W^{p_{n \to z}} = e^{-\gamma \times d} \quad [\gamma : \text{forgetting coefficient } (0 \le \gamma \le 1)]$$

(e.g., $\gamma = 0.2$)

$$W^{p_{n \to p_i}} = e^{-0.2 \times 5}$$

$$\vdots$$

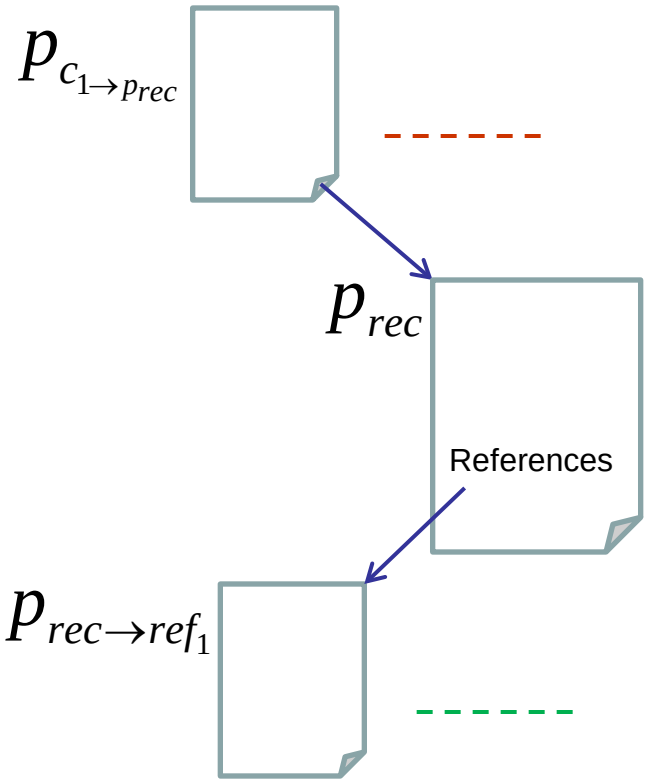
$$W^{p_{n \to p_2}} = e^{-0.2 \times 7}$$

$$W^{p_{n \to p_1}} = e^{-0.2 \times 8}$$

$$\mathbf{P}_{user} = \mathbf{F}^{p_n} + \cdots + e^{-0.2 \times 5} \cdot \mathbf{F}^{p_i} + \cdots + e^{-0.2 \times 7} \cdot \mathbf{F}^{p_2} + e^{-0.2 \times 8} \cdot \mathbf{F}^{p_1}$$

(2) Feature Vector Construction for Candidate Papers

- **Basically, TF-IDF**
- **Also use information about citation and reference papers**



$$\mathbf{F}^{p_{rec}} = \mathbf{f}^{p_{rec}} + W^{p_{c_1 \to p_{rec}}} \cdot \mathbf{f}^{p_{c_1 \to p_{rec}}} + \cdots$$
$$+ W^{p_{rec \to ref_1}} \cdot \mathbf{f}^{p_{rec \to ref_1}} + \cdots$$

Weighting scheme

$$W^{p_{1 \to ref_i}} \ (i = 1, \cdots, l)$$

- LC
- SIM
- RPY

(3) Recommendation of Papers

- **Compute cosine similarity**

$$sim(\mathbf{P}_{user}, \mathbf{F}^{p_{rec}}) = \frac{\mathbf{P}_{user} \cdot \mathbf{F}^{p_{rec}}}{|\mathbf{P}_{user}| \cdot |\mathbf{F}^{p_{rec}}|}$$

$\mathbf{P}_{user}$: User profile

$\mathbf{F}^{p_{rec}}$: Feature vector for candidate paper to recommend

- **Then, recommend the top *n* papers to the user**
  - *n*=5,10

Experiments

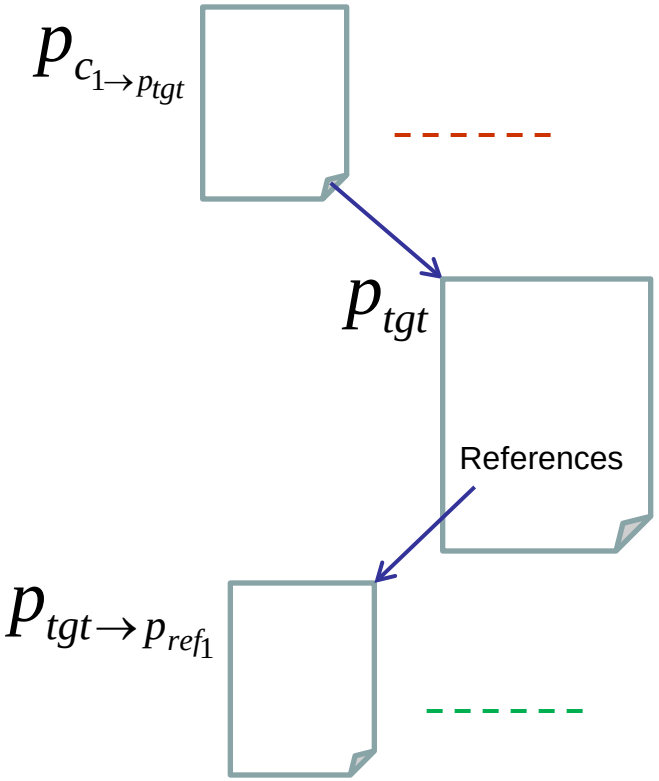
Experimental Data

- **Researchers**

Natural Language Processing
Information Retrieval

	Junior researchers	Senior researchers
Number of subjects	15	13
Average number of DBLP papers	1.0	9.5
Average number of relevant papers in ACL'00 – '06	28.6	38.7
Average number of citation papers	0	10.5 (max. 199)
Average number of reference papers	18.7 (max. 29)	19.4 (max.79)

Experiments



Experimental Data

- **Candidate Papers to Recommend**
  - **ACL Anthology Reference Corpus**
  - **[Bird et al., LREC'08]**

Information about
citation and reference papers

$$p_{tgt} <= p_{c_{1\to p_{tgt}}}$$
$$p_{tgt\to p_{ref_1}} <= p_{tgt}$$

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

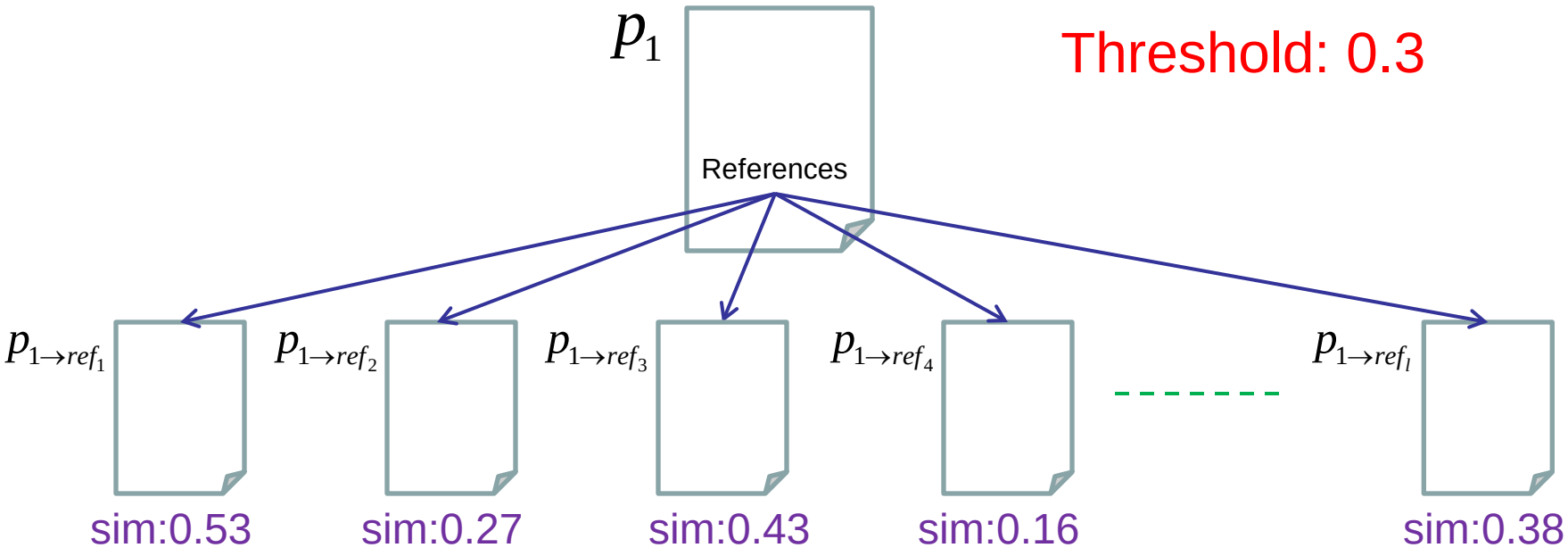
Junior Researchers

The most recent paper only

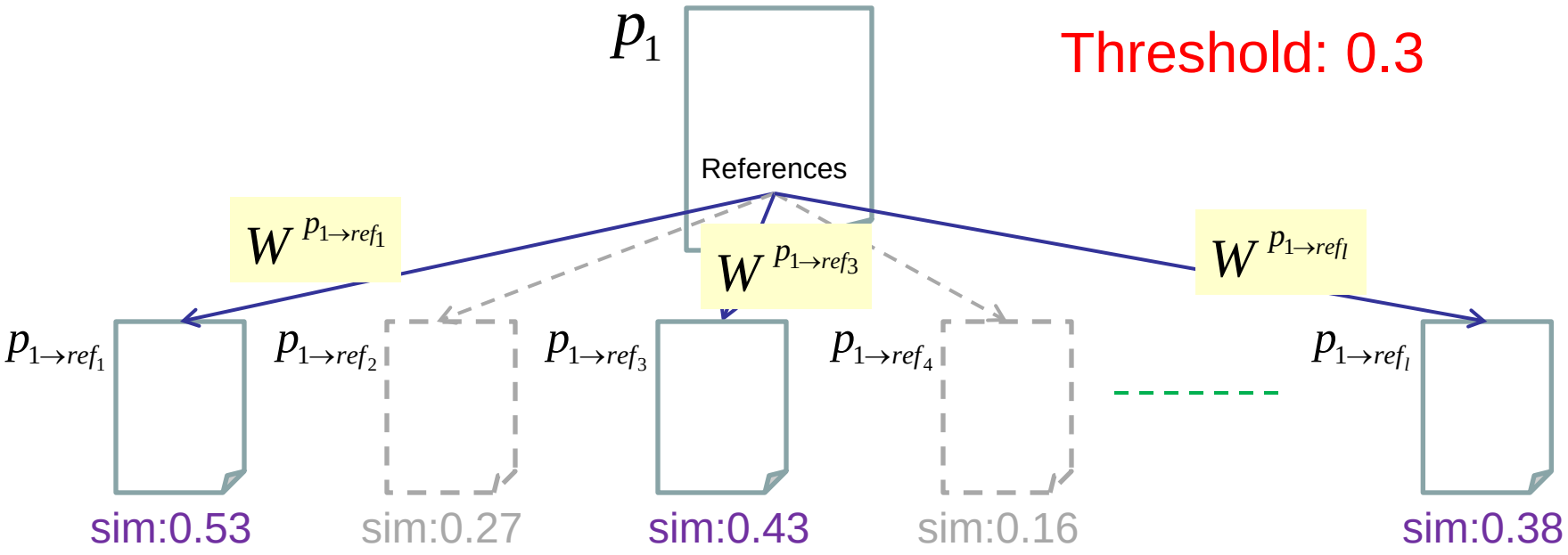
NDCG@5		The most recent paper in user profile (MP)					
		Weight "**LC**"		Weight "**SIM**"		Weight "**RPY**"	
		MP	MP+R	MP	MP+R	MP	MP+R
ACL papers to recommend (AP)	AP	0.382	0.442	0.382	0.443	0.382	0.431
	AP+C	0.388	0.429	0.390	0.435	0.389	0.438
	AP+R	0.402	0.405	0.427	0.451	0.404	0.440
	AP+C+R	0.418	0.445	0.435	**0.457**	0.423	0.452

MRR		The most recent paper in user profile (MP)					
		Weight "**LC**"		Weight "**SIM**"		Weight "**RPY**"	
		MP	MP+R	MP	MP+R	MP	MP+R
ACL papers to recommend (AP)	AP	0.455	0.505	0.455	0.522	0.455	0.520
	AP+C	0.450	0.477	0.452	0.525	0.448	0.489
	AP+R	0.453	0.494	0.490	0.524	0.469	0.492
	AP+C+R	0.472	0.538	0.521	**0.568**	0.515	0.526

Is Pruning of Reference Papers Effective?



Is Pruning of Reference Papers Effective?



Weighting scheme

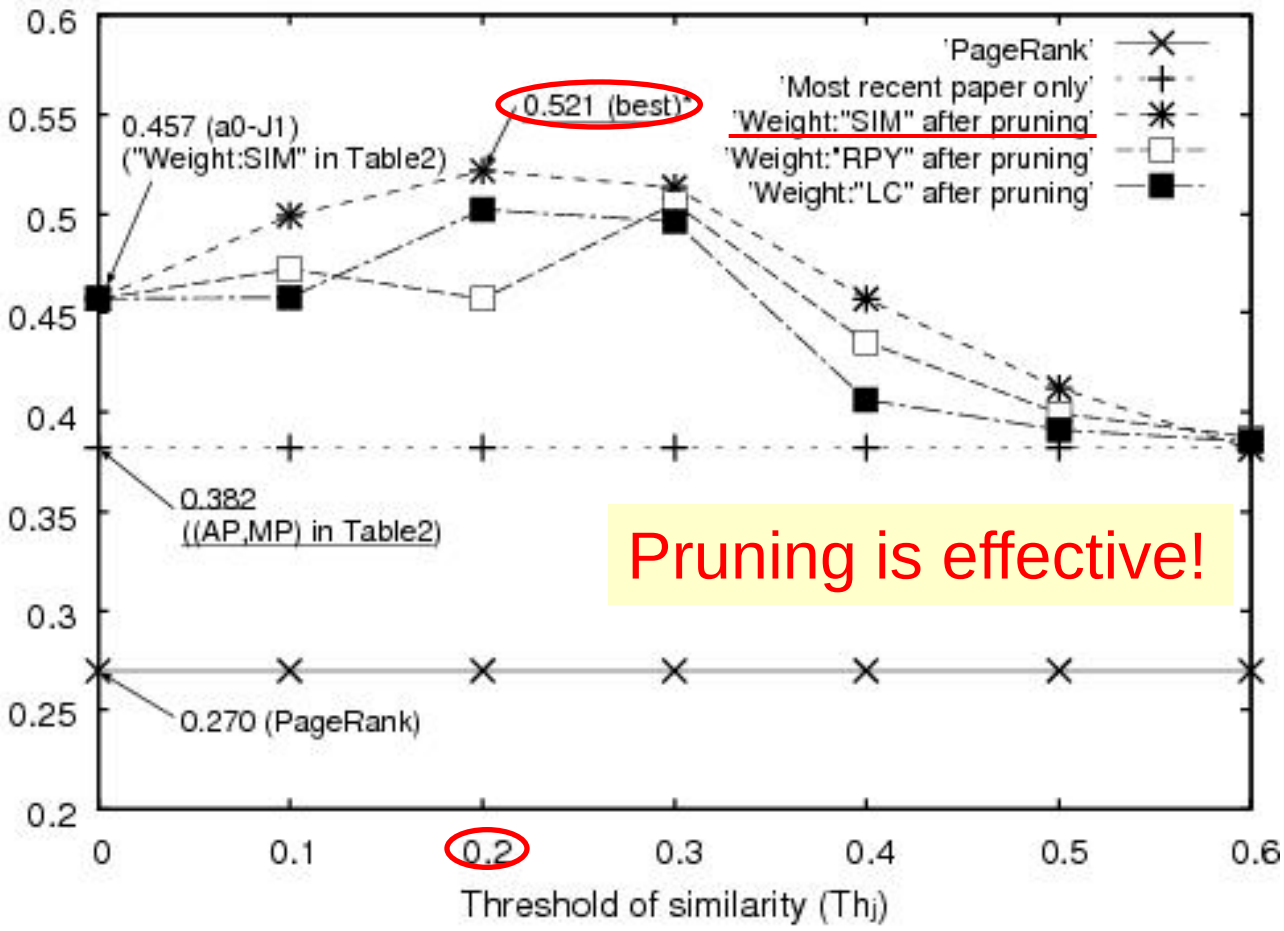
$W^{p_{1\to ref_i}}\ (i=1,\cdots,l)$

- LC
- SIM
- RPY

Junior Researchers

The most recent paper with pruning its reference papers

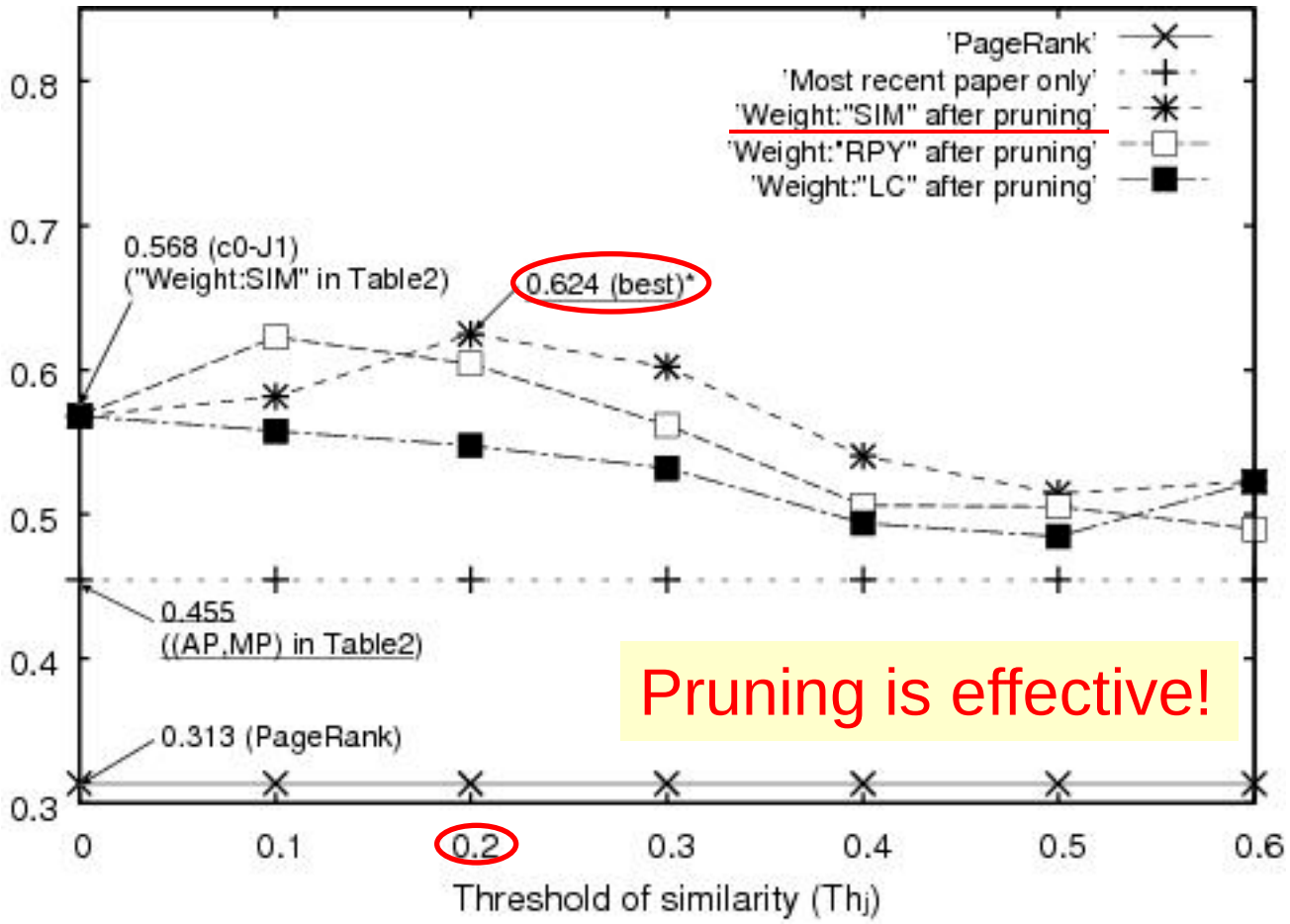
[NDCG@5]



Junior Researchers

The most recent paper with pruning its reference papers

[MRR]



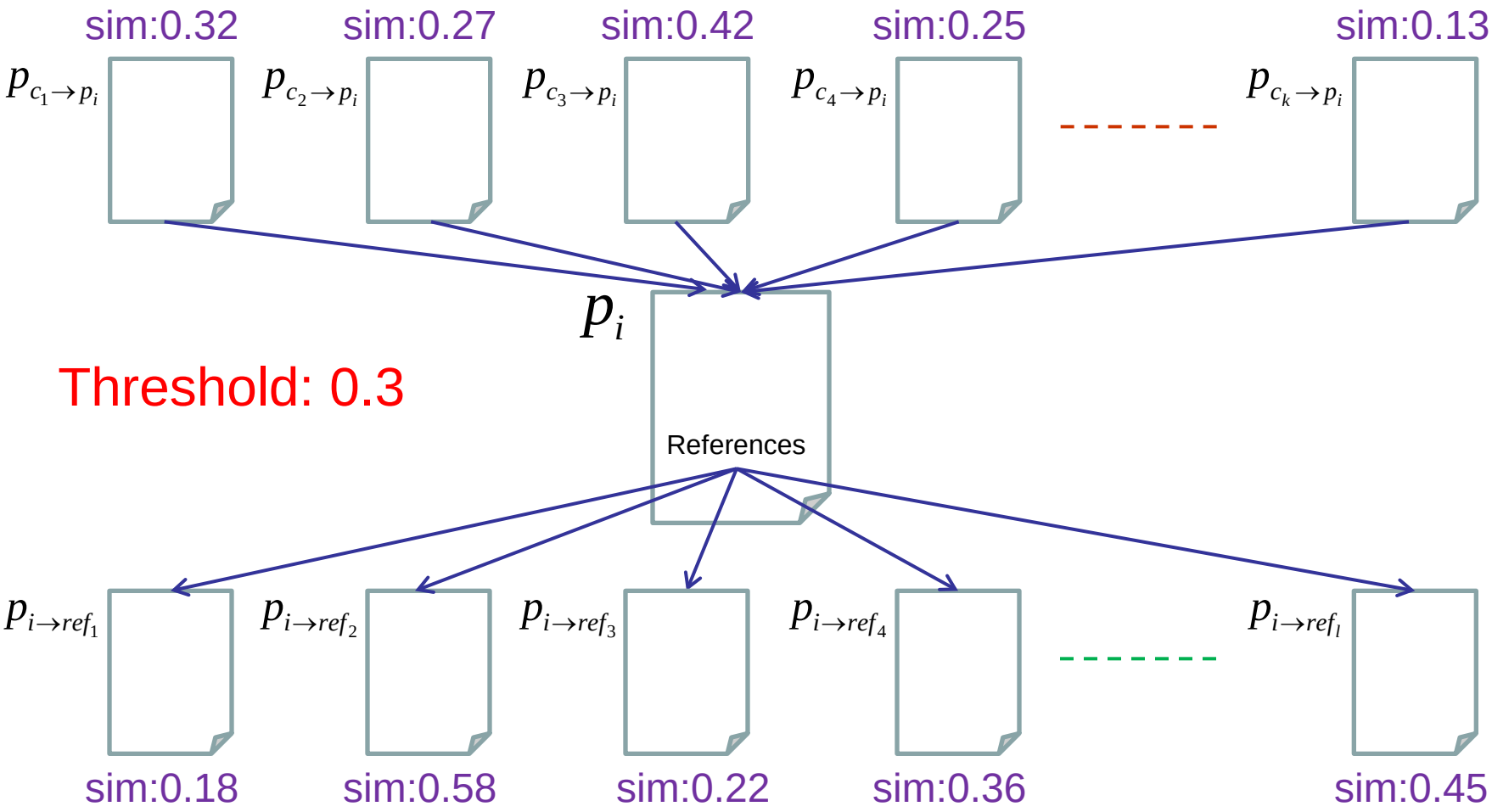
Senior Researchers

The most recent paper only

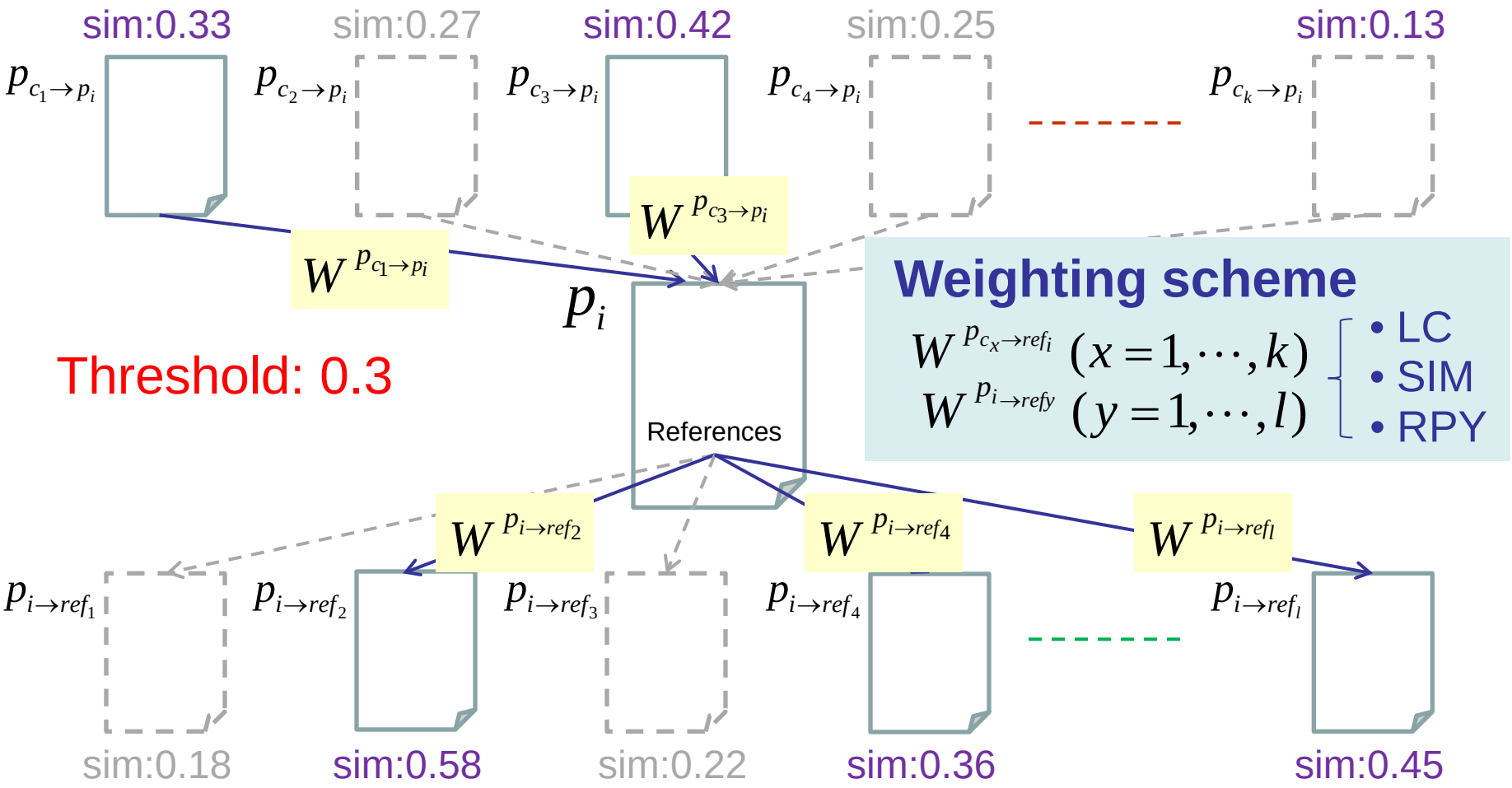
NDCG@5		The most recent paper in user profile (MP)											
		Weight: "**LC**"				Weight: "**SIM**"				Weight: "**RPY**"			
		MP	MP+C	MP+R	MP +C+R	MP	MP+C	MP+R	MP +C+R	MP	MP+C	MP+R	MP +C+R
ACL papers to recommend (AP)	AP+R	0.325	0.334	0.390	0.401	0.325	0.351	0.406	0.401	0.325	0.338	0.395	0.401
	AP+C	0.332	0.341	0.378	0.384	0.335	0.383	0.399	0.406	0.334	0.381	0.401	0.404
	AP+R	0.345	0.408	0.353	0.410	0.374	0.373	0.416	0.418	0.348	0.393	0.402	0.408
	AP+C+R	0.367	0.390	0.390	0.417	0.384	0.402	0.419	**0.421**	0.374	0.415	0.413	0.418

MRR		The most recent paper in user profile (MP)											
		Weight: "**LC**"				Weight: "**SIM**"				Weight: "**RPY**"			
		MP	MP+C	MP+R	MP +C+R	MP	MP+C	MP+R	MP +C+R	MP	MP+C	MP+R	MP +C+R
ACL papers to recommend (AP)	AP+R	0.621	0.657	0.670	0.709	0.621	0.696	0.688	0.709	0.621	0.696	0.688	0.709
	AP+C	0.615	0.696	0.688	0.696	0.621	0.696	0.692	0.727	0.615	0.696	0.656	0.696
	AP+R	0.618	0.651	0.659	0.696	0.658	0.657	0.648	0.697	0.637	0.657	0.661	0.657
	AP+C+R	0.637	0.709	0.709	0.710	0.689	0.696	0.728	**0.739**	0.681	0.688	0.696	0.709

Is Pruning of Citation and Reference Papers Effective?



Is Pruning of Citation and Reference Papers Effective?

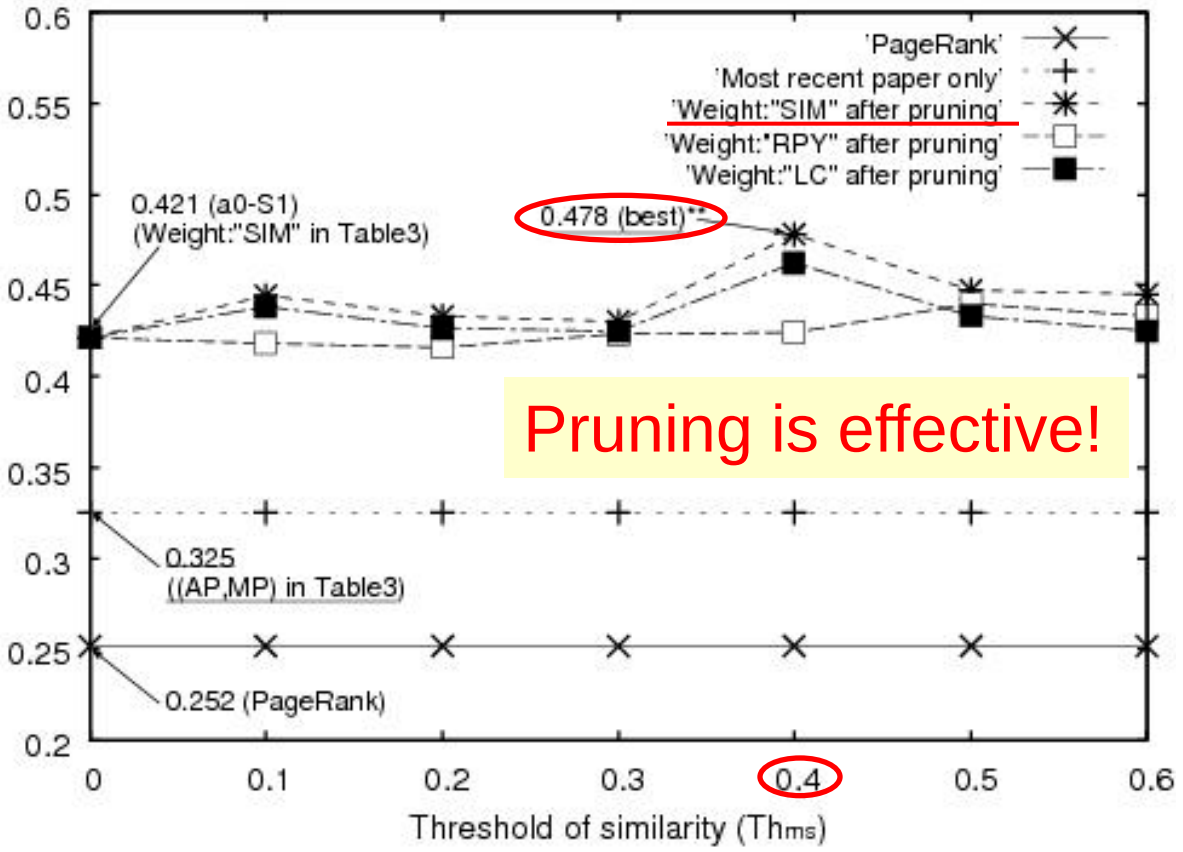


Senior Researchers

The most recent paper with pruning

its citation and reference papers

[NDCG@5]

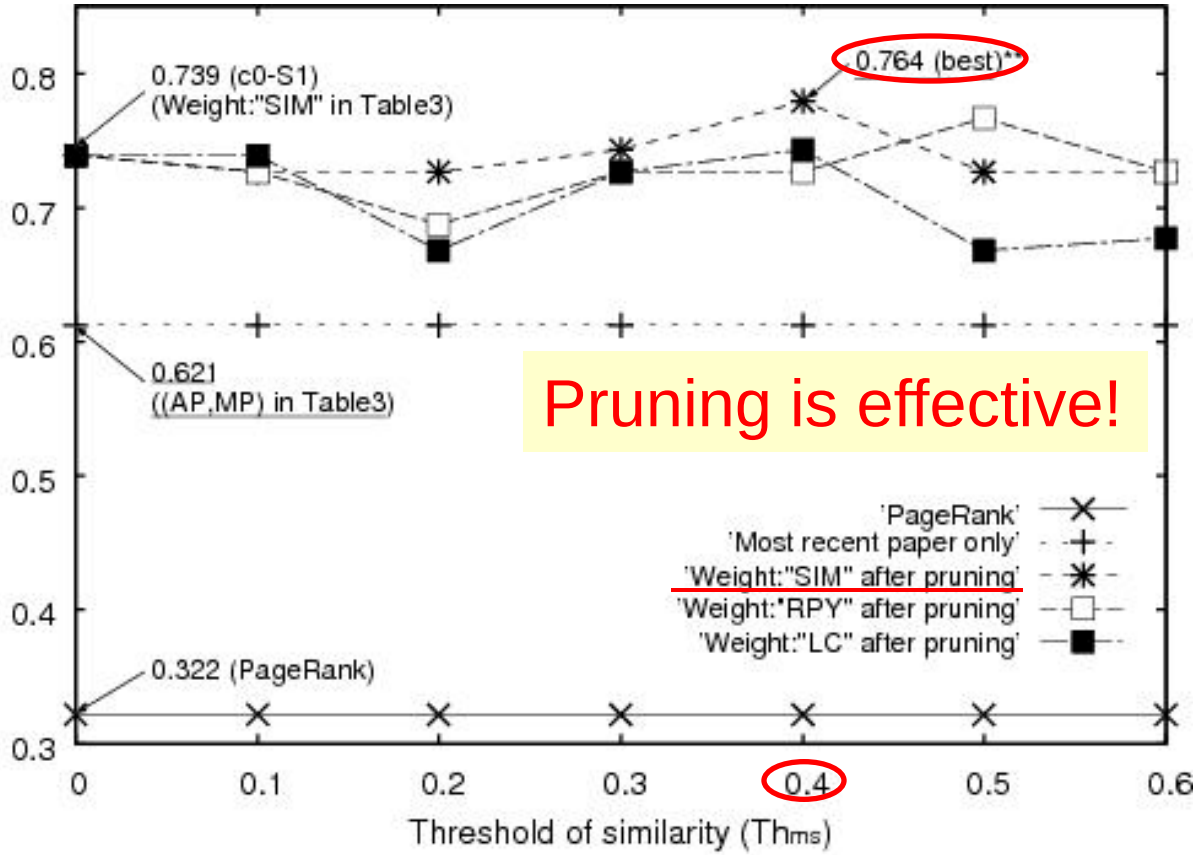


Senior Researchers

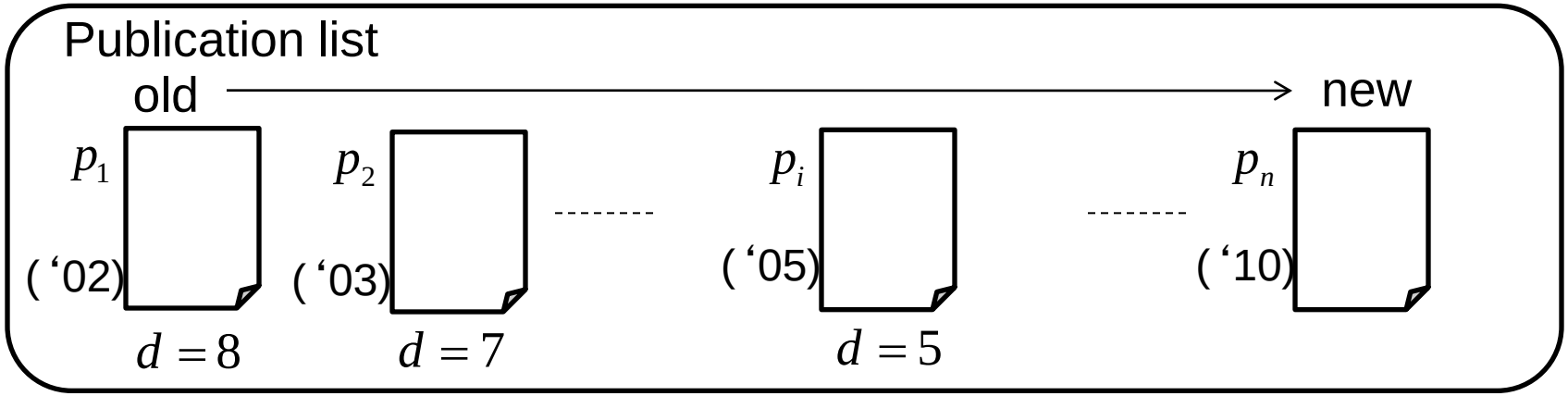
The most recent paper with pruning

its citation and reference papers

[MRR]



Is Forgetting Factor (FF) Effective?



$$W^{p_{n \to z}} = e^{-\gamma \times d} \quad [\gamma : \text{forgetting coefficient } (0 \le \gamma \le 1)]$$

(e.g., $\gamma = 0.2$)

$$W^{p_{n \to p_i}} = e^{-0.2 \times 5}$$

$$\vdots$$

$$W^{p_{n \to p_2}} = e^{-0.2 \times 7}$$

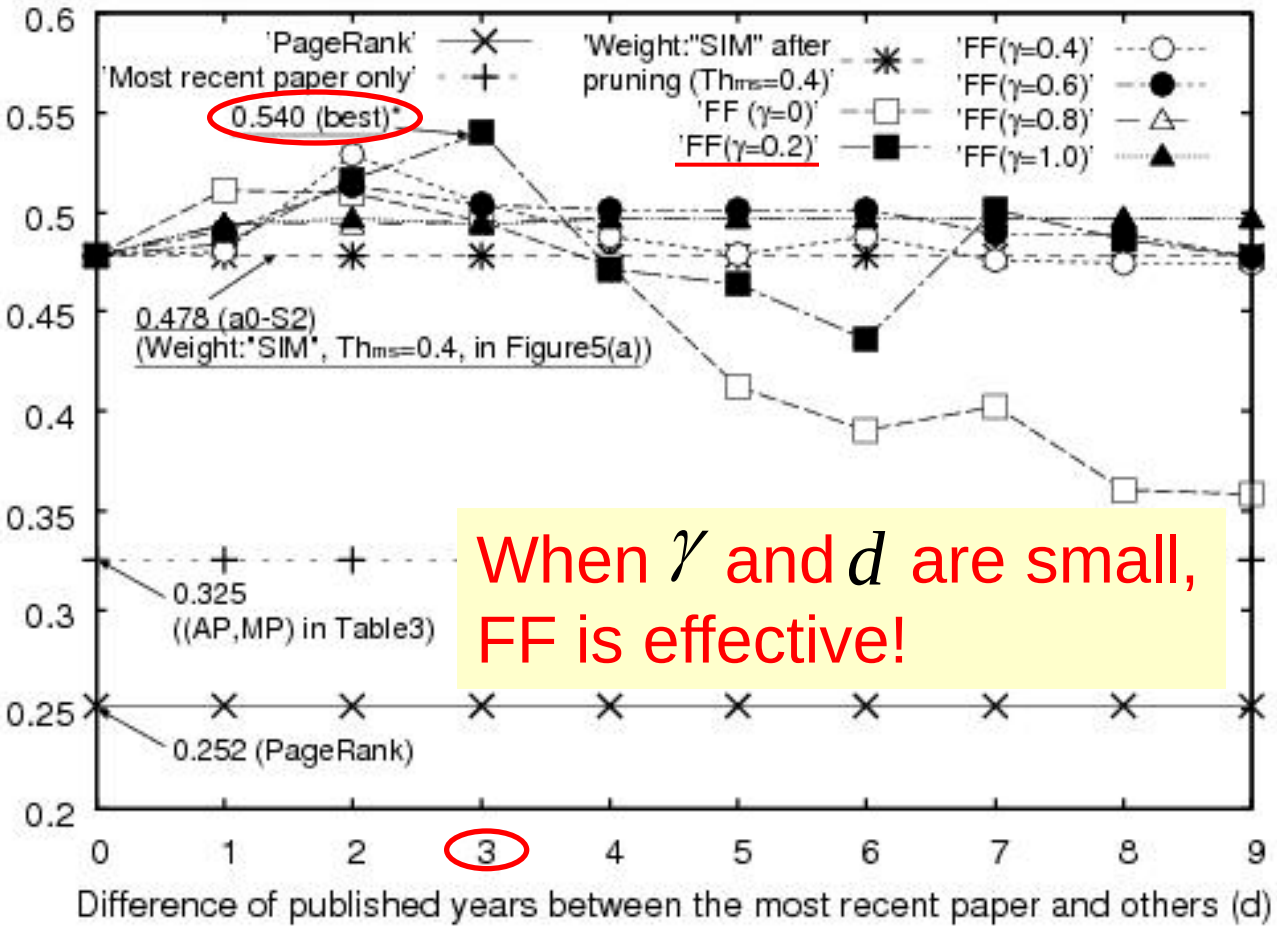
$$W^{p_{n \to p_1}} = e^{-0.2 \times 8}$$

$$\mathbf{P}_{user} = \mathbf{F}^{p_n} + \cdots + e^{-0.2 \times 5} \cdot \mathbf{F}^{p_i} + \cdots + e^{-0.2 \times 7} \cdot \mathbf{F}^{p_2} + e^{-0.2 \times 8} \cdot \mathbf{F}^{p_1}$$

Senior Researchers

Past published papers with forgetting factor

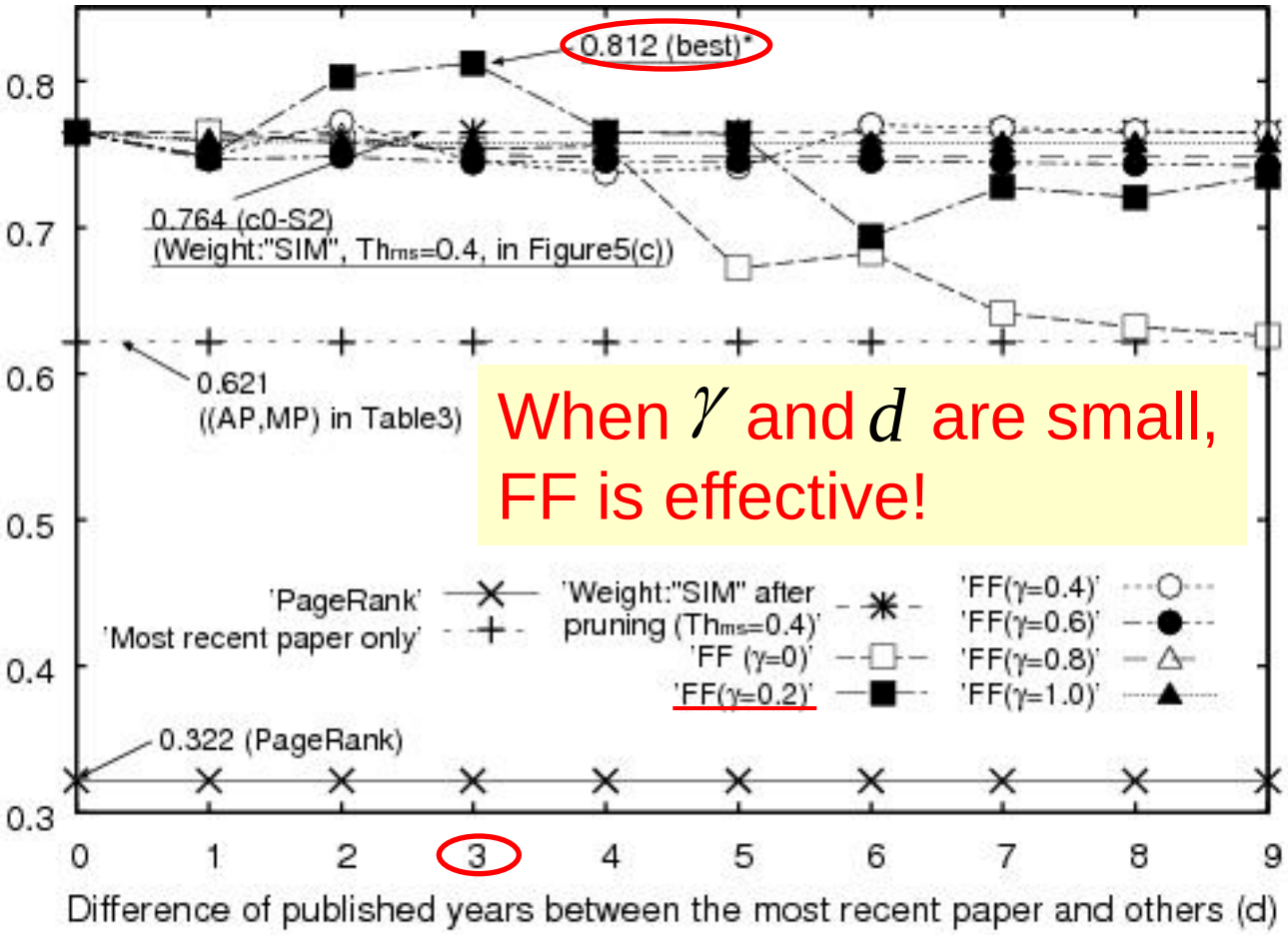
[NDCG@5]



Senior Researchers

Past published papers with forgetting factor

[MRR]



Conclusion

- **Propose a generic model towards recommending scholarly papers relevant to junior and senior researcher's interests**
  - Use past publications to capture the researcher's interests
  - Our user model also incorporates its neighboring papers (citation and reference papers) as context
    - Also employ this scheme to characterize candidate papers to recommend

Conclusion

- **Achieve higher recommendation accuracy**
  - When our model prunes neighboring papers with low similarity (for both junior and senior researchers)
    - This scheme can enhance the signal of the original topic of the paper to recommend and user profile
  - When we construct user profile using past papers within 3 years from the most recent paper (for senior researchers)

Future Work

We plan to develop methods for:

- **Helping recommend interdisciplinary papers that could encourage a push to new frontiers for senior researchers**

- **Recommending papers that are easier to understand to quickly acquire knowledge about intended research for junior researchers**

***Thank you very much*!**