Incorporating Satellite Documents into Co-citation Networks for Scientific Paper Searches

Masaki Eto

Gakushuin Women’s College

Tokyo, Japan

masaki.eto@gakushuin.ac.jp

BIRNDL 2016

Outline of this presentation

1. Background

    Co-citation and network model

    Outline of co-citation network searching

2. Research question

    Satellite documents

3. Proposed Retrieval Method

    Specifying satellite documents

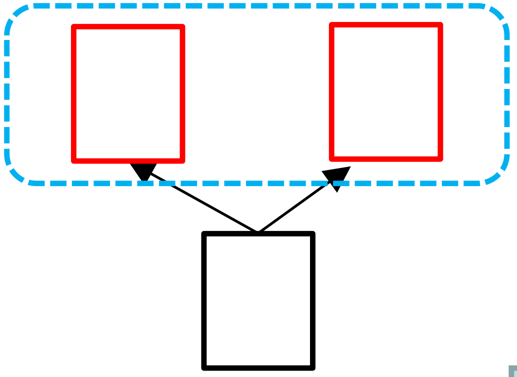
    Incorporating satellite documents

    Ranking documents in the network

4. Experiment

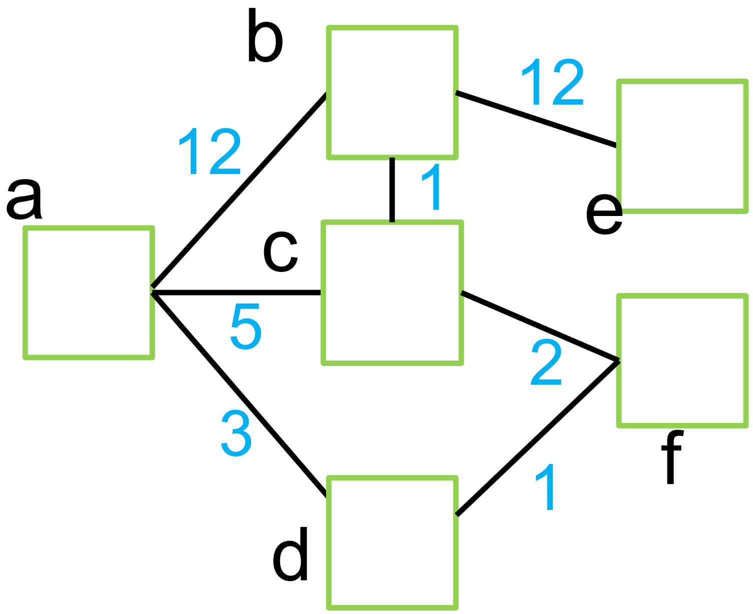
    Evaluating the proposed method

Co-citation Network



Co-citation
=a linkage between a pair of documents concurrently cited by a third document

Network model



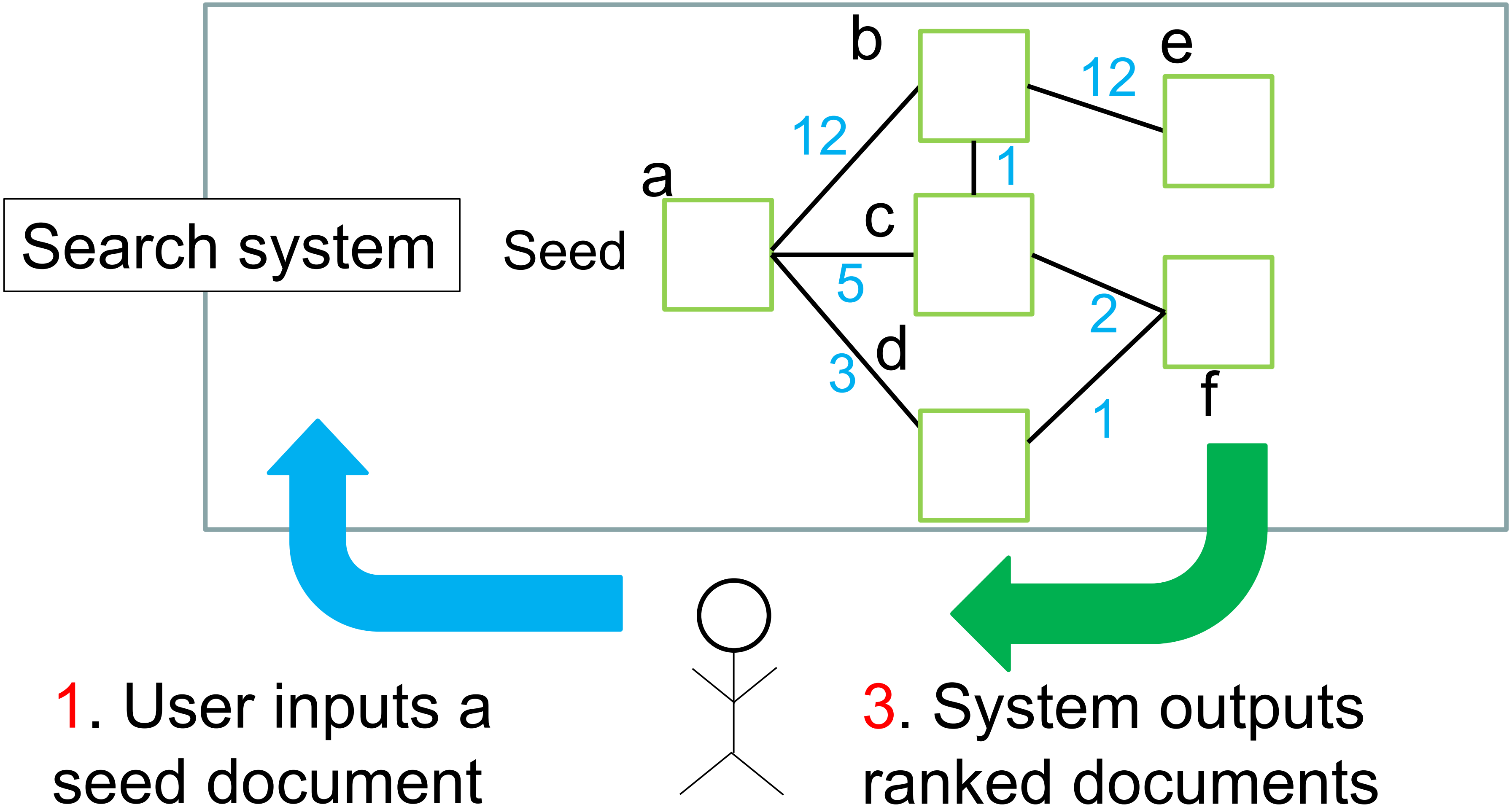
Node = cited document

Edge = co-citation linkage

Weight =
number of co-citing documents

Outline of Co-citation Network Searching

2. System creates a network and ranks the documents in the network



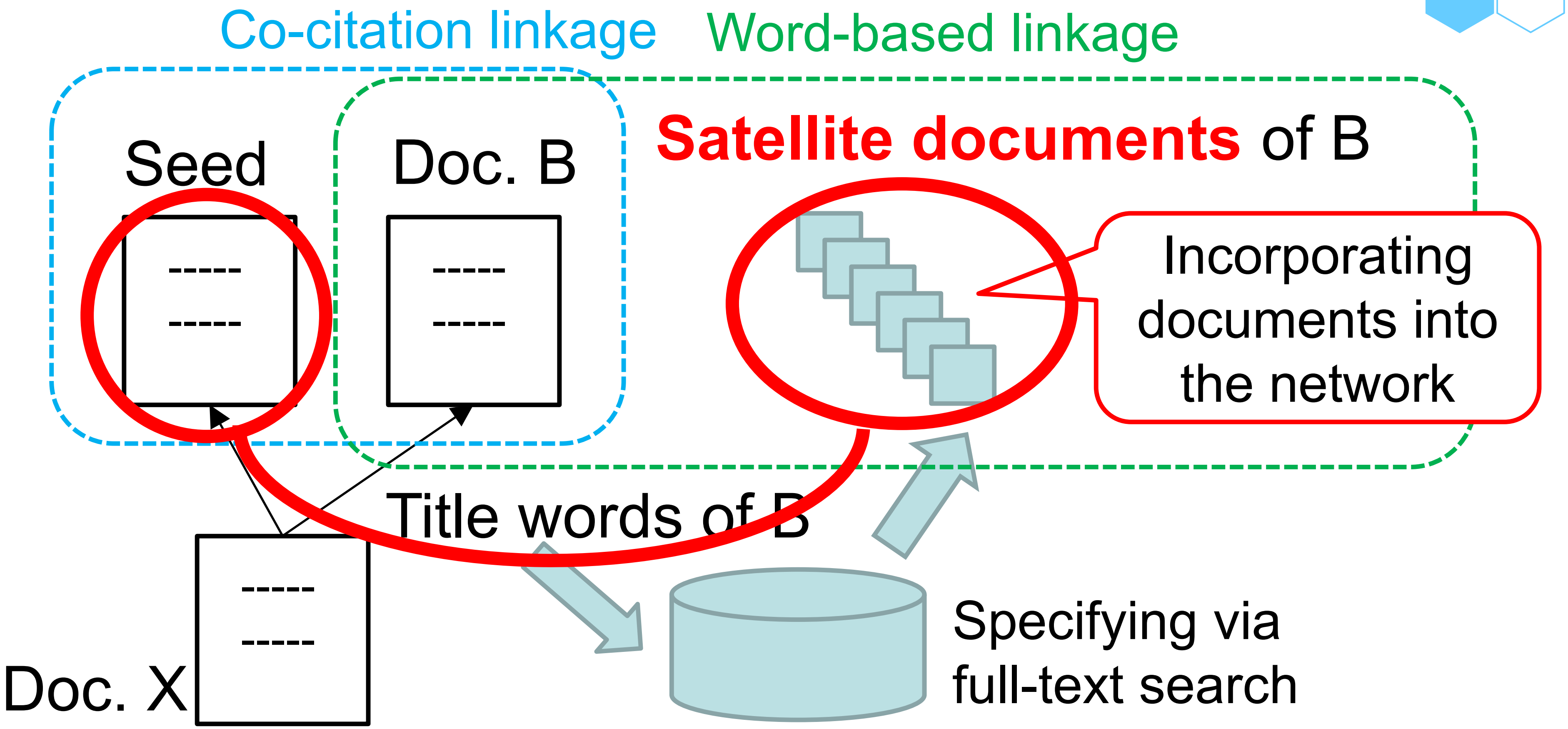
1. User inputs a seed document

3. System outputs ranked documents

Outline of this presentation

1. Background
   - Co-citation and network model
   - Similar document search
2. Research question
   - Satellite documents
3. Proposed Retrieval Method
   - Specifying satellite documents
   - Incorporating satellite documents
   - Ranking documents in the network
4. Experiment
   - Evaluating the proposed method

Enlarging the Co-citation Networks so as to Include New Relevant Documents



Research question

Do satellite documents have relevant linkages to the seed that are not identified by co-citation linkages?

Outline of this presentation

1. Background

    Co-citation and network model

    Similar document search

2. Research question

    Satellite documents

3. Proposed Retrieval Method

    Specifying satellite documents

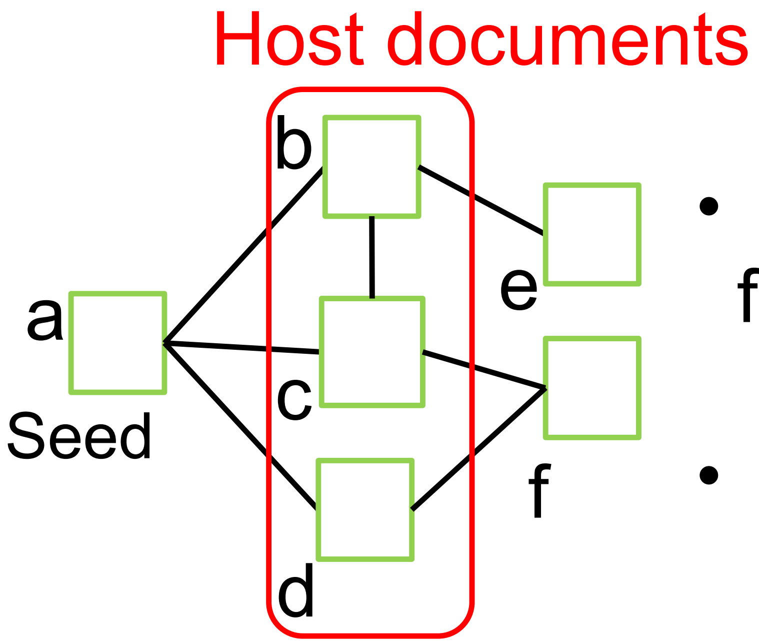
    Incorporating satellite documents

    Ranking documents in the network

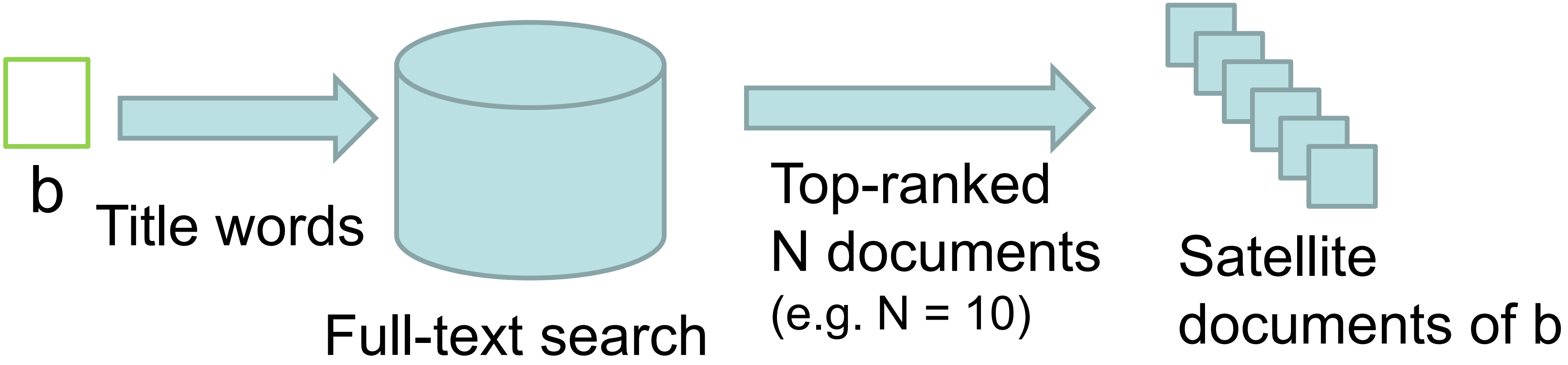
4. Experiment

    Evaluating the proposed method

Specifying Satellite Documents



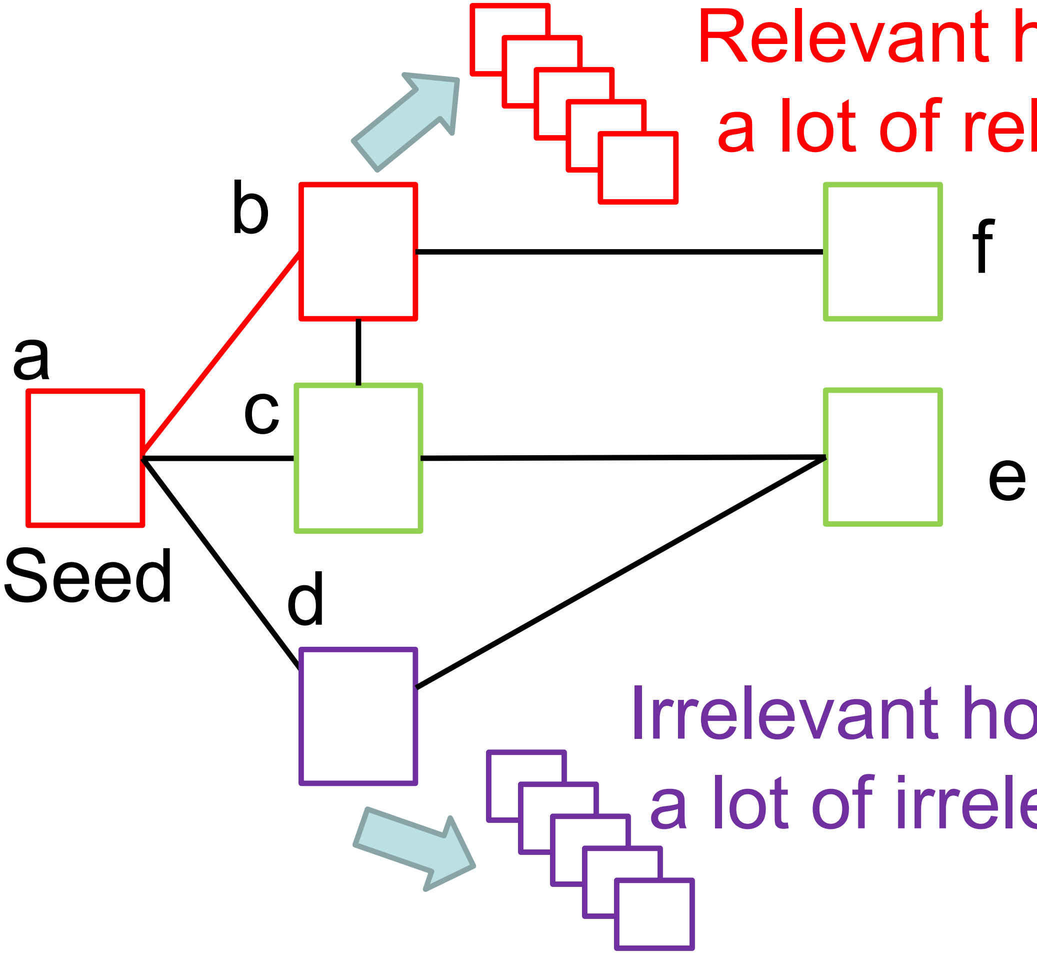
- Host documents are sources for specifying satellite documents
- Each host document is one hop from the seed



Tf-idf (Indri Search Engine by Lemure project)

Problem of Satellite Documents

Not all co-citation linkages are relevant

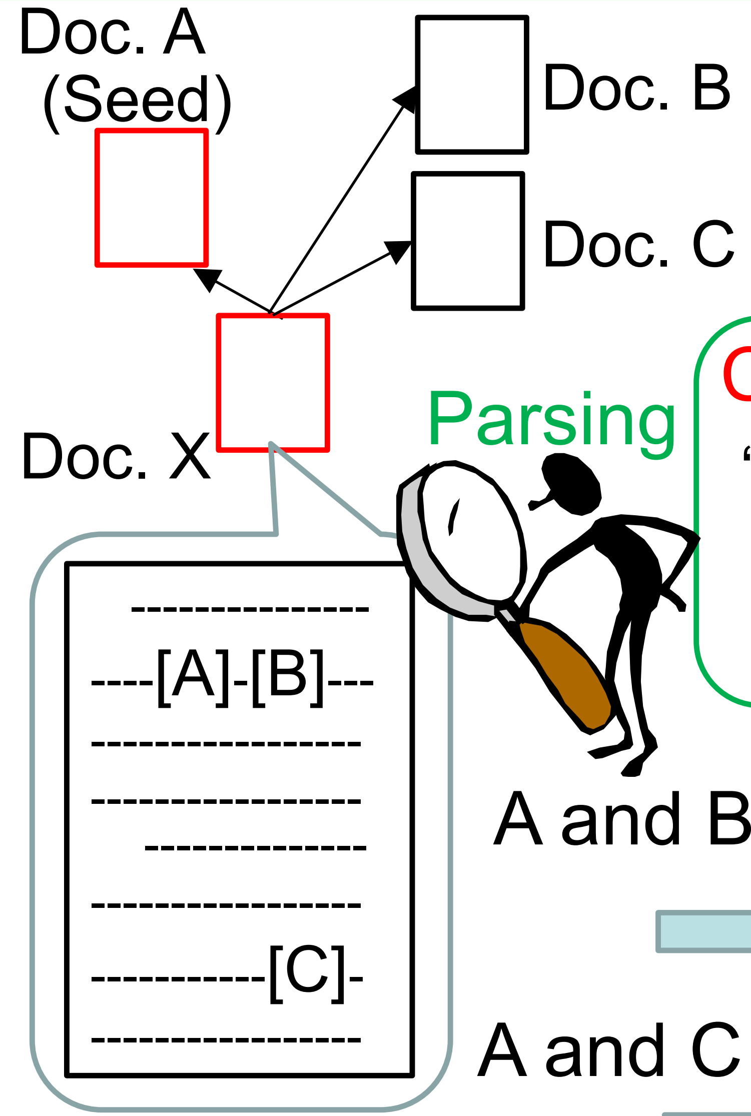


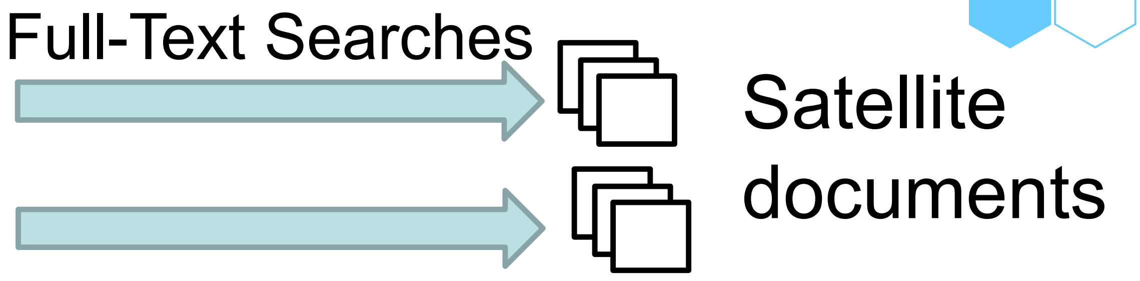
Relevant host yields
a lot of relevant satellite documents

Irrelevant host yields
a lot of irrelevant satellite documents

Checking the appropriateness of host documents

Checking the Appropriateness of Host Documents (optional process)





Co-citation contexts are analyzed

"Co-citation in the same paragraph has strong relationship"
(Eto 2013, Gipp & Beel 2009)

A and B are cited in the same paragraph → Doc. B is selected as host

A and C are cited in different paragraphs → Doc. C is not selected as host

Citing document

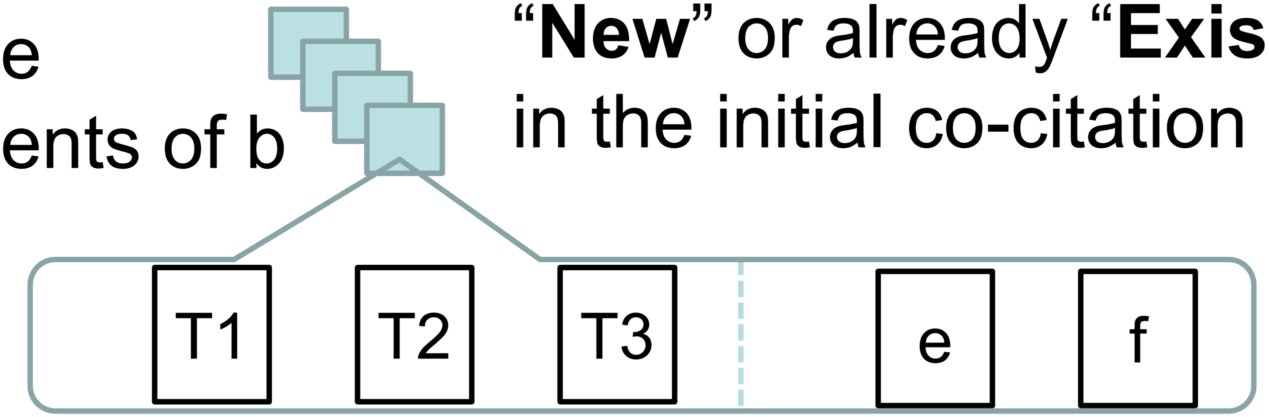
Outline of this presentation

1. Background
    - Co-citation and network model
    - Similar document search
2. Research question
    - Satellite documents
3. Proposed Retrieval Method
    - Specifying satellite documents
    - Incorporating satellite documents
    - Ranking documents in the network
4. Experiment
    - Evaluating the proposed method

Incorporating Satellite Documents

Satellite documents of b

"**New**" or already "**Existing**" in the initial co-citation network

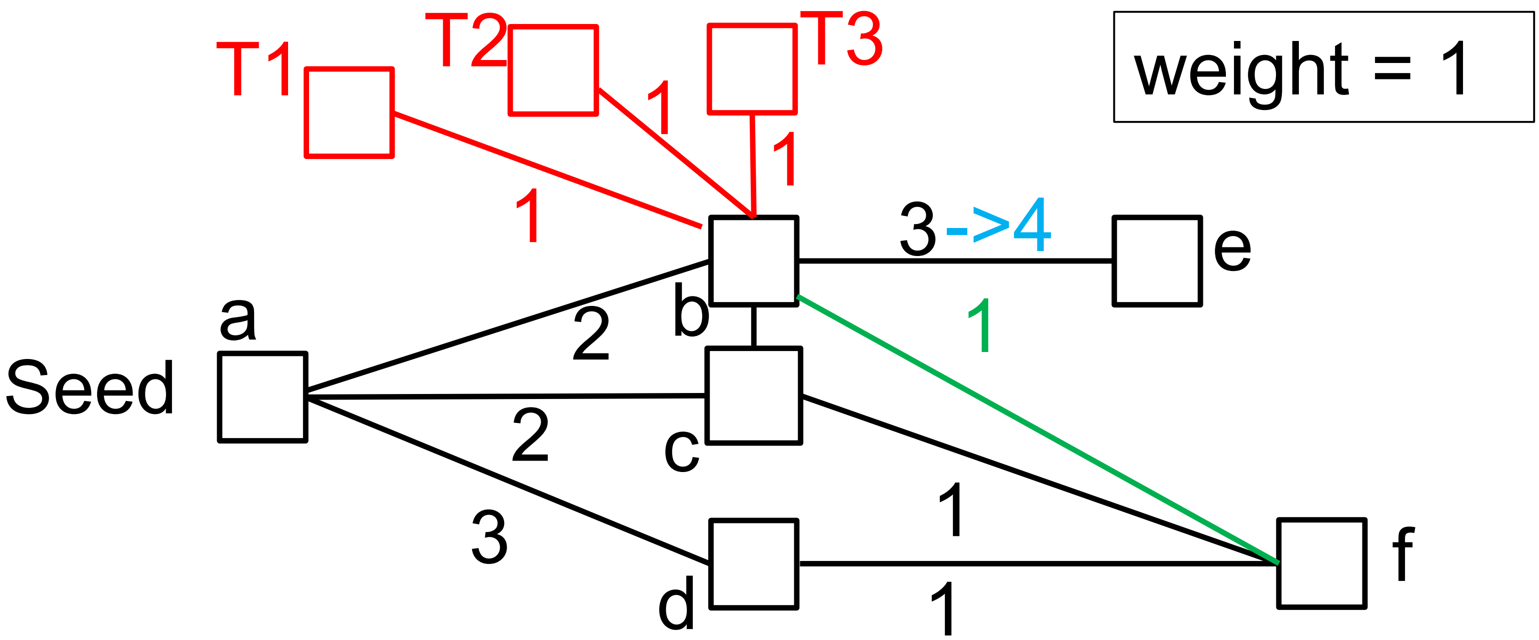


New

New node and new edge

Existing

Added weight or New edge



Outline of this presentation

1. Background

    Co-citation and network model

    Similar document search

2. Research question

    Satellite documents

3. Proposed Retrieval Method

    Specifying satellite documents

    Incorporating satellite documents

    Ranking documents in the network

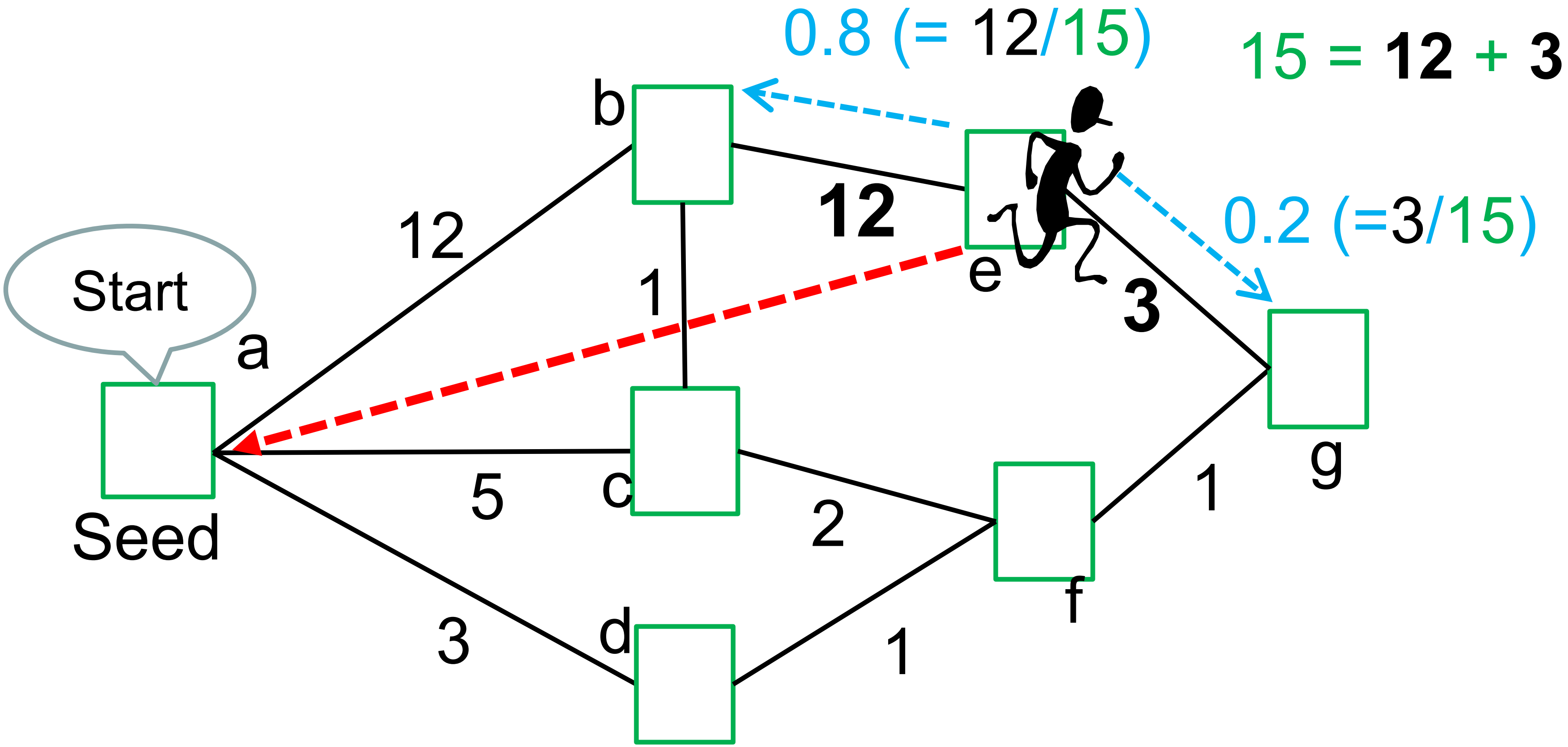
4. Experiment

    Evaluating the proposed method

Ranking Documents in the Network by the RWR (Random walk With Restart) Algorithm (Tong, 2008)

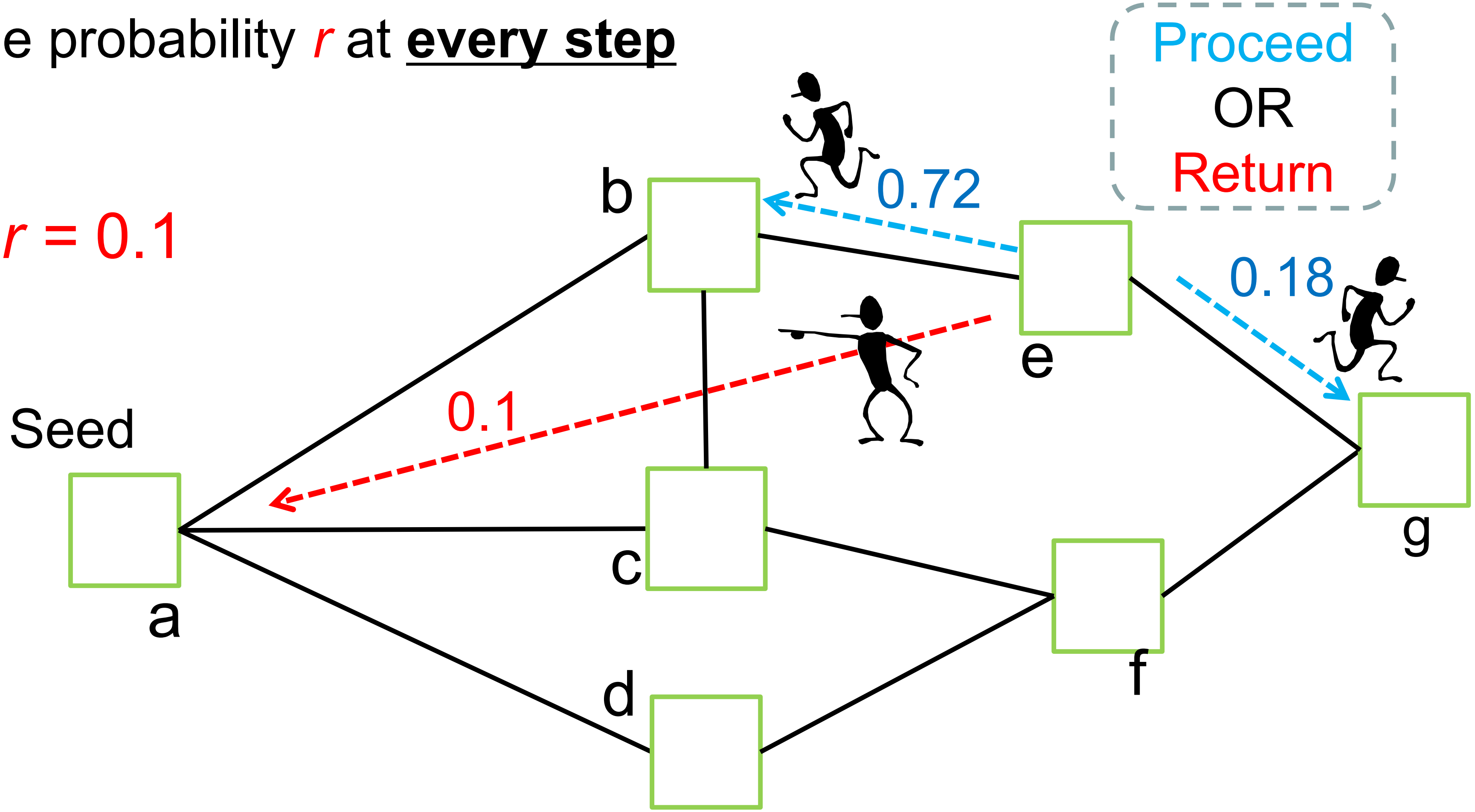
Simple random walk

The walker proceeds to the connected documents based on transition probabilities calculated by weights of edges



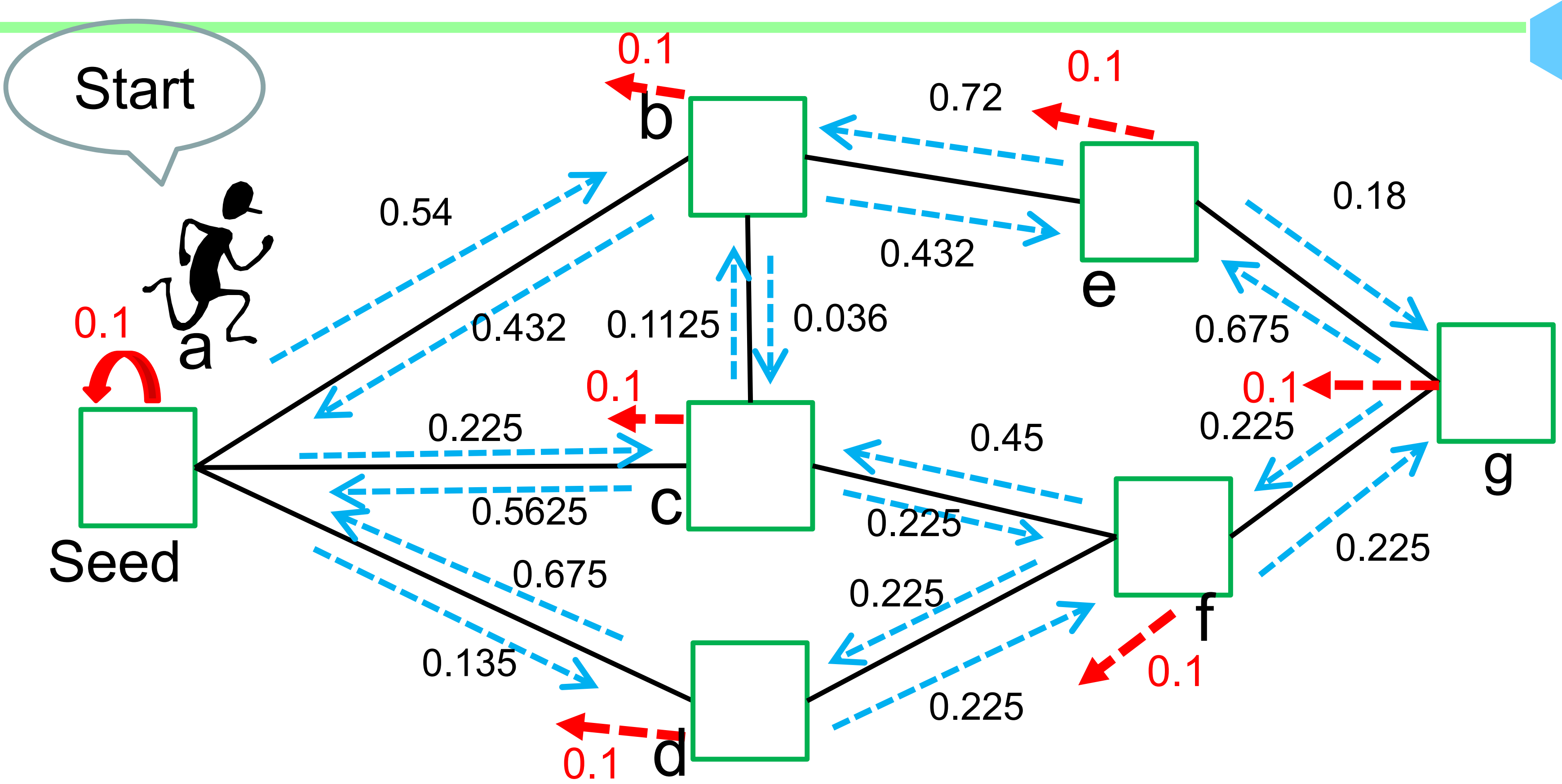
RWR: What is ‘*Restart*’?

The walker returns to the seed document with the probability *r* at **every step**



r ≒ parameter of the penalty for distance from the seed
(If *r* is high, documents near the seed have high document scores)

RWR: How are document scores calculated?

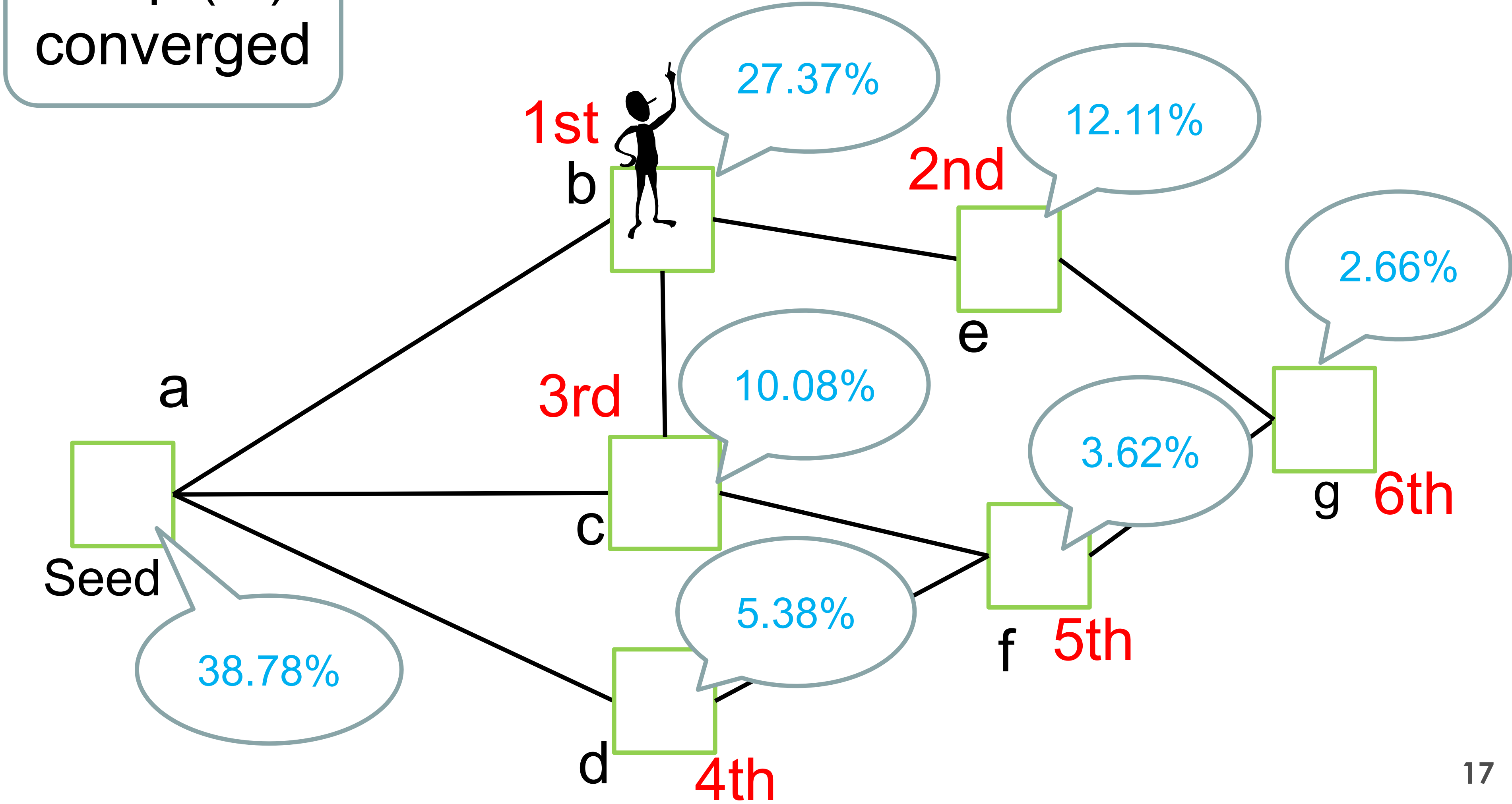


- The position of the walker at Step (*t*) can be estimated by the transition probabilities
- When *t* is low, the position probability is unstable. As the number of *t* increases, the position probability may converge

RWR: How are documents ranked?

Step (∞)
converged

Converged position probability =
Document score



Outline of this presentation

1. Background

    Co-citation and network model

    Similar document search

2. Research question

    Satellite documents

3. Proposed Retrieval Method

    Specifying satellite documents

    Incorporating satellite documents

    Ranking documents in the network

4. Experiment

    Evaluating the proposed method

Information Retrieval Experiment

Retrieval Methods

- Baseline (initial co-citation network)

  Network created by taking up to two hops from the seed

- Proposed Method (all)

  All one hop documents from the seed are host documents

- Proposed Method (context)

  Host documents are selected by co-citation context

Test Collection

- 152,000 documents (XML) (Pubmed central dataset)
- Each document has MeSH descriptors
- 100 seed documents

Evaluation metric

- nDCG@K (K = 5, 10, 50, 100)

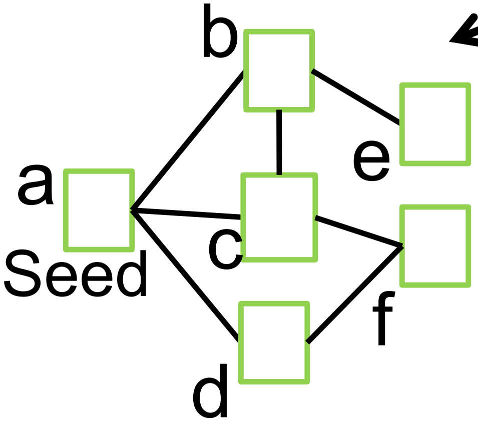
Search Run

Seed

Input a seed document

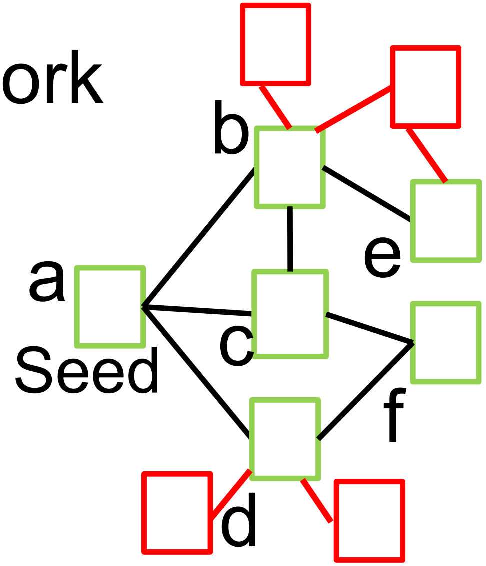
152,000 documents

Create an initial co-citation network



Baseline

Incorporating satellite documents

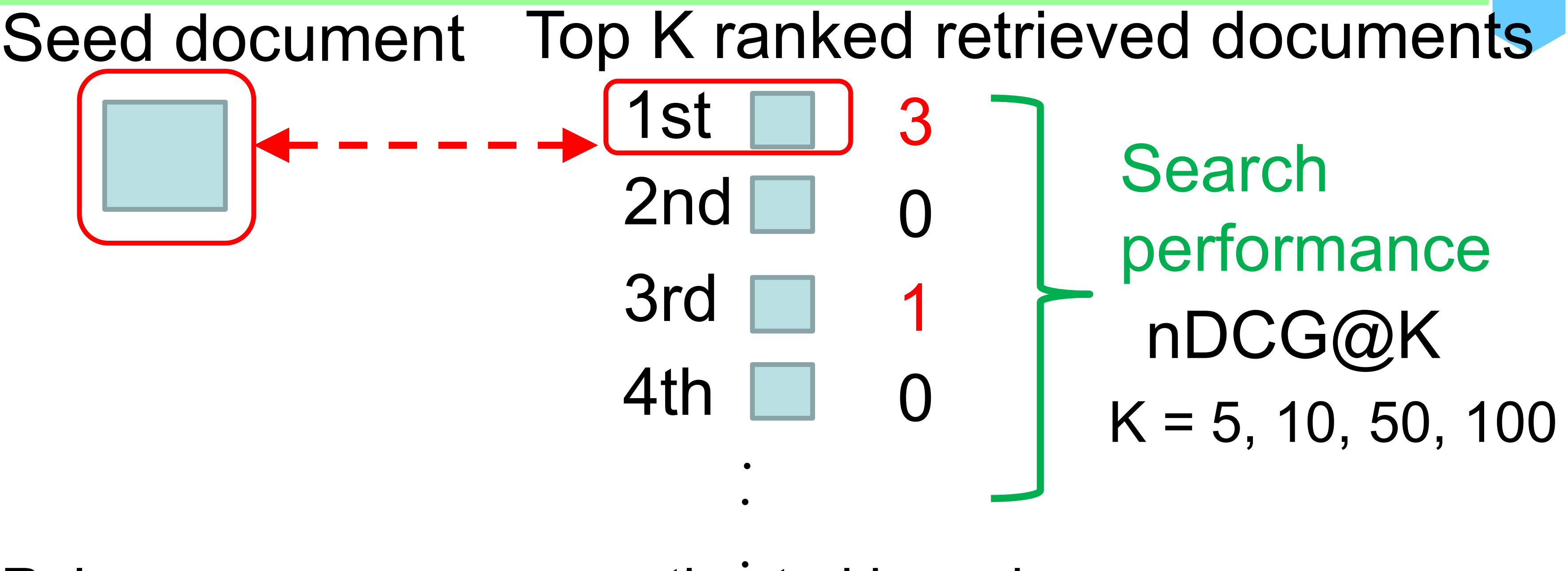


Proposed methods

- All
- Context

Ranked results by RWR are compared

Relevance Assessment



Seed document

Top K ranked retrieved documents

1st 3

2nd 0

3rd 1

4th 0

:

Search performance

nDCG@K

K = 5, 10, 50, 100

Relevance scores were estimated based on **similarity** between the seed and each retrieved document

Jaccard Coeffiecinet based on MeSH descriptors

Jaccard Coeffiecinet	Relevance Score
>= 0.3	3
>= 0.2	2
>= 0.1	1

Result (averaging results of 100 seed)

		Proposed N = 10		Proposed N = 100	
K	Baseline	all	context	all	context
5	.226	.226	.232*	.224	.234**
10	.223	.221	.227**	.226	.230**
50	.188	.191*	.189**	.197**	.191
100	.174	.181**	.177*	.188**	.180**

* P < .05, ** P < .01

- The maximum scores at each K are the results of Proposed with N = 100

⇨ Proposed methods tended to outperform the baseline

- The scores of Proposed (context) are higher than those of the baseline method in all cases

⇨ The checking process had a stable and positive impact on improving the search performance

Conclusion

This study proposed a technique to enlarge co-citation networks by incorporating satellite documents in scientific paper searches

Retrieval methods using the proposed technique tended to outperform the baseline method, which was based on the initial co-citation network

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP26730163

Q and A

Thank you!