# Prioritizing and Scheduling Conferences for Metadata Harvesting in dblp

M. Neumann<sup>1</sup> C. Michels<sup>2</sup> P. Schaer<sup>1</sup> R. Schenkel<sup>2</sup>

<sup>1</sup>Department of Information Science TH Köln (University of Applied Sciences)

<sup>2</sup>Department of Computer Science University of Trier

Joint Conference on Digital Libraries, 2018

## Outline

- 1 Motivation
- 2 Research Question
- 3 Method
- 4 Our Results/Contribution



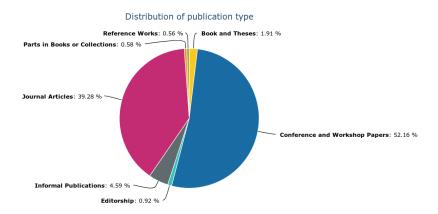
## Outline

- Motivation
- Research Question
- Method
- Our Results/Contribution



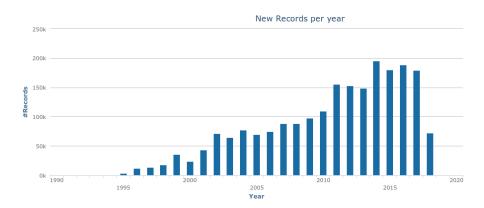
### The dblp computer science bibliography:

- on-line reference for bibliographic information on CS
- free access to high-quality bibliographic meta-data
- >4 million publication records
- $\blacksquare$  originating from  $\approx$ 5.400 conferences and  $\approx$ 1.500 journals

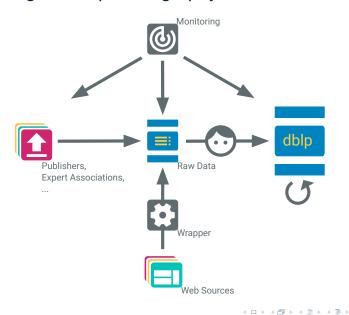




New entries to the database per year: conference and workshop papers









Mandy Neumann (TH Köln)

## Motivation

- limited resources
- conferences: arbitrary intervals
- not all records equally important to dblp
  - ightarrow identify and prioritize missing data in the acquisition process



## Outline

- 1 Motivation
- 2 Research Question
- 3 Method
- 4 Our Results/Contribution



### Research Question

How can we find a prioritization mechanism for conference series with regard to their expected urgency for the data acquisition process at a given point in time?

→ Ranking problem: rank the set of conferences in descending order according to their relevance to the database



## Outline

- 1 Motivation
- 2 Research Question
- 3 Method
- 4 Our Results/Contribution



- base ranking primarily dependent on temporal patterns
  - relation between past event dates and dates of entry to dblp
- add additional factors to study influence on ranking
  - loosely based on information quality / dblp quality criteria



### Temporal patterns

- basic date-based calculation of expectancy
  - → delay as base scoring factor
- publication date of proceedings not available use entry date to dblp database as approximation



January	February	March	April
1 2 3 4 5 6 7 8 9 10111213141516	1 2 3 4 5 6 7 8 910111213 14151617181920 21222324252627	1 2 3 4 5 6 7 8 910111213 14151617181920 21222324252627	1 2 3 4 5 6 7 8 910 11121314151617
31	20	20293031	232027202930

May	June	July	August
1	1 2 3 4 5		1 2 3 4 5 6 7
		4 5 6 7 8 910	
		11121314151617	
16171819202122	20212223242526	18192021222324	22232425262728
23242526272829	27282930	25262728293031	293031
3031			

September	October	November	December
1 2 3 4	1 2		1 2 3 4
5 6 7 8 91011	3 4 5 6 7 8 9	7 8 910111213	5 6 7 8 91011
121314151617 <u>18</u>	10111213141516	14151617181920	12131415161718
19202122232425	17181920212223	21222324252627	19202122232425
2627282930	24252627282930	282930	262728293031
	31		



January	February	March	April
2 3 4 5 6 7 8	1 2 3 4 5	1 2 3 4	1 2 3 4 5 6 7 8 9101112131415 16171819202122 23242526272829 30
9101112131415	6 7 8 9101112	5 6 7 8 91011	
16171819202122	13141516171819	12131415161718	
23242526272829	20212223242526	19202122232425	
3031	272829	262728293031	
May	June	July  2 3 4 5 6 7 8 9101112131415 16171819202122 23242526272829 3031	August
1 2 3 4 5 6	1 2 3		1 2 3 4 5
7 8 910111213	4 5 6 7 8 910		6 7 8 9101112
14151617181920	111213 <mark>14</mark> 151617		13141516171819
21222324252627	1819202122324		20212223242526
28293031	252627282930		2728293031
September 1 2 3 4 5 6 7 8 9 101 122 3141516 17181920212223 24252627282930	October 1 2 3 4 5 6 7 8 91011121314 15161718192021 22232425262728 293031	November 1 2 3 4 5 6 7 8 91011 12131415161718 1920212232425 2627282930	December 1 2 3 4 5 6 7 8 9 10111213141516 1718192021223 24252627282930 31



1 2 3 4 5 6       1 2 3       1 2 3       1 2 3 4 5         7 8 910111213       4 5 6 7 8 910       4 5 6 7 8 910       8 9101112         14151617181920       11121314151617       11121314151617       1516171818         21222324252627       18192021222324       18192021222324       2232425262         28293031       25262728       25262728293031       2930	21314 92021

May	June	July	August
1 2 3 4 5	1 2	1 2 3 4 5 6 7	1 2 3 4
6 7 8 9101112	3 4 5 6 7 8 9	8 91011121314	5 6 7 8 91011
13141516171819	10111213141516		121314 <u>15</u> 161718
20212223242526		22232425262728	19202122232425
2728293031	24252627282930	293031	262728293031

September	October	November	December
1	1 2 3 4 5 6	1 2 3	1
	7 8 910111213	4 5 6 7 8 910	2 3 4 5 6 7 8
9101112131415	14151617181920	11121314151617	9101112131415
16171819202122		18192021222324	16171819202122
23242526272829	28293031	252627282930	23242526272829
30			3031



### 2014

January 1 2 3 4 5 6 7 8 9101112 13141516171819 20212223242526 2728293031	February 1 2 3 4 5 6 7 8 9 10111213141516 17181920212223 2425262728	March 1 2 3 4 5 6 7 8 9 10111213141516 17181920212223 24252627282930 31	
May 1 2 3 4 5 6 7 8 91011 12131415161718 1920212232425 262728293031	June 1 2 3 4 5 6 7 8 9101 12 31415 16171819202122 23242526272829 30	July 1 2 3 4 5 6 7 8 910111213 14151617181920 21222324252627 28293031	August 1 2 3 4 5 6 7 8 910 11121314151617 18192021222324 25262728293031
September 1 2 3 4 5 6 7 8 91011121314 15161718192021 22232425262728 2930	October 1 2 3 4 5 6 7 8 9101112 13141516171819 20212223242526 2728293031	November 1 2 3 4 5 6 7 8 9 10111213141516 17181920212223 24252627282930	December 1 2 3 4 5 6 7 8 91011121314 15161718192021 22232425262728 293031





January	February	March	April
1 2 3 4 5 6 7 8 91011 12131415161718 19202122232425 262728293031	1 2 3 4 5 6 7 8 9101112131415 16171819202122 232425262728	2 3 4 5 6 7 8 9101112131415 16171819202122 23242526272829 3031	1 2 3 4 5 6 7 8 9101112 13141516171819 20212223242526 27282930
May 1 2 3 4 5 6 7 8 910 11121314151617 1819202122324 25262728293031	June 1 2 3 4 5 6 7 8 91011121314 15161718192021 22232425262728 2930	July 1 2 3 4 5 6 7 8 9101112 13141516171819 20212223242526 2728293031	August 1 2 3 4 5 6 7 8 9 10111213141516 1718192021223 24252627282930 31
September 1 2 3 4 5 6 7 8 910111213 1415 617181920 21222324252627 282930	October 1 2 3 4 5 6 7 8 91011 12131415161718 19202122232425 262728293031	November 1 2 3 4 5 6 7 8 9101112131415 16171819202122 23242526272829 30	December 1 2 3 4 5 6 7 8 910111213 14151617181920 21222324252627 28293031



- Example conference:
  - interval: 1
  - usual month: June
  - usual delay: 3 months
  - → expected: September 2016
- 177 other conferences also due in September
- base scoring: raw delay between expected and current date; mapping of raw delay to intervals to smooth out high delays



### Additional factors to refine priority ranking:

- conference rating
- citation counts
- discontinuity indicator
- internationality
- author prominence





- conference rating
- citation counts
- discontinuity indicator
- internationality
- author prominence



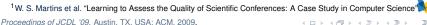
- conference rating: CORE; Martins et al.¹
- citation counts
- discontinuity indicator
- internationality
- author prominence

¹W. S. Martins et al. "Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science ↑

Proceedings of JCDL '09. Austin, TX, USA: ACM, 2009.

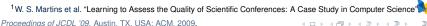


- conference rating: CORE; Martins et al.¹
- citation counts: Microsoft Academic Graph (MAG)
- discontinuity indicator
- internationality
- author prominence

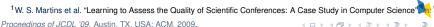




- conference rating: CORE; Martins et al.<sup>1</sup>
- citation counts: Microsoft Academic Graph (MAG)
- discontinuity indicator: self-defined, in terms of #years since last appearance in dblp
- internationality
- author prominence



- conference rating: CORE; Martins et al.<sup>1</sup>
- citation counts: Microsoft Academic Graph (MAG)
- discontinuity indicator: self-defined, in terms of #years since last appearance in dblp
- internationality: self-defined, in terms of #countries of conference venues
- author prominence



- conference rating: CORE; Martins et al.<sup>1</sup>
- citation counts: Microsoft Academic Graph (MAG)
- discontinuity indicator: self-defined, in terms of #years since last appearance in dblp
- internationality: self-defined, in terms of #countries of conference venues
- author prominence: dblp data



#### Gold standard:

- human judgments hardly practicable
- pseudo-relevance:
  - distance in months between current month and month of ingestion into dblp
  - mapped onto intervals
  - inverted to give higher values to more recent entries



## Outline

- 1 Motivation
- 2 Research Question
- 3 Method
- 4 Our Results/Contribution



every factor outperforms baseline

Table 1: Overview on ndcg-100 values for each month and the year's average.

system	jan	feb	mar	apr	 dec	avg
baseline	0.240	0.338	0.353	0.434	 0.605	0.505
conf. rating	0.230	0.378	0.524	0.627	 0.736	0.645
internationality	0.226	0.331	0.507	0.610	 0.679	0.608
discontinued	0.291	0.411	0.615	0.727	 0.711	0.643
citations	0.225	0.333	0.442	0.517	 0.643	0.554
prominence	0.248	0.423	0.568	0.637	 0.696	0.608

every factor outperforms baseline

Table 1: Overview on ndcg-100 values for each month and the year's average.

system	jan	feb	mar	apr	 dec	avg
baseline	0.240	0.338	0.353	0.434	 0.605	0.505
conf. rating	0.230	0.378	0.524	0.627	 0.736	0.645
internationality	0.226	0.331	0.507	0.610	 0.679	0.608
discontinued	0.291	0.411	0.615	0.727	 0.711	0.643
citations	0.225	0.333	0.442	0.517	 0.643	0.554
prominence	0.248	0.423	0.568	0.637	 0.696	0.608

Table 2: Comparison of ndcg values on different cut-offs. Statistical differences to the baseline tested with two-sided t-test (\* \* \* = p < 0.001, \*\* = p < 0.05).

system	ndcg-10	ndcg-20	ndcg-100	ndcg-200
baseline	0.530	0.545	0.505	0.439
conf. rating	0.739**	0.716**	0.645***	0.597***
internationality	0.616	0.632	0.608***	0.575***
discontinued	0.713**	0.686***	0.643***	0.594***
citations	0.588	0.575	0.554***	0.548***
prominence	0.681**	0.662**	0.608***	0.577***



Table 2: Comparison of ndcg values on different cut-offs. Statistical differences to the baseline tested with two-sided t-test (\* \* \* = p < 0.001, \*\* = p < 0.05).

system	ndcg-10	ndcg-20	ndcg-100	ndcg-200
baseline	0.530	0.545	0.505	0.439
conf. rating	0.739**	0.716**	0.645***	0.597***
internationality	0.616	0.632	0.608***	0.575***
discontinued	0.713**	0.686***	0.643***	0.594***
citations	0.588	0.575	0.554***	0.548***
prominence	0.681**	0.662**	0.608***	0.577***



Table 2: Comparison of ndcg values on different cut-offs. Statistical differences to the baseline tested with two-sided t-test (\* \* \* = p < 0.001, \*\* = p < 0.01, \* = p < 0.05).

system	ndcg-10	ndcg-20	ndcg-100	ndcg-200
baseline	0.530	0.545	0.505	0.439
conf. rating	0.739**	0.716**	0.645***	0.597***
internationality	0.616	0.632	0.608***	0.575***
discontinued	0.713**	0.686***	0.643***	0.594***
citations	0.588	0.575	0.554***	0.548***
prominence	0.681**	0.662**	0.608***	0.577***



## Interpretation

### Best performing factors in terms of information quality:

- credibility:
  - expressed through ratings
- currency:
  - expressed through penalty by discontinuity
- popularity:
  - expressed through citation, internationality and prominence scores



# Summary

- We can use information quality-related features to rank conferences for data ingestion routines.
- All proposed features outperform the baseline derived from ingestion delays.

- Outlook
  - combine features
  - separate workshops
  - extend approach to journals etc.



## Discussion

Thank you for your attention! Feel free to ask any questions now!

#### Contact us:

mandy.neumann@th-koeln.de
michelsc@uni-trier.de
philipp.schaer@th-koeln.de
schenkel@uni-trier.de

Visit http://dblp.uni-trier.de



### Table of contents

- 1 Motivation
- 2 Research Question
- 3 Method
- 4 Our Results/Contribution

