Prioritizing and Scheduling Conferences for Metadata Harvesting in dblp

M. Neumann[1] C. Michels[2] P. Schaer[1] R. Schenkel[2]

[1] Department of Information Science
TH Köln (University of Applied Sciences)

[2] 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

1 Motivation

2 Research Question

3 Method

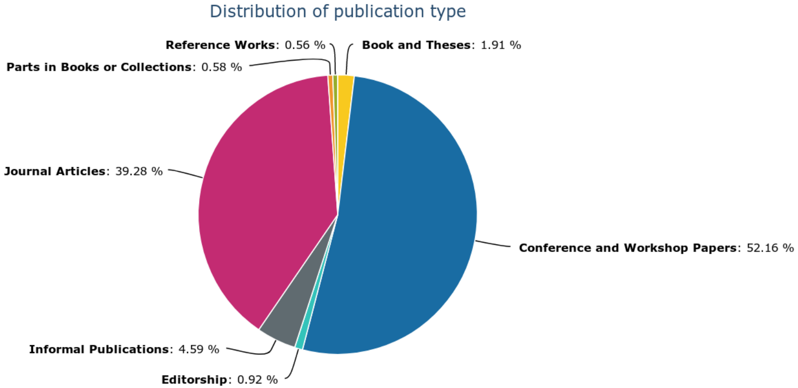
4 Our Results/Contribution

Maintaining the dblp Bibliography

The dblp computer science bibliography:

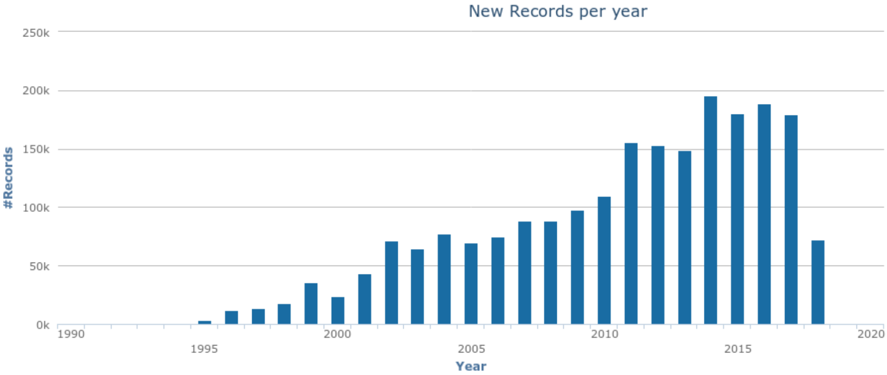
- on-line reference for bibliographic information on CS
- free access to high-quality bibliographic meta-data
- >4 million publication records
- originating from ≈5,400 conferences and ≈1,500 journals

Maintaining the dblp Bibliography

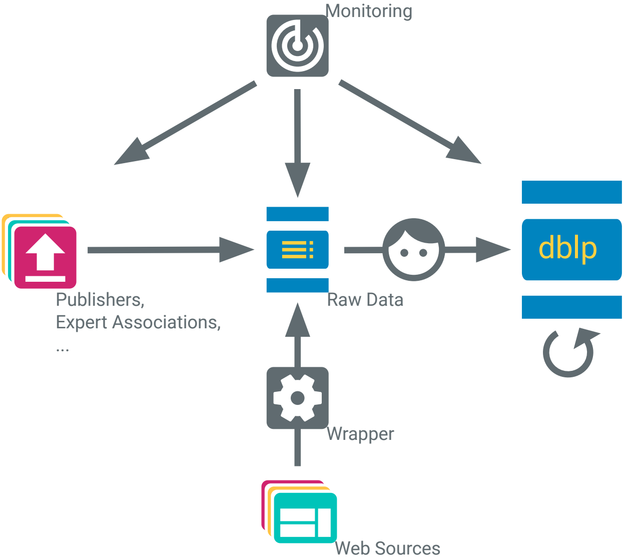


Maintaining the dblp Bibliography

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



Maintaining the dblp Bibliography



Motivation

- limited resources
- conferences: arbitrary intervals
- not all records equally important to dblp

  → 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

Method

- 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

Method

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

2011

January

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

February

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28						

March

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

April

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

May

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

June

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

July

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

August

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

September

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

October

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

November

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

December

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

2012

January

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

February

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29				

March

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

April

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

May

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

July

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

August

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

September

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

October

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

November

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

December

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

2013

January

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

February

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28			

March

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

April

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

May

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

June

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

July

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

August

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

September

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

October

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

November

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

December

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

2014

January

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

February

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

May

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

June

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

July

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

August

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

September

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

October

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

November

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

December

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

2015

```
    January
       1 2 3 4
 5 6 7 8 91011
12131415161718
19202122232425
262728293031

    February
             1
 2 3 4 5 6 7 8
 9101112131415
16171819202122
232425262728

     March
             1
 2 3 4 5 6 7 8
 9101112131415
16171819202122
23242526272829
3031

     April
     1 2 3 4 5
 6 7 8 9101112
13141516171819
20212223242526
27282930

      May
         1 2 3
 4 5 6 7 8 910
11121314151617
18192021222324
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
17181920212223
24252627282930
31

   September
   1 2 3 4 5 6
 7 8 910111213
14151617181920
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
```

Method

- 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

Method

Additional factors to refine priority ranking:

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

Method

Data sets:

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

Method

Data sets:

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

[1] W. S. Martins et al. “Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science”. In: *Proceedings of JCDL '09*. Austin, TX, USA: ACM, 2009.

Method

Data sets:

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

[1] W. S. Martins et al. “Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science”. In: *Proceedings of JCDL ’09*. Austin, TX, USA: ACM, 2009.

Method

Data sets:

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

[1] W. S. Martins et al. "Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science". In: *Proceedings of JCDL '09*. Austin, TX, USA: ACM, 2009.

Method

Data sets:

- conference rating: CORE; Martins et al.[1]
- 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

[1] W. S. Martins et al. “Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science”. In: *Proceedings of JCDL ’09*. Austin, TX, USA: ACM, 2009.

Method

Data sets:

- conference rating: CORE; Martins et al.[1]
- 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

[1] W. S. Martins et al. "Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science". In: *Proceedings of JCDL '09*. Austin, TX, USA: ACM, 2009.

Method

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

Main Results

- 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

Main Results

- 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

Main Results

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***

Main Results

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***__

Main Results

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