Conference Indexing in Digital Libraries
A Ranking Model and Case Study on dblp

Christopher Michels[1][0000−0003−0523−8547], Mandy Neumann[2][0000−0003−3694−4997], Philipp Schaer[2][0000−0002−8817−4632], and Ralf Schenkel[3][0000−0001−5379−5191]

[1] Schloss Dagstuhl LZI, dblp
christopher.michels@dagstuhl.de
[2] TH Köln - University of Applied Sciences
firstname.lastname@th-koeln.de
[3] Trier University, dblp
schenkel@uni-trier.de

Abstract. Digital library curators make relevance decisions in their daily work to prioritize the most urgent metadata updates. In this work, we propose a complex relevance and ranking model to support the decision and prioritization process of digital library curators. Our approach incorporates different aspects of relevance decisions into a framework for feasible data quality management in digital libraries. A case study demonstrates the effects of the factors we use to model these aspects.

Keywords: bibliometrics · conferences · database curation · dblp · digital library · ranking models

1 Introduction

Digital libraries[4] need to retrieve and keep up with the most recent relevant research to fulfill the information needs of their target audience in the scientific community. Their workload is dominated by indexing the tables of content with missing old or expected new publications. With limited time and workforce, curators have to tackle an extensive, complex, and dynamic pool of heterogeneous data sources, ranging from numerous individual hints in e-mails on a small scale to harvesting websites and data feeds of publishers on a larger scale. This bottleneck requires bibliographic prioritization: The most important indexing updates for missing or upcoming publications need to be identified and addressed first.

In terms of information retrieval, we can model this prioritization process as a ranking problem. A given set of data sources has to be ranked according to specified criteria. The features used for ranking should reflect the curators' relevance

[4] We use the term digital libraries to consolidate different systems like reference databases or online bibliographies.

decisions. These relevance decisions of digital library curators are governed by higher standards of quality than those for other information access systems [1]. In contrast to web search engines such as Google Scholar, a digital library such as dblp[5] is expected to provide structured authority data as well as more consistent and coherent system responses, relying on a higher understanding of the data, tasks, communities, and the specific information needs involved.

In computer science and related disciplines, conferences constitute the main channel of sharing results with the research community [5]. Dynamic, community-driven lifecycles and event structures are essential hallmarks of conferences, in comparison with other publication venues such as journals or book series.

Bibliometric analyses are a source of context knowledge which is of particular importance to library curators. Lee [3] use several conference-related factors to predict citation rates on conference papers. Some factors they investigate are name, age, size, and internationality of conference series. Size is operationalized in terms of the number of papers presented, which is a commonly available factor for conferences as opposed to other possible operationalizations like the number of submissions, attendees, sponsors, or conference profits. To operationalize internationality, they use the degree of international collaborations in papers. They also look at the age of conferences, trying to answer the question if longer running conference series are able to gather more citations on papers than shorter running ones. They find that internationality is one of the factors that significantly contribute to citation rates. Size has a negative correlation with citation: the fewer papers are presented at a conference, the more they are cited.

Besides conference-related metadata such as size, and external bibliometric data such as citation numbers, archive-internal factors also influence curator decisions. For example, the prominence of an author within the archive might also influence which pending metadata updates are addressed first.

In this work, we propose a complex ranking and relevance model to support the decision and prioritization process of digital library curators. Our approach incorporates different aspects of curation decisions into a framework for feasible data quality management in digital libraries. It incorporates both bibliometric and retrieval-related elements. While the factors itself are motivated through bibliometric research and findings, all proposed methods can easily be integrated into an actual ranking model of a retrieval or recommendation system. This work is based on the prioritization mechanism for conference metadata updates from our previous work presented at JCDL 2018 [7]. We extend our previous model and describe the different ranking factors and the corresponding data sources at a much higher level of detail and include a case study that illustrates the feasibility of the approach. Our main goal is to thoroughly describe the ranking factors to allow the readers to understand the mechanics behind them entirely.

In Section 2, we describe the components we employ to rank conferences with pending metadata updates. A case study on ranking conference updates in dblp with these components is presented in Section 3. The work closes with a short discussion and an outlook on future work in Section 4.

[5] `https://dblp.org`

2 Components of the Ranking Model

The task at hand can be modeled as a ranking problem. All conferences listed in the archive have to be ranked according to an information need inherent in archive curation. The curators need to have the conferences ranked highest for which an update is expected. In case of multiple conferences being due simultaneously, their ranking should reflect their priority for the archive. Usually, curators are guided by several criteria when deciding which conference to index next. In describing and evaluating a set of such rank-establishing factors, the following notation applies.

Each conference c (like CIKM, ECIR, JCDL, etc.) of the conference set C is constituted of conference events $E(c) := (e_n, \ldots, e_1)$. In bibliographic terms, an event e groups all volume-level event members $V(e) := (v_1, \ldots, v_n)$, e.g., proceedings, workshops, or other named parts of published content[6]. Modeling the variety in conference event members is simplified here. Event members are simply attributed to a conference by their date $\textsc{date}(v)$. The date $\textsc{date}(v)$ consists of the year $\textsc{year}(\textsc{date}(v))$ and the month $\textsc{month}(\textsc{date}(v))$ of the last day of the event, i.e., $\textsc{date}(e) = \textsc{date}(v)\ \forall v \in V(e)$.

The set $V_f(c) := (v \mid \exists e\colon e \in E(c), v \in V(e))$ contains all event members of a given conference, regardless of the event to which they belong, for a partial function f. When f depends on external metadata sources, e.g., citation links, no value might be provided for a given event member $v \notin V_f(c)$.

2.1 Base delay score

The primary criterion for ranking conferences by urgency is the delay between the expected next indexing date of a new record and the current date NOW. Conference events usually occur at regular intervals and at roughly the same time of the year. In the archive, there is a delay between the event and archiving date $\delta_{\textsc{indexed}}(v) := \textsc{indexed}(v) - \textsc{date}(v)$ for each proceedings record v.[7] Given these regularities and the edit history of an archive, we can estimate the next event and when it is expected to be indexed.

We assume that the events $E(c)$ of a conference c are in decreasing order of their date $\textsc{date}(e)$, with the most recent event being denoted by e_n. The limited set of up to u most recent events is referred to as $\mathit{recent}_u(c)$.We use $\mathit{recent}_6(c)$ to determine the characteristic interval between the last five events in months (unit M), $\delta_{\textsc{event}}(c)$, with a default of $12M$. Furthermore, the archive delay $\delta_{\textsc{indexed}}(c)$ is the median of $\delta_{\textsc{indexed}}(v)$ for $\mathit{recent}_5(c)$. Finally, to determine the expected event month $\textsc{expmonth}(c)$, we take the mode of $\textsc{month}(\textsc{date}(e))$ of $\mathit{recent}_5(c)$[8].

Then, the expected date of the next conference event is $\textsc{date}(e_{n+1}) = (\textsc{expmonth}(c),\ \textsc{year}(\textsc{date}(e_n))) + \delta_{\textsc{event}}(c)$. The recording delay of c is approximated based on this estimated next event and the archive delay in Eq. (1).

[6] In the following referred to as proceedings.

[7] $\delta_{\textsc{indexed}}(v)$ may be negative in the case of pre-proceedings.

[8] In case of a multi-modal distribution, we take the most recent most frequent month.

If no new entry is expected yet ($\text{DATE}(e_{n+1}) > \mathsf{NOW}$), the conference scores 0 in the ranking. Otherwise, $delay(c)$ is log-smoothed and inverted in Eq. (2) to compute the delay scoring factor. Conferences for which new entries are expected promptly then rank highest, while extremely high delays are practically ignored. The base score and the factors below have the range $[1, 2]$.

$$delay(c) = \mathsf{NOW} - (\text{DATE}(e_{n+1}) + \delta_{\text{INDEXED}}(c)) \tag{1}$$

$$w_{delay}(c) = 1 + \frac{1}{1 + \log_2(delay(c) + 1)}, delay(c) > 0 \tag{2}$$

2.2 Ranking Factors

We define the following ranking factors, which reflect the extent to which there is a need to index a conference and help determine a corresponding score for the ranking. The estimated delay of the expected next event constitutes a basic factor that is combined with a boosting factor for each of the remaining criteria of a given conference. Each combination results in a corresponding ranking score:

$$\begin{gathered} score_\phi(c) = w_{delay}(c) \times w_\phi(c), \\ \phi \in \Phi = \{active, rate, size, intl, affil, cite, prom\} \end{gathered} \tag{3}$$

Activity. A conference might cease to be organized without any notice reaching the archive. Conferences that are likely to be discontinued should thus receive a lower score than active ones. The following scoring takes the relation between the time since the last entry and the regular event interval into account (Eq. 4). The activity scoring factor in Eq. (5) then boosts conferences that have a recently active life-cycle according to their archive history.

$$age(c) = \frac{\mathsf{NOW} - \text{DATE}(e_n)}{\delta_{\text{EVENT}}(c)} \tag{4}$$

$$w_{active}(c) = 1 + \frac{1}{(1 + age(c)^2)} \tag{5}$$

Ratings. External ratings usually guide update cycles if they coincide with the set of conferences that are relevant to an archive. If such ratings can be integrated into a single list for a given conference c, a list of corresponding numeric rating values attributed to c is yielded by $rated(c)$ in Eq. (6). With the average numeric rating $rate(c)$ the conference rating weighting factor is given in Eq. (7).

$$rate(c) = \frac{\sum_{r \in rated(c)} r}{|rated(c)|} \tag{6}$$

$$w_{rate}(c) = 1 + \frac{rate(c)}{max_{rate} C} \tag{7}$$

Size. Similar to Lee [3], we define the size of a conference in terms of the average number of papers for each event $e(c)$. The set of papers for a given event member v is denoted by $papers(v)$. The size factor is normalized by the maximum value present in the data set (Eq. 9).

$$size(c) = \frac{\sum_{v \in V_{size}(c)} |papers(v)|}{|V_{size}(c)|} \tag{8}$$

$$w_{size}(c) = 1 + \frac{size(c)}{max_{size}C} \tag{9}$$

Internationality. The number of locations where events are organized can be an indicator of the internationality of a conference. The multiset $located(c)$ contains all locations of all event members of a given conference, whereas $locations(c)$ contains all distinct elements in $located(c)$. The internationality of a conference $intl(c)$ in Eq. (10) is defined as the number of distinct event countries, divided by the total of their occurrences. Normalization for the corresponding scoring factor is trivial since there cannot be more locations than event members for a single conference.

$$intl(c) = \frac{|locations(c)|}{|located(c)|} \tag{10}$$

$$w_{intl}(c) = 1 + \frac{intl(c)}{max_{intl}C} = 1 + \frac{intl(c)}{1} = 1 + intl(c) \tag{11}$$

Affiliations. Lee [3] measures the internationality of a conference via the affiliation countries of attendees rather than event locations. We introduce a factor to approximate the internationality of a conference audience by the affiliations of its published authors as they are known to the archive. For each event member, given the set of authors with affiliation information $affiliated(v)$, and the set of distinct affiliation countries of its authors' $affiliations(v)$, we compute the internationality of affiliation histories $affil(v)$. The first quotient in Eq. (12) serves as a weight, considering that the set of authors with a known affiliation history $affiliated(v)$ might only constitute a small part of all distinct authors of an event member $authors(v)$. The second quotient then models the actual audience internationality, dividing the distinct locations of the affiliation histories of the event member v by all affiliation locations known to the archive (COUNTRIES) .

$$affil(v) = \frac{|affiliated(v)|}{|authors(v)|} \cdot \frac{|affiliations(v)|}{\text{COUNTRIES}} \tag{12}$$

$$affil(c) = \frac{\sum_{v \in V_{affil}(c)} affil(v)}{|V_{affil}(c)|} \tag{13}$$

$$w_{affil}(c) = 1 + \frac{affil(c)}{max_{affil}C} \tag{14}$$

Citations. Incoming citations, by analogy to incoming links on the web [8], are commonly employed to quantify the scientific impact of a publication. For each v, given its publication year $\text{PUBLISHED}(v)$, we get the number of incoming citations $\text{CITED}(v, y)$ from records published in year y, and $\text{CITED}(y)$ as the total of all incoming citations from the publication year y regardless of their target. The set Y_v in Eq. (15) contains all publication years for all citation origins known to the archive for an event member v. Eq. (17) describes the weighted average of the number of incoming citations across Y_v for a given event member v and each publication year $y_i \in Y_v$, relative to the number of papers in v and to the total of incoming citations in y_i; it uses the normalized weights described in Eq. (16). With the oldest publication year given by $y_1 \in Y_v$, incoming citations receive higher weights the closer in time the publication year of their origin is to the event year of their target. Thus, we consider power-law effects immanent to online networks: Similar to the influence of older, established, and frequently linked web documents, older, well-known events potentially accumulate more citations than more recent events [2].

$$Y_v = \{y \mid \exists t : t \ \textit{cites}\ v,\ \text{PUBLISHED}(t) = y\},\ |Y_v| = m \tag{15}$$

$$\sum_{i=1}^{m} w_i = 1,\ w_i = \frac{m - i + 1}{\sum_{i=1}^{m} i} \tag{16}$$

$$\textit{cite}(v) = \sum_{i=1}^{m} w_i \ \frac{\text{CITED}(v, y_i)}{|\textit{papers}(v)|\ \text{CITED}(y_i)} \tag{17}$$

$$\textit{cite}(c) = \frac{\sum_{v \in V_{\textit{cite}}(c)} \textit{cite}(v)}{|V_{\textit{cite}}(c)|} \tag{18}$$

$$w_{\textit{cite}}(c) = 1 + \frac{\textit{cite}(c)}{\textit{max}_{\textit{cite}} C} \tag{19}$$

Author Prominence. The total number of published works of an author a is considered an indicator of their prominence in this scientific field. In Eq. (20) the number of publications known to an archive $|\textit{papers}(a)|$ is summed for all distinct authors of an event member $\textit{authors}(v)$ and put in proportion to the number of distinct authors. The average prominence value of all event members of a conference in Eq. (21) forms the basis of the prominence scoring factor.

$$\textit{prom}(v) = \frac{\sum_{a \in \textit{authors}(v)} |\textit{papers}(a)|}{|\textit{authors}(v)|} \tag{20}$$

$$\textit{prom}(c) = \frac{\sum_{v \in V_{\textit{prom}}(c)} \textit{prom}(v)}{|V_{\textit{prom}}(c)|} \tag{21}$$

$$w_{\textit{prom}}(c) = 1 + \frac{\textit{prom}(c)}{\textit{max}_{\textit{prom}} C} \tag{22}$$

3 Case Study on dblp Conference Updates

3.1 Data Sets

Most of the functions described in Section 2 are based on fields of dblp records. Others rely on external data sets. Both types of data sources are described in this section. The subset of covered conferences varies among the computed weighting factors since the required data is sometimes not available for all conferences. If there is no data available for some conference in a specific factor, it contributes the neutral weight 1 to the combined score of that conference. Augmenting metadata in any case entails an existing mapping from the identifiers of third-party data to the identifiers inherent in the target literature database. In addition to external ratings and a citation graph, signatures marked up with affiliation locations are integrated into the ranking factors introduced above.

dblp. The data set we use for our case study is the dblp collection[9] as of 2018-12-17. Of more than 4.4 million distinct records in total (excluding author homepages), this data set contains about 40,000 records of proceedings of about 4,600 different conferences to be considered.

Since the relation between conference event dates and some other date is essential to our approach, we need to make sure that this information is available for the records under consideration. Exact date information is available in fewer titles than month and year information. Therefore, we parse only the event year and month values for DATE() from event member titles with simple pattern matching. There are 4,395 conferences in the set that have at least one event member for which these data fields could be parsed. Thus, we only take this subset into account for our analysis. The next important date is the creation date of a record. Whenever a record is modified in dblp, its timestamp is updated. The creation date of a record thus corresponds to its earliest modification date. The date of publication does not necessarily coincide with the date of the event but may be several days or weeks in advance or even distinctly later. We use the creation date as an approximation of the publication date.

Some of the proposed methods rely on aggregation over simple fields, such as keys for conferences, event members, publications, author profiles, or record creation dates, to determine sets of distinct papers or authors and their respective sizes for the citation-, prominence-, and size-related factors. Other, more complex fields of dblp records require parsing for the computation of the factors discussed above. Geographical information[10] for *located*(), for example, are extracted from the title field of proceedings records, if possible. Of all conferences suitable for evaluation, 4,000 have at least one country information available (91%). A size score > 0 is present for 4,153 of the evaluated conference streams (94.4%). Prominence scores are available for 4,149 conferences (94.4%).

Ratings. The rating factor is based on several local, external conference ratings. The rating CORE originates from Australia and its ratings from 2008

[9] `https://doi.org/10.5281/zenodo.3051910`

[10] For parsing geographical information, the Python library geotext (`https://pypi.python.org/pypi/geotext`) by Yaser Martinez Palenzuela was used.

and 2017 have been mapped to dblp. A similar, compatible mapping local to the Brazilian computer science research community [4] has been integrated as well. The integration process is checked for rare instances of disparate conference-substructure modeling. For example, if dblp attributes two separate conference identifiers from one of the ratings to the same conference, the rating values are ignored. The alphabetical rating ranks are mapped to numerical ranks to compute $rate(c)$ for 791 conferences (18%).

Citations. The *Open Academic Graph* (OAG) includes the *Microsoft Academic Graph* (MAG) enhanced by AMiner[11], comprising approx. 166 million records. These records are mapped to dblp based on their DOIs, if possible, falling back to matching the record titles otherwise. About 3.1 million incoming citation edges from this graph have their citation targets and origins in dblp. This set of incoming citations is used to compute $\text{CITED}(v, y_i)$ and $\text{CITED}(y_i)$ above. With the OAG as of 2017-06-09, a citation-based score is computed for 3,900 evaluated conferences $(88,7\%)$.

Affiliations. To mark up the author-publication links in dblp with the country labels corresponding to the author's affiliation location at that time, integrated data sets from OAG as of 2017-06-09 and a set of institutions derived from Wikidata are used. The institution country labels from Wikidata are matched based on unnormalized affiliation strings of publication authors in MAG and sanitized by checking if the parsed city exists in the parsed country. The country labels are attributed to 3.3 million signatures in dblp based on the DOIs and titles of the publications. For each conference event, the ratio of distinct country labels which distinct event authors have had up to the year of aggregation to the number of distinct country labels in the entire data set is computed, resulting in affiliation-based scores for 3,734 evaluated conferences (85%).

3.2 Detailed Example

In this section, we provide an example in the form of a case study on how to calculate a ranking score for a specific conference at a given point in time. The example describes the components from the perspective of their use in archive curation: How do the individual factors incorporate latent features for each scoring? How can they assist curators in differentiating the various relevant aspects of the archive update process? In the following, a comparison of five different conferences answers these questions while also illustrating how the applied models cope with arising difficulties and necessary adaptations. We define NOW as 2018-07.

In Table 1, the estimation of the next event entry as well as the delay, in relation to an evaluation date, for the International Conference on Data Engineering (ICDE, unique stream key: conf/icde) is exemplified. It appears to be an annual conference ($\delta_{\text{EVENT}}(c) = 12M$) with records being added to the archive with a rather small delay ($\delta_{\text{INDEXED}}(c) = 2M$). The expected next entry of 2018 ($\text{DATE}(e_{n+1}) = 2018\text{-}06$) is one month overdue.

[11] `https://aminer.org/open-academic-graph`

Function	Input / Body	Output
$limited_6(c)$	$E(c)$	$\{2017\text{-}04, 2016\text{-}05, \dots\}$
$\delta_{\text{EVENT}}(c)$	$\{11, 13, 12, 12, 12\}$	$12M$
$\delta_{\text{INDEXED}}(c)$	$\{1, 1, 1, 2, 2, 1, 5, 2, 3\}$	$2M$
$\text{DATE}(e_{n+1})$	$2017\text{-}04 + 12M + 2M$	2018-06
$delay(c)$	$2018\text{-}07 - 2018\text{-}06$	$1M$

Table 1: Delay-Score Example of ICDE with NOW = 2018-07

In Table 2, raw values and final scores for each factor defined above are demonstrated in comparison to four other conferences – the Joint Conference on Digital Libraries (JCDL), the conference on Automata Theory and Formal Languages (Automata), the International Conference on Web-Age Information Management (WAIM), and the Symposium on Principles of Database Systems (PODS). Since the year of the date of consideration is 2018, all conference records that have been added to dblp up to and including 2017 are taken into account. Values for weighting factors are also calculated using data available up to this point in time.

One can see that all conferences share the same delay of one month, as all of them were expected to be updated in June 2018. Thus, the base score is the same for all of them[12].

Examining the differences in scores for the proposed ranking factors exemplifies how they will be ranked in relation to each other in different settings.

First of all, there is no big difference in the activity factor between the conferences, since all of them are active in a sense where records are added regularly to the archive. ICDE gets the lowest score because the last indexed event has already taken place 15 months before NOW, whereas for WAIM it is only 12 months.

A rating-based ranking will put ICDE and PODS slightly higher than JCDL as they received one more A^* rating in the past. WAIM has a much lower score here since it has received only one C-level rating. There is no rating data available at all for Automata; thus $w_{rate}(c)$ is 1.0, the lowest score in this example. ICDE also scores highest in size with over 100 papers per proceeding on average, while Automata has only about 21 papers per proceeding. On the other hand, Automata gets the highest citation score in our example since papers from this conference are being cited more than the others.

When it comes to the internationality of a conference, our example highlights that this notion depends on its definition. When defined in terms of diversity of event locations, Automata outscores all the other conferences, with all eight event venues having taken place in different countries. In this setting, WAIM receives the lowest score, as all conference events of the past have taken place in one country. On the other hand, when looking at the affiliation countries of

[12] To break ties, conferences with the same score are sorted alphabetically.

c	conf/icde	conf/jcdl	conf/automata	conf/waim	conf/pods
$delay(c)$	$1M$	$1M$	$1M$	$1M$	$1M$
$\boldsymbol{w_{delay}(c)}$	**1.5**	**1.5**	**1.5**	**1.5**	**1.5**
$\delta_{\text{EVENT}}(c)$	12	12	12	12	12
$\mathsf{NOW} - \text{DATE}(e_n)$	15	13	13	12	14
$age(c)$	1.250	$1.08\overline{3}$	$1.08\overline{3}$	1.0	1.167
$\boldsymbol{w_{active}(c)}$	**1.390**	**1.460**	**1.460**	**1.5**	**1.424**
$rated_{abc}(c)$	$\{A, A^*, A^*\}$	$\{A, A^*\}$	$\{\}$	$\{C\}$	$\{A, A^*, A^*\}$
$rated_{num}(c)$	$\{3, 4, 4\}$	$\{3, 4\}$	$\{\}$	$\{1\}$	$\{3, 4, 4\}$
$rate(c)$	$3.\overline{6}$	3.5	0.0	1.0	$3.\overline{6}$
$\boldsymbol{w_{rate}(c)}$	$\mathbf{1.91\overline{6}}$	**1.875**	**1.0**	**1.250**	$\mathbf{1.91\overline{6}}$
$\lvert papers(v)\rvert$	5004	1821	214	1554	1182
$\lvert V_{size}(c)\rvert$	48	24	10	29	36
$size(c)$	104.250	75.875	21.400	53.586	$32.8\overline{3}$
$max_{size}C$			1654.16		
$\boldsymbol{w_{size}(c)}$	**1.063**	**1.046**	**1.013**	**1.032**	**1.020**
$\lvert locations(c)\rvert$	13	5	8	1	7
$\lvert located(c)\rvert$	47	23	8	28	36
$intl(c)$	0.277	0.217	1.0	0.036	0.194
$max_{intl}C$			1.0		
$\boldsymbol{w_{intl}(c)}$	**1.277**	**1.217**	**2.0**	**1.036**	**1.194**
$affil(c)$	0.0152	0.0255	0.010	0.014	0.014
$\boldsymbol{w_{affil}(c)}$	**1.116**	**1.195**	**1.074**	**1.110**	**1.108**
$cite(c)$	$4.566E-6$	$8.731E-7$	$2.359E-5$	$3.991E-7$	$1.605E-5$
$max_{cite}C$			$5.708E-4$		
$\boldsymbol{w_{cite}(c)}$	**1.008**	**1.002**	**1.041**	**1.001**	**1.028**
$prom(c)$	57.434	33.289	26.200	55.733	76.931
$max_{prom}C$			102.434		
$\boldsymbol{w_{prom}(c)}$	**1.561**	**1.325**	**1.256**	**1.544**	**1.751**

Table 2: Score examples for five different conferences with $\mathsf{NOW} = 2018\text{-}07$.

publishing authors, JCDL scores highest by far, followed by ICDE and WAIM that also have a rather diverse set of publishing authors in terms of affiliations.

In a scenario where the prominence of publishing authors w. r. t. the archive guides the prioritization of indexing, the PODS conference will be on top of the ranking just before ICDE. JCDL and Automata, on the other hand, appear to have less prominent authors and will thus be ranked lower.

In this case study, we looked at the influence of the different factors by calculating them on a given set of example conferences. This is not a formal evaluation of the ranking factors, and the combination of these factors remains as future work. However, we see the general feasibility and plausibility within this case study, which was the primary concern in this work.

4 Discussion and Future Work

In this paper, we have presented a complex set of ranking factors to support the decision and prioritization process of digital library curators. Building upon our previous work, we have elaborated on the definition of several factors that influence the ordering of the pending metadata updates for conferences. A detailed example in the form of a case study on dblp data demonstrated the effects of each of these factors. It exemplifies how different factors influence the ranking of conferences with the same base score, i. e. that are due at the same time.

Our example also replicates a finding of the bibliometric analysis of Lee [3]: The conference with the fewest number of papers per proceeding got the highest citation score in the example set. While this is no result of a formal evaluation, it still shows the general plausibility of this factor.

While our ranking factors might look very over-specified and very much tailored to dblp, we believe that our work can be of use to other digital libraries and use cases. This is due to the fact that we solely rely on publicly available metadata like MAG, CORE, or Wikidata. The indexing data needed to compute the base delay score should be available in other digital libraries as well.

In this work, we focus on the use case of recommending most urgent conferences to database curators. Of course, these factors can be used in other use cases, such as retrieval tasks when searching for conference-related resources. Another use case might be performing bibliometric studies on conferences, such as finding the most influential or prestigious conferences in a field.

There are some limitations to our work. Data aggregation needed for most of the factors may suffer from flawed metadata, like unresolved ambiguous author names. Even though derived data sets for author-name disambiguation have proven high standards of quality for bibliographies such as dblp [6], errors are unavoidable. And even in the case of dblp, this issue may remain despite several methods of daily on- and offline curation methods from automated tests to consulting human experts. Not only the quality of metadata is a limiting factor, but also the availability of data to calculate the different factors. As most of the used metadata is imported from external sources, this issue might be neglectable

as these data sets are publicly available. Nevertheless, in our case, we could aggregate only about 20% rating information for conferences. The other data had much better coverage, but an exhaustive matching could not be achieved.

Future work consists of the evaluation of our reworked factors against a gold standard built from actual indexing times, as we did in [7]. This way, we will see which of our factors (or which combination) models the relevance decisions of curators best. We also want to focus more on the usefulness of the digital library for the users, e. g. by incorporating latent signals of unfulfilled information needs from the web logs of digital libraries. In the end, this work could lead to a system that produces a ranking based on this model on demand and guides digital library curators into updating the most pressing archive deficits.

Acknowledgments

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG, project no. 217852844).

References

1. Agosti, M., Crivellari, F., Nunzio, G.M.D.: Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Min. Knowl. Discov. **24**(3), 663–696 (2012). https://doi.org/10.1007/s10618-011-0228-8
2. Dai, N., Davison, B.D.: Freshness matters: In flowers, food, and web authority. In: Proceeding of the 33rd SIGIR 2010, Geneva, Switzerland, July 19-23, 2010. pp. 114–121. ACM (2010). https://doi.org/10.1145/1835449.1835471
3. Lee, D.H.: Predictive power of conference-related factors on citation rates of conference papers. Scientometrics **118**, 281–304 (2018). https://doi.org/10.1007/s11192-018-2943-z
4. Martins, W.S., Gonçalves, M.A., Laender, A.H.F., Pappa, G.L.: Learning to assess the quality of scientific conferences: a case study in computer science. In: Proceedings of the JCDL 2009, Austin, TX, USA, June 15-19, 2009. pp. 193–202. ACM (2009). https://doi.org/10.1145/1555400.1555431
5. Michels, C., Fu, J.: Systematic analysis of coverage and usage of conference proceedings in web of science. Scientometrics **100**, 307–327 (2014). https://doi.org/10.1007/s11192-014-1309-4
6. Müller, M., Reitz, F., Roy, N.: Data sets for author name disambiguation: an empirical analysis and a new resource. Scientometrics **111**(3), 1467–1500 (2017). https://doi.org/10.1007/s11192-017-2363-5
7. Neumann, M., Michels, C., Schaer, P., Schenkel, R.: Prioritizing and scheduling conferences for metadata harvesting in dblp. In: Proceedings of the 18th JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018. pp. 45–48. ACM (2018). https://doi.org/10.1145/3197026.3197069
8. Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: Proceedings of the 23rd SIGIR 2000, Athens, Greece, July 24-28, 2000. pp. 288–295. ACM (2000). https://doi.org/10.1145/345508.345602