The References of References: Enriching Library Catalogs via Domain-Specific Reference Mining.

Giovanni Colavizza, Matteo Romanello, and Frédéric Kaplan

École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. The advent of large-scale citation services has greatly impacted the retrieval of scientific information for several domains of research. The Humanities have largely remained outside of this shift despite their increasing reliance on digital means for information seeking. Given that publications in the Humanities probably have a longer than average life-span, mainly due to the importance of monographs in the field, we propose to use domain-specific reference monographs to bootstrap the enrichment of library catalogs with citation data. We exemplify our approach using a corpus of reference monographs on the history of Venice and extracting the network of publications they refer to. Preliminary results show that on average only 7% of extracted references are made to publications already within such corpus, therefore suggesting that reference monographs are effective hubs for the retrieval of further resources within the domain.

Keywords: Bibliometrics, Citation Extraction, Information Retrieval, History of Venice.

1 Introduction

The Humanities are the Cinderella of sciences with respect to citation-driven information retrieval. The lack of citation data not only prevents the quantitative analysis of the field's communication practices (2), but hinders the daily work of researchers, for whom the manual lookup of reference lists is still the only reliable way to collect the state of the art on a topic of interest. Library sections dedicated to domain-specific reference works are an important component of any Humanities research library, whose selection is mainly done by librarians and domain experts.

There are several reasons for this state of affairs, yet the lack of citation data is the most renown problem, lamented several times over (5; 8; 14). For these and other reasons the use of citations as a means to evaluate research in the Humanities has also been questioned (15), with alternatives being proposed (4; 10). Coverage of services such as Web of Science and Scopus is still far from satisfying, albeit improving over time (11), both for journals (12) and monographs (20). Monographs are especially important as the practice in the Humanities still favors them over other kind of publications in order to get recognition within the field (17).

The extraction of references to scholarly publications is instead a widely developed area of research. Recent developments include fully fledged architectures to extract and use citation data resulting into open digital libraries (18). Several reference extraction services exist, such as ParsCit(3), BILBO[1], GROBID(9) and FreeCite[2], but none is unfortunately flexible enough to work with our requirements.

Referencing in the Humanities is a less standardized practice than in other sciences. More specifically, reference lists at the end of a publication are optional, as references are commonly made in footnotes. Furthermore, humanists developed elaborated practices for the abbreviation and encoding of references, which also entail making use of formatting features such as italics or variations in type module. Lastly, it is common in the Humanities to refer to both primary and secondary sources. A primary source is a documentary evidence used to support a claim, a secondary source is a scholarly publication (16). The necessity to cope with these issues meant we could not re-use existing tools for reference extraction.

We propose to leverage domain-specific reference works as a means to enrich library catalogs with citation data. Such reference works can be identified by their physical location in the library (consultation shelves), catalog subjects or scholarly bibliographies. In order to demonstrate the viability of our proposal, we (i) report on the development of a pipeline for the extraction and look-up of references into the library catalog; and we (ii) show that a set of reference monographs individuated using library finding aids might be an effective hub to most of the relevant scholarship within a domain of interest.

This paper is organized as follows: section 2 presents our approach, section 3 presents our preliminary results and section 4 concludes the paper.

2 Approach

We sketch here the main steps of our approach, as illustrated in Fig. 1. The corpus of reference works is first selected, then digitized and OCRed. The manual annotation of a subset of references is then followed by their automatic extraction over the whole corpus. Finally, the look-up module finds matches into the library catalog, and connects paired resources (cited with citing monographs, or co-cited monographs). Eventually, a second look-up module is introduced in order to evaluate the cohesiveness of the selected corpus, defined as the fraction of references made to resources within the corpus over the total extracted resources.

2.1 Corpus Selection

Library catalogs provides a first means to identify all the publications of interest within a domain of study. We queried the catalog in several ways, in order to extract:

[1] http://bilbo.hypotheses.org/
[2] http://freecite.library.brown.edu/.

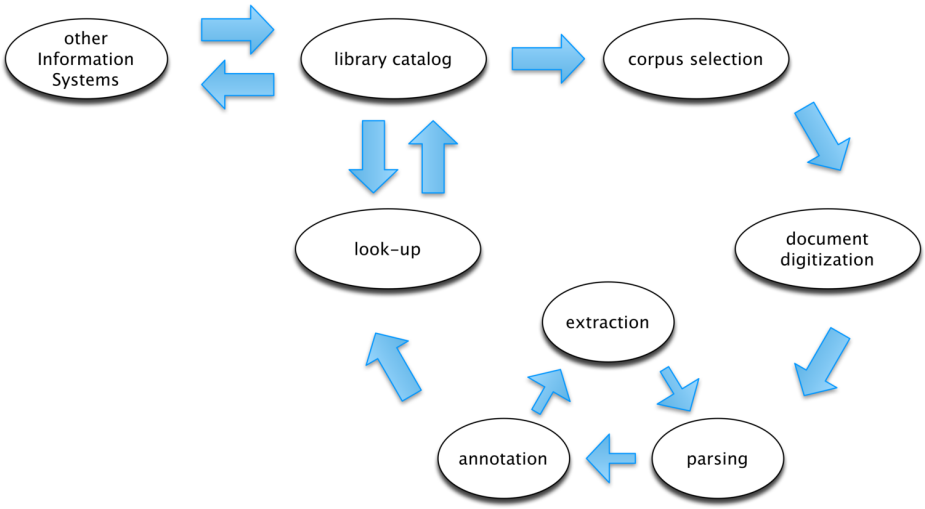

Fig. 1. The proposed pipeline for the enrichment of library catalogs with domain-specific citation data.

1. all the resources in the consultation shelves devoted to the History of Venice;
2. all the resources under subject History of Venice (e.g. Dewey code 945.31);
3. expand by keyword search over the title (e.g. using words as "Venice" in multiple languages) and by using scholarly bibliographies such as Zordan's repertory (19).

The outcome is a set of 1904 monographs and 10 journals. The number of monographs with a list of references is 836 (201 in consultation, cat. 1), 44% of the corpus of monographs, equally distributed over time as shown in Fig. 1. Of these, 701 (184 cat. 1) have structured lists of references, as opposed to end notes, which have been manually individuated and will be used in what follows.

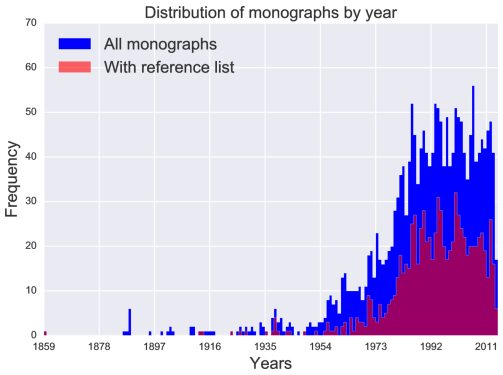

Fig. 2. Number of monographs in the corpus per year (blue/black), over the monographs with a reference list (red/grey): reference lists are equally distributed over time.

The second step in our pipeline is the classification of reference styles and the manual annotation of a sub-set of references for each individuated

class. A reference style is a specific combination of elements in a reference, such as author and title, encoded in a predefined way (e.g. using quotations for the title). We grouped styles in *classes* and *families*. For example:

> "De Virine, Theodore Low. Notable Printers of Italy during the Fifteenth Century. New York: The Grolier Club, 1910."

Is a reference presenting the author's surname, then name separated by comma, title, place of publication, publisher and date. The punctuation and capital letters in use are particularly relevant. A different class stems from the elimination of at maximum one element, or one change in encoding. E.g. removing the publisher would create a new class of the same family. A different family is identified by at least two removals or additions of elements, and/or sensible changes in the encoding of the same information. For example:

> "De Virine, T. L. Notable Printers of Italy during the Fifteenth Century. The Grolier Club, 1910."

Would stem a different class in a separate family as the author's name is now abbreviated and the publisher has been dropped. Classes and their families are useful as a feature for parsing since the references from a specific publication all belong to a unique class/family combination.

In total we individuated 33 classes and 6 families.

Manual annotation was then done over a set of references for each class.[3] Annotations are distinguished into two categories: *generic* and *specific*. A generic annotation distinguishes the completeness of a reference (if full or abbreviated) and the type of referred object (if a monograph or a contribution, such as a journal article). Specific annotations identify the components of generic categories. Examples of specific annotation tags are: "author", "title", "publisher".

Approximately 27% of the 701 monographs have been annotated, 2 pages of references each on average. As a consequence, circa 3.8% of all available pages with references have been annotated. We total 49580 annotations, of which 8646 are generic (i.e. full references) and 40934 specific (i.e. their components).

2.2 Reference Extraction and Parsing

The following component of the pipeline is a parser and reference extraction module, which performs two tasks:

1. Reference parsing: given a text stream of lists of references, parse the text to assign the most likely specific tag to each token.
2. Reference extraction and categorization: given a stream of tokens with specific tags, decide where a reference begins and ends, and assign a generic category to the reference ("monograph", "abbreviated" reference and "contribution").

[3] Using the Brat annotation environment available at http://brat.nlplab.org/.

Both parsers use Conditional Random Fields with the same set of features—except for specific tags resulting from task 1 that are used in task 2—a technique commonly adopted for similar tasks, introduced by (7). The order of the tasks has been determined empirically to maximize performance on a sub-set of specific tags (crucially author, title and year of publication): the most relevant for the look-up module. We used 8051 annotated references for training and testing, for a total of 122612 tokens, or circa 15 tokens per reference, plus 35124 negative tokens (outside of references).

2.3 Catalogue Look-up

Extracted references need to be disambiguated in order to be used in the catalog. This task is performed by a look-up system that tries to match the components of the extracted reference against a bibliographic database, e.g. a library catalog.

Given the nature of the data at hand, such look-up system had to: have a good coverage of the domain; have the ability to work with a limited set of metadata fields as input; and have a degree of tolerance for errors from OCR.

The solution we implemented builds upon the unofficial API of the electronic library catalog of the Italian library system (SBN).[4] This API provides a good coverage of the publications within our monographs, which can be easily explained in light of the focus of our materials on the history of Venice.

3 Preliminary Results

We present results for the two main components of our pipeline: the module for reference parsing and extraction, and the look-up. We also briefly discuss the cohesiveness of our corpus, and its usefulness as a hub for finding further resources within the domain.

3.1 Reference Parsing and Extraction

This module performs the following: given a stream of text likely to contain a list of references, it initially tags every token with specific tags. A second model then parses the text again in order to attribute generic and begin-end tags at the same time. Eventually, all individuated references for each monograph are exported for the look-up module. We do all our implementation in Python, using the CRFSuite (13). The set of features includes[5]:

– The token, itself lowercase, its position in the line, its shape and type, according to a set of predefined classes (e.g. for shape: “UUDDDD” for “AD1900” meaning two uppercase characters and four digits. For classes, in this case we would have “AllUpperDigits”, “InitUpper”).

[4] For a description of the API see `http://literarymachin.es/sbn-json-api/`.
[5] The full list of features is available upon request.

- Suffixes and Prefixes from 1 to 4 characters included.
- A set of indicator features, for example: if the token contains two digits, if four digits, if it could be an abbreviation or contain Roman numbers, etc.
- The reference style category (unique combination of class and family).
- The specific token tag, only for model 2.

For both models we began by keeping a validation set of 25% of references on a side, on which we base our final evaluations. We then experimented with cross-validation on the remaining 75%, in order to find the best parameters and combinations of training approaches. We tested: 1- reducing the features by removing the token and its lowercase version, plus all suffixes and prefixes; 2- removing references to primary sources; 3- training separate models for each family of reference styles; 4- splitting the training data in different sizes (sets of references to parse contiguously); and 5- changing the order of the parsing tasks. Test 2 was positive and kept, test 4 gave us a windows of slices of text containing 5 references as optimal for splitting annotated pages for training. Tests 1 and 5 slightly reduced performance, while test 3 produced overfitted models, probably because of the lack of sufficient and balanced annotated data for every family.

Once the tasks were configured, we searched the parameter space for the best configuration of our CRFs. Using a quasi-Newton gradient descent method (L-BFGS), we have two main parameters: c1 for L1 and c2 for L2 regularizations respectively. Good parameters were found to be:

- Model 1, c1: 0.0289; c2: 0.0546.
- Model 2, c1: 1.53; c2: 0.002.

Intuitively, model 2 benefits from sparse regularization much more than model 1. The result is a set of 181699 references, 8632 of which were part of the golden set and 173067 were newly parsed and extracted.

A 5-fold validation over the whole dataset gives a flat and weighted F1-score of 0.77 and 0.85 for task 1 and 2 respectively, while validation scores on the validation set are summarized in tables 1 and 2, which should be read along with confusion matrices in Figure 3.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0) null	0.679	0.553	0.609	9033
1) pagination	0.900	0.905	0.902	811
2) publisher	0.780	0.688	0.731	1029
3) author	0.847	0.862	0.855	5464
4) title	0.839	0.911	0.873	18834
5) publication number-year	0.772	0.835	0.802	466
6) publication place	0.860	0.873	0.867	1729
7) year	0.882	0.880	0.881	1744
avg / total	0.805	0.812	0.806	39110

Table 1: Extraction results for task 1: parsing.

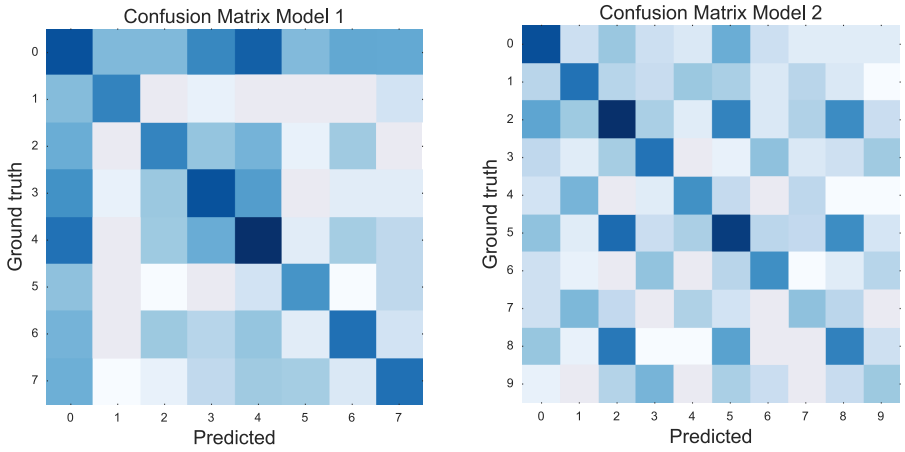

Fig. 3. Confusion matrices for models 1 and 2. Identifiers should be compared with tables 1 and 2 respectively. A darker square means more matches. For example, in matrix 1, the *null* class is very noisy. In matrix 2, errors are consistent with expectations, e.g. *in monograph* mistaken for *in contribution* or *in abbreviated*.

Class	**Precision**	**Recall**	**F1-score**	**Support**
0) out	0.936	0.958	0.947	4815
1) begin monograph	0.846	0.903	0.873	1349
2) in monograph	0.841	0.911	0.874	15683
3) end monograph	0.862	0.894	0.878	1352
4) begin contribution	0.812	0.759	0.785	523
5) in contribution	0.892	0.802	0.845	10930
6) end contribution	0.823	0.820	0.822	523
7) begin abbreviated	0.418	0.266	0.325	192
8) in abbreviated	0.418	0.362	0.388	1963
9) end abbreviated	0.325	0.193	0.242	192
avg / total	0.841	0.845	0.842	37522

Table 2: Extraction results for task 2: extraction and classification.

For Model 1 the main source of errors are null tokens (without tag). Several tags have been removed due to being either under-represented or too varied to be properly captured. This explains the difficulty of the parser to properly fit the null tag. Model 2 instead behaves consistently with the availability of data, meaning the abbreviated references are not as well captured as monographs and contributions. It is nevertheless important to note that begin tags mostly get mistaken for other begin tags, and the same for inside and end tags, all of which are weak errors.

3.2 Catalogue Look-up

The catalog look-up attempts to match the metadata fields of the input extracted reference against the bibliographic metadata accessible via the SBN API. Potential disambiguation candidates are retrieved from the API by performing a search across the whole catalog by using the title of the publication as the search key.

Once some candidates are found, the look-up further expands its search trying to match the input reference against the metadata corresponding fields of each returned catalog record. The number of successfully matched fields is then used to calculate a confidence score, ranging from 0 to 1, which allows us to rank the disambiguation candidates. The confidence score is especially useful in those cases where there are several correct matches for a given input reference. This is often the case given that multiple editions of a given work correspond to multiple records in the catalog.

Several factors needed to be taken into account when performing matching on metadata fields:

1. the title field of catalog records may contain other details in addition to the title (e.g. the editor of a collective volume). To overcome this issue we check whether each token in the title of the reference is also contained in the title of the disambiguation candidate.
2. Names of authors are usually abbreviated in our references, whereas they are given in their full form in catalog records. Therefore, when matching the "author" we strip name initials and try to match on family names.

We carried out a preliminary evaluation of the accuracy of the catalog look-up in order to assess the feasibility of our approach and identify problem areas.

We took a random sample of 2000 references out of the total 181699, equally distributed between manually annotated and automatically extracted references. For each reference we verified whether the look-up result with highest confidence score constitutes a correct disambiguation of the input reference. We considered a result correct also when a different edition of the same work was returned.

The evaluation showed the following results:

- in 41.7% of the cases the look-up does not return any candidate result. Given that the title is used as a search key this issue may be due to at least four different reasons:
  1. the title contains an OCR error that prevents it from matching against the catalog (e.g. "La pittura e la scultura veronese dal secolo Vili al secolo XIII" where "VIII" was wrongly recognized as "Vili");
  2. the title is wrongly segmented or even absent due to a parsing error;
  3. the cited publication is not contained in the catalog;
  4. some words in the title are spelled differently in the reference and in the catalog record (e.g. "Una famiglia veneziana dal X al XIII secolo" as opposed to "una famiglia veneziana dal 10. al 13. secolo").

– in the remaining cases (58.3%), for 72.3% of the references the first disambiguation candidate was correct.

While the precision showed by the look-up is encouraging, improving its recall constitutes a crucial area for further work. A qualitative evaluation of the pipeline at the end of the look-up is currently undergoing in order to individuate all classes of errors and then quantify them.

3.3 Cohesiveness of the Selected Corpus

An important assumption on which our approach rests is the structural hub role of the selected corpus within the citation network of the domain at hand. If the selected corpus is not sufficiently spanning outside of itself, at the same time being well-connected internally (presenting a giant component), then we might find it not effective in order to connect different research areas within the same domain of study. Our assumption was not immediately supported by previous work, which in general highlighted great variability in citation patterns among different disciplines in the Humanities. Co-citation structures by domain and by research themes can both be found (see e.g. (1)), and the proportion of monographs and journal articles is quite varied in different domains (6).

We adapted the lookup module presented in section 3.2 in order to look references up within the corpus itself. The details of this adaptation are beyond the scope of this paper, suffice to say we tuned it to maximize precision in order to avoid miss-matches. The adapted look-up module has been manually evaluated on a small set of 500 extracted references, resulting in a precision score of nearly 1.00 and a recall score above 0.95.[6]

As a reminder, the extracted references of 701 or 37% monographs (of which 184 or 9.7% in consultation, cat. 1) have been matched against the whole corpus of 1904 (100%) selected monographs. We considered only extracted references to monographs, which are 96607 (over the total of 181699).

Firstly, we investigated the cohesiveness of our corpus, defined as the proportion of references inside of the corpus itself, over the extracted total. Results are summarized in Table 3. Overall, only 7% of the extracted references are to monographs within the corpus, slightly more for the monographs in consultation (8%).

Reference Set	**Proportion**	**Matched(Extracted) References**
Consultation, cat. 1	0.0802	1861(21337)
Without cat. 1	0.0669	5398(75270)
All set	0.0699	7259(96607)

Table 3: Citation span of the elected corpus: most of the references are to the outside.

[6] These high scores should not be taken as final: the evaluation was carried out on extracted references, which have errors from previous steps. The evaluation of the whole pipeline from the beginning to the end is still ongoing.

Secondly, we investigated the connectedness of the corpus itself by the extracted references. This is equivalent to the proportion of monographs from the corpus which are in the giant component of the co-citation network resulting from the look-up procedure. The giant component is well-individuated and comprises circa 59% of the corpus. The coverage drops to 32.5% using only the 184 monographs in consultation.

These preliminary results suggest that most of the selected corpus could be useful as a collection of hubs pointing to the relevant literature in the domain, also being strongly connected internally.

4 Conclusions and Future Work

We proposed to use a selection of domain-specific reference monographs in order to enrich Humanities library catalogs with citation data. Our main goal is to allow users to rapidly find the most relevant publications on a topic of interest. Our contribution in this respect is twofold. Firstly, we developed a robust pipeline for the extraction and disambiguation of references contained within publications from the Humanities, evaluating it on a dataset from the domain of the History of Venice. This system constitutes the first necessary step towards the envisaged enrichment of library catalogs. Secondly, we briefly investigated the citation structure of the same dataset in order to assess how effective it may be in serving as a hub to access the domain literature. We found that only 7% of the references made from such corpus are to monographs already within the corpus, suggesting that a wide span over the literature might be achieved from a limited set of selected reference works.

The work presented in this paper constitutes but one aspect of our project. Other aspects that are currently being developed are 1) the extraction of references from footnotes contained in journal articles and 2) the extraction of references to primary sources, such as archival documents, that are often found in publications on the history of Venice. The latter, in particular, will allow us to transform those references into links to archival information systems.

Acknowledgments

We thank Martina Babetto and Silvia Ferronato for the annotation work. The Library of the Ca' Foscari University of Venice for collaborating with bibliographical resources and logistics support. This project is funded by the Swiss National Fund under Division II, project number 205121_159961.

Bibliography

[1] Ahlgren, P., Pagin, P., Persson, O., Svedberg, M.: Bibliometric analysis of two subdomains in philosophy: free will and sorites. Scientometrics 103, 47–73 (2015)

[2] Ardanuy, J.: Sixty years of citation analysis studies in the humanities (1951-2010). Journal of the American Society for Information Science and Technology 64(8), 1751–1755 (2013)

[3] Councill, I.G., Giles, C.L., Kan, M.Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In: LREC (2008)

[4] Hammarfelt, B.: Using altmetrics for assessing research impact in the humanities. Scientometrics 101(2), 1419–1430 (2014)

[5] Heinzkill, R.: Characteristics of references in selected scholarly English literary journals. The Library Quarterly pp. 352–365 (1980)

[6] Knievel, J.E., Kellsey, C.: Citation Analysis for Collection Development: A Comparative Study of Eight Humanities Fields. The Library Quarterly: Information, Community, Policy 75(2), 142–168 (2005)

[7] Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML pp. 282–289 (2001)

[8] Linmans, A.J.M.: Why with bibliometrics the Humanities does not need to be the weakest link: Indicators for research evaluation based on citations, library holdings, and productivity measures. Scientometrics 83(2), 337–354 (2009)

[9] Lopez, P.: GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Research and Advanced Technology for Digital Libraries, pp. 473–474. Springer (2009)

[10] Marchi, M.D., Lorenzetti, E.: Measuring the impact of scholarly journals in the humanities field. Scientometrics 106(1), 253–261 (2015)

[11] Mingers, J., Leydesdorff, L.: A review of theory and practice in scientometrics. European Journal of Operational Research 246(1), 1–19 (2015)

[12] Mongeon, P., Paul-Hus, A.: The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106(1), 213–228 (2015)

[13] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), `www.chokkan.org/software/crfsuite`

[14] Sula, C.A., Miller, M.: Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. Literary and Linguistic Computing 29(3), 452–464 (2014)

[15] Thelwall, M., Delgado, M.M.: Arts and humanities research evaluation: no metrics please just data. Journal of Documentation 71(4), 817–833 (2015)

[16] Wiberley Jr, S.E.: Humanities Literatures and Their Users. In: Encyclopedia of Library and Information Sciences. pp. 2197–2204 (2010)

[17] Williams, P., Stevenson, I., Nicholas, D., Watkinson, A., Rowlands, I.: The role and future of the monograph in arts and humanities research. Aslib Proceedings 61(1), 67–82 (2009)
[18] Wu, J., Williams, K., Chen, H.H., Khabsa, M., Caragea, C., Ororbia, A., Jordan, D., Giles, C.L.: Citeseerx: Ai in a digital library search engine. In: Innovative Applications of AI Conference (2014)
[19] Zordan, G.: Repertorio di storiografia veneziana : testi e studi. Il Poligrafo, Padova (1998)
[20] Zuccala, A., Guns, R., Cornacchia, R., Bod, R.: Can we rank scholarly book publishers? A bibliometric experiment with the field of history. Journal of the Association for Information Science and Technology 66(7), 1333–1347 (2014)