Integrating and exploiting public metadata sources in a bibliographic information system

Extended Abstract

Ralf Schenkel

Trier University, Trier, Germany
schenkel@uni-trier.de
https://orcid.org/0000-0001-5379-5191

Abstract. Bibliographic information systems need to rely on metadata provided by various sources in various forms and with various quality. The talk gives some insights how the dblp bibliography as an example for such a system is maintained and improved. It shows how metadata can be automatically harvested from publisher websites and how the harvesting process can be steered. It also discusses some open sources of bibliographic metadata and how they can be used to enrich existing bibliographic data, but also how varying their quality is.

1 Introduction

Bibliographic information systems are a valuable source for searching, exploring and accessing scientific publications, but also for assessing scientific performance and impact of individual researchers, institutions and publication venues. Important examples for systems include commercial providers like Google Scholar and Microsoft Academic, academic initiatives like CiteSeer[1], ResearchGate[2], and Bibsonomy[3], publisher portals like the ACM Digital Library[4] and Elsevier Scopus[5], and domain-specific portals like PubMed[6], Semantic Scholar[7], and dblp[8]. These systems often differ in the type and volume of metadata provided, in the services they offer on-top of the metadata, and in the scientific domains they cover. For example, the dblp bibliography focuses on providing metadata provided by publishers or collected from public sources for different types of publications in aggregated form, i.e., grouped by authors or venues. dblp does not provide further information like abstracts or citations that can only be extracted from publisher Web sites or even from the full-text of publications. ResearchGate

[1] http://citeseerx.ist.psu.edu/index
[2] https://www.researchgate.net/
[3] https://www.bibsonomy.org/
[4] https://dl.acm.org/
[5] https://www.elsevier.com/solutions/scopus
[6] https://www.ncbi.nlm.nih.gov/pubmed/
[7] https://www.semanticscholar.org/
[8] https://dblp.org/

and Bibsonomy provide data contributed by their users; Google Scholar, MS Academic, and Semantic Scholar collect publications and the associated metadata on the Web and extract further information from the full-text; the ACM DL and Elsevier Scopus combine publisher-provided metadata with additional information extracted from the full-text.

Building and maintaining such a bibliographic information system incurs a number of difficult challenges: Publication venues to be included need to be selected and monitored on a regular basis for new publications; information on publications needs to be extracted from their original source, often a Web site; authors need to be disambiguated. Some of these steps can be supported by additional sources of metadata like ORCID. We will now sketch how dblp solves some of these challenges. We will also highlight how citation information may become available in the future.

2 Monitoring, Selecting and Prioritizing Venues

The completeness of a bibliographic information system is clearly an issue. There is a huge number of publication venues today that has clearly outgrown the capacity of metadata providers. As dblp attempts to provide a certain level of data quality that requires at least a partial manual check, it is impossible to cover all published works. Thus, the decision to include a venue in dblp is made based on a certain set of minimal requirements, documented at `https://dblp.org/faq/`. These requirements include a certain prominence within computer science of people involved in the venue, a certain level of quality control, and a long-time availability of the corresponding publications. The high diversity of publication types makes this decision even more difficult – for example, some technical reports are cited very frequently and thus seem to be very important, however, they usually have neither gone through any academic quality checking nor, especially for commercial technical reports, are not provided in a way that would ensure their long-term availability.

Once a publication venue like a conference or a journal is selected, it is important that new publications appearing at this venue are included with short delay. For some of the big publishers, dblp receives regular updates in form of structured metadata on new publications. For other publishers and even for some important conferences, this information needs to be crawled and extracted from the corresponding Web sites. To reduce overhead, monitoring of these sources is done based on a schedule that takes into account, among other things, the typical publication frequency; Neumann et al. [5] recently discussed possible prioritization schemes that include features like author prominence, delay since the publication appeared, and citation frequency; similar heuristics are used in dblp. The actual extraction is then done either with hand-written wrappers or, more recently, with wrappers written in OXPath [2, 4] that turned out to be more robust against changes of the underlying HTML design.

3 Author Disambiguation

A critical problem for the data quality of a bibliographic information system is author disambiguation, i.e., deciding which individuals (as opposed to names) have authored which publications. Common problems include synonyms, i.e., different individuals with the same name, and homonyms, i.e., different names for the same individual. A typical example for a synonym is the author name "Wei Wang", which corresponds to more than 130 individuals in dblp[9]; still, more than 1,000 publications have not yet been assigned to one of these individuals. Many automated solutions have been proposed in the literature for this problem [1], taking into account information like the co-author graph, topics and venues of the publications, time of the publication, author affiliations, etc. However, no existing method provides a quality good enough for a production system, and many methods cannot efficiently deal with incrementally growing collections. Thus dblp still relies on manual disambiguation, supported by some automated recommendations.

With the recent success of the ORCID initiative[10], more and more authors of publications are now annotated with their ORCID, which makes disambiguation a trivial task. In addition, the ORCID author profiles can help to identify individuals and their publications. Adding ORCID information, dblp could identify and correct a few thousand author profiles that were not corresponding to individuals; however, as ORCID feeds its profiles partly from other metadata providers, not all information in a profile may be correct, thus again manual observation is required.

4 Citations

An important class of bibliographic meta information is citations. Citations are an important ingredient of many bibliometric measures, and a large number of potential applications like citation recommendation, collaboration recommendation, reviewer recommendation, and venue recommendation, but also author disambiguation rely or can at least benefit from the availability of citation information. For many years, anyone who wanted to make use of citations had to extract them from the full-text of publications, which in turn was not available for most publications, or had to rely on a manual collection. Well-known examples of citation collections created with automated methods are CiteSeer, the Open Academic Graph[11], Microsoft Academic Graph[12] (a snapshot of which is included in the Open Academic Graph), or Semantic Scholar. For all automated methods, the quality of the extracted citations is usually not perfect, and often not all citations of a publication can be extracted and mapped to a reference

[9] https://dblp.org/pers/hd/w/Wang:Wei
[10] https://orcid.org/
[11] https://aminer.org/open-academic-graph
[12] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

collection of publications due to errors already present in the original document, introduced by an intermediate OCR process, or errors in the extraction process. Manual citation collection has the potential to provide data with better quality, but does not scale to large volumes of publications. For dblp, an initial effort was undertaken to manually collect a small number of citations for important publications; however, the manual effort for this was orders of magnitude larger than collecting other metadata of publications, hence this effort was not continued.

The recent Initiative for Open Citations[13] has changed the picture here. With the agreement of many important publishers to openly publish citation information for their publications on Crossref[14], there is for the first time a large bulk of citation information available for further analysis and usage. I4OC reports that today more than 50% of all citations on Crossref are freely available. The OpenCitations initiative, in turn, has made this information available in a format than can be easily processed by applications [6, 7]. As of March 12, 2018, it provides citations for more than 300,000 publications from various domains, with overall more than 12 million citations. Other initiatives, especially the Springer-Nature SciGraph initiative[15], have announced to provide citation along with other publication metadata; however, the latest release from November 2017 does not yet include any citation information.

However, despite these highly useful activities, only a small fraction of citations of computer science publications is freely available today. To better understand this, we made an attempt to collect citations from Crossref for all publications in dblp (as of January 27, 2018). The experiment was done with Crossref data from late February 2018. Of the roughly 4 million publications in dblp, about 3.2 million have a DOI assigned and could thus be looked up with the Crossref API; of these, only about 70,000 were unknown to Crossref. For slightly less than 600,000 publications, Crossref returned citation information, which is slightly less than 20% of the available publications (and less than 15% of all publications in dblp). This is not nearly enough to be useful for an analysis or recommendation. Moreover, of the approximately 16 million citations retrieved, only slightly more than 4 million could be mapped to dblp based on the provided DOI; the other citations either did not provide a DOI or could not be mapped to a paper in dblp. So while the initiative is clearly extremely useful, the citation data available at this time is not yet extensive enough to be useful in a bibliographic information system. Extracting missing citations from the full-text of publications, but also mapping citations without DOI information to a reference collection such as dblp based on their reference text is thus still an important problem.

An interesting research question in this context is estimating when a collection includes enough citation information to provide stable estimates for the impact of authors and publications. Is the citation data now available in Crossref already sufficient, or do we need more citations (and probably citations from

[13] https://i4oc.org/
[14] https://www.crossref.org/
[15] https://www.springernature.com/de/researchers/scigraph

more diverse sources)? Do we need new bibliometric measures that take missing information into account, and that can deal with the inherent uncertainty?

5 Availability

A collection of bibliographic needs to be published under a license that is as open as possible to be useful for scientific purposes. Many of the collections mentioned above are available under such licences. The collection of metadata included in dblp, for example, has been released under the Open Data Commons ODC-BY 1.0 license. A daily updated XML dump can be found at `http://dblp.org/xml/`; the data format is explained in [3]. For repeatable scientific experiments, the persistent monthly snapshots should be used which are available under `http://dblp.org/xml/release`. The dblp example has shown that the use of such a collection goes way beyond standard bibliometric analysis; there are several hundreds of scientific papers that use the dblp data for experiments on large-scale graphs, social networks, author disambiguation, or topic mining.

6 Outlook and Perspectives for Future Work

Recently, the work on dblp has focused on improving the volume of new publications per year, while at the same time improving overall data quality by cleaning author profiles. A current activity focuses on networking with other data providers, including ORCID and Wikidata. Future activities on the instance level will include increasing the coverage of historic publications and monographs (including PhD theses).

From a broader perspective, integrating validated bibliographic information from the diverse providers and from all scientific disciplines and publishing this information under an open licence would clearly be a significant improvement over the status quo. There clearly is a large overlap of authors between computer science and neighbouring disciplines like mathematics and biology, but also seemingly distant disciplines like psychology. On the modelling side, dblp currently lacks a good representation of conference series and individual conference events. This problem is more complex than it may seem as conferences may change publishers, may pause for some time, or may change name (like the WWW conference), or even may merge with others or split. On the data side, an important aspect of future work is indexing scientific data sets, which are now often published at providers like DataCite[16] and need to be made accessible through information providers; ideally, also their their use in scientific literature should be tracked and reported as metadata of the publications.

An interesting long-term goal for data providers would be to collect additional metadata on conferences and journals beyond classical bibliographic metadata. This may include, for example, information on the members of the program committee, organizers in different roles, keynote speakers, etc. This information

[16] `https://www.datacite.org/`

would allow for a better estimation of the reputation of scientists in their corresponding scientific community. Most of this data is already available on the Web, but needs to be collected, cleaned, and completed before it can be useful.

References

1. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Record 41(2), 15–26 (2012), http://doi.acm.org/10.1145/2350036.2350040
2. Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.J.: Oxpath: A language for scalable data extraction, automation, and crawling on the deep web. VLDB J. 22(1), 47–72 (2013), https://doi.org/10.1007/s00778-012-0286-6
3. Ley, M.: DBLP - some lessons learned. PVLDB 2(2), 1493–1500 (2009), http://www.vldb.org/pvldb/2/vldb09-98.pdf
4. Michels, C., Fayzrakhmanov, R.R., Ley, M., Sallinger, E., Schenkel, R.: Oxpath-based data acquisition for dblp. In: 2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017. pp. 319–320 (2017), https://doi.org/10.1109/JCDL.2017.7991609
5. Neumann, M., Michels, C., Schaer, P., Schenkel, R.: Prioritizing and scheduling conferences for metadata harvesting in dblp. In: JCDL (2018)
6. Peroni, S., Dutton, A., Gray, T., Shotton, D.M.: Setting our bibliographic references free: towards open citation data. Journal of Documentation 71(2), 253–277 (2015), https://doi.org/10.1108/JD-12-2013-0166
7. Shotton, D.: Open citations. Nature 502(7471), 295–297 (2013), https://doi.org/10.1038/502295a