Extraction, analysis and publication of bibliographical references within an institutional repository

Götz Hatop
University Library, Philipps University Marburg, Marburg, Germany

Received 7 January 2016
Revised 22 February 2016
15 March 2016
Accepted 6 April 2016

Abstract

Purpose – The academic tradition of adding a reference section with references to cited and otherwise related academic material to an article provides a natural starting point for finding links to other publications. These links can then be published as linked data. Natural language processing technologies are available today that can perform the task of bibliographical reference extraction from text. Publishing references by the means of semantic web technologies is a prerequisite for a broader study and analysis of citations and thus can help to improve academic communication in a general sense. The paper aims to discuss these issues.

Design/methodology/approach – This paper examines the overall workflow required to extract, analyze and semantically publish bibliographical references within an Institutional Repository with the help of open source software components.

Findings – A publication infrastructure where references are available for software agents would enable additional benefits like citation analysis, e.g. the collection of citations of a known paper and the investigation of citation sentiment.The publication of reference information as demonstrated in this article is possible with existing semantic web technologies based on established ontologies and open source software components.

Research limitations/implications – Only a limited number of metadata extraction programs have been considered for performance evaluation and reference extraction was tested for journal articles only, whereas Institutional Repositories usually do contain a large number of other material like monographs. Also, citation analysis is in an experimental state and citation sentiment is currently not published at all. For future work, the problem of distributing reference information between repositories is an important problem that needs to be tackled.

Originality/value – Publishing reference information as linked data are new within the academic publishing domain.

Keywords Digital libraries, Library services, Linked data, Citation analysis, Reference extraction, Semantic publishing

Paper type Research paper

Introduction

The most common purpose of a citation is to show that a referenced work has influenced the author's work. The reference section of an academic paper has its value not only in giving credit to authors whose work was used, but is useful to support assertions and arguments of the author and also helps readers to find more information on the subject and beyound. Besides these reasons, the further development of tools and techniques for citation analysis and other applications like author-topic evolution or co-authorship graph analysis depends on published references.

As a consequence, the publication of reference information by the means of semantic web technologies is of paramount importance.

Unfortunately, because of the amount of work required to manually create machine readable reference information, the publication of this information is almost constantly

Library Hi Tech
Vol. 34 No. 2, 2016
pp. 259-267

0737-8831
DOI 10.1108/LHT-01-2016-0003

neglected. Moreover, automatic reference extraction is a task with high complexity and has only recently seen some progress by the use of machine learning approaches based on Conditional Random Fields (CRF). Groza *et al.* (2012) summarizes the research performed over the last years on automatic metadata extraction from scientific publications.

The process of extracting and semantically publishing references from scholarly work can be divided into the following tasks:

(1) reference extraction;

(2) citation analysis; and

(3) reference publication.

The following sections introduce these processing phases and suitable open source software components in more detail.

The software written in the context of this project utilizes open source reference extraction libraries to find bibliographical references and uses the Apache Jena library to support reference publishing. The software has been published online under the name METABLOCK and is available from github.

Reference extraction

The term "reference" is sometimes used with unclear semantics. Throughout this paper, this term is used exclusively to describe a bibliographic resource as mentioned in the reference section of an article. Such a reference appears in the list of works containing the full bibliographic information such as "Author Name(s)," "Year of Publication," "Title of Work" and "Journal Name" about a cited work.

The bibliographic data from a paper itself like title, author and date of publication is distinct from data within the reference section. The part of text where the work of another author is mentioned is essential for further text analysis. This part of text will not be counted as part of a reference here but be denoted as "citation" or explicitly as "reference context" when necessary.

Due to the vast amount of languages and cultures where scientific research is undertaken, there has never been a single rigid understanding on how a reference to a related academic work should be written down. Therefore, the correct detection and parsing of bibliographical references from scientific papers is an interesting field for natural language processing approaches.

The success of machine learning approaches applied to this task is impressive. Tkaczyk *et al.* (2015) presents an overview about the state of the art in metadata extraction from scientific literature. Two of the more advanced and also open source tools in this field are CERMINE (Tkaczyk *et al.*, 2014) and Grobid (Lopez, 2010).

Both programs make use of CRF and are build upon free mathematical libraries for this approach. CRF is a probabilistic graphical model for classification, an introduction to the mathematical background can be found in Sutton and McCallum (2011).

The processing phases necessary for reference extraction can be described as follows:

(1) reference localization;

(2) reference segmentation; and

(3) reference chunking.

The entire metadata extraction process as well as an evaluation of CERMINE is described in more detail by Tkaczyk *et al.* (2012).

Reference localization

Usually, in an academic document every reference block starts with a keyword, which is different with various languages. Typical words such as "References," "Bibliography," "Works Cited," etc. are used to introduce a reference section. Thus, a very simple approach for reference localization is to start from the end of a paper until one of the keywords can be found. The reliance on keyword alone will mean that there will always be a chance that the program might miss a keyword not present in the database. Another problem is, that larger academic work like monographs are more difficult to process, since the reference section may be followed by appendices or other related material.

Reference segmentation

With reference segmentation, the process of finding individual reference strings from a collection of reference lines is addressed. This splitting can be done by the use of heuristics build upon assumptions on the length of lines or ending punctuations (Kern, 2013).

Reference chunking

Reference chunking represents the process of label sequencing a reference string. The parts of the reference containing the authors, the title, the publication year, etc. needs to be tagged. The presence of abbreviations, inconsistent formatting and semantically overloaded punctuation and separators present difficulties.

Reference extraction in an institutional repository

To get an impression about what is currently possible in this discipline, the reference extraction process was tested with 100 articles in PDF format, mostly from the domain of information science and mathematics. Since the overall aim of the procedure is to find references which can be published as linked data, a reference is counted as successfully extracted, if a complete bibliographical citation could be found without side effects like missing separators between two consecutive references. Errors occur, if two or more references have not been separated by the extraction process or if short sentences like an acknowledgment has incorrectly been marked as a bibliographic description.

The evaluation of such systems is usually done in terms of precision, recall and the F1 metrics for the individual metadata fields considered by the system. With this notion, recall is a measure of completeness and consist in the fraction of correctly detected references among the total sum of references and precision is a measure of fidelity and consists in the fraction of correct results among those that the system believes to belong to the relevant subset.

The F1 measure is calculated as the harmonic mean between precision and recall and is a measure for the accuracy of a classification test (Table I).

Both Grobid and CERMINE are written in Java and actively developed, and due to the ongoing development process the overall success rate of reference extraction should eventually improve over time.

Table I. Reference extraction performance

Program	True positives	False negatives	False positives	Recall	Precision	F1
Grobid	3,047	662	62	0.82	0.98	0.89
CERMINE	2,497	1,212	8	0.67	0.99	0.80

Citation analysis

When using natural language processing technology to investigate citations, the identification of citation context is the first important task to tackle. If a citation ontext can be extracted, it is possible to find out more about the referenced paper. One application that immediately follows from context detection is the structured summarization of a paper based on the citing authors opinion (Tandon and Jain, 2012). Citation sentiment detection is another attractive task as it can help researchers to identify shortcomings and problems of a particular approach.

Citation context detection

A citation context of a referenced work can be defined as the set of sentences about the work in a paper. Such a citation context should contain valuable information about the cited paper due to the high quality of the researchers analysis of the work. However, it is not clear beforehand, as to what extend a citing sentence has to be considered context. The results of citation analysis are dependent on the size of context sentences (Athar and Teufel, 2012), and considering context windows of different widths can be useful.

Qazvinian and Radev (2010) distinguish between explicit and implicit citations, where an explicit citation is a sentence which explicitly mentions a referenced work and an implicit citation contains information about a work without containing any bibliographical information from the reference list. While explicit citations are relatively easy to find in text, it is much harder to detect implicit citation. Murray (2015) has investigated this subject and released open source software which again uses machine learning technology to detect implicit citations.

However, within this project, the additional complexity of detecting implicit citations has been omitted. Citation context extraction is supported only for explicit references and the referencing sentences are extended in both directions to build a context.

Citation sentiment analysis

If a citation context can be identified in text, it might be possible to determine the authors opinion about the referenced work. Citation sentiment analysis is an attractive task because sentiment information can give better insight of how a research field is structured, as positive citations indicate influence in a way that negative citations clearly do not.

Current approaches for sentiment analysis are based on machine learning technologies and thus do require annotated data for training purposes. Athar (2014) addresses this task and presents a large corpora for citation sentiment analysis.

Sula and Miller (2014) proposes a polar classification of positive and negative citation context for citation sentiment analysis. They describe citation classification of articles from the humanities based on a naive Bayes classifier. The classifier is built upon two training sets, one positive and one negative. The authors report highly successful citation extraction, but limited success on citation polarity detection.

This is broadly consistent with the results from this project. When training a naive Bayes classifier with 75 percent of the data annotated by Athar (2014), the polarity detection on the remaining 25 percent (the test set) is accurate for only 74 percent of the citation sentences. However, the classifier currently only makes use of a simple bag-of-word model and does not try to use more sophisticated feature sets, which might improve the results.

Although it is worth mentioning that detecting citation sentiment is an important step toward a better text analysis, this is currently only implemented within METABLOCK to support further tests on this subject, and sentiment information is not published as linked data.

Reference publishing

The advantages of semantic web technologies are significant with respect to metadata publishing. Publishing bibliographical references together with traditional metadata promises to take the linking between academic papers one step further to a level where essential connections between distinct academic work could be understood by machines. In order to enable linked data applications to discover datasets as well as to ease the integration of data from multiple sources, linked data publishers should comply with a set of best practices (Heath and Bizer, 2011):

- Vocabulary usage: the best practices advise publishers to use terms from widely-used vocabularies. In the case at hand, the Dublin Core vocabulary (DCTerms) provides the verb "references" for a bibliographic relation, and with this term the statement that one paper references another paper can be expressed.
- Metadata provision: linked data should be self-descriptive and thus include metadata. An important form of metadata is provenance metadata, which refers to the sources of information, such as entities and processes involved in producing or delivering an artifact. This information is crucial for decisions about whether information is trusted, how to integrate diverse information sources, and how to give credit to originators when reusing information. The DCTerms vocabulary reserves the term "provenance" for this information, and METABLOCK inserts information about the metadata extraction program with this term.
- Linking: by setting RDF links, a connection between distributed data can be established. Applications are then in a position to navigate and discover additional information. The basic requirement for such a link is, that an URI can be assigned to an extracted bibliographic record.

Reference identification

As a first step for reference publishing in a linked data scenario, URIs are needed to identify two papers which are connected by a reference. While the paper where the reference is extracted from already has an URL from its hosting repository, this is in general not the case for the referenced paper. Some extracted bibliographical references may already contain an identifier such as a DOI, an URN or an URL which can be used or easily transformed to a http-based URI suitable for publishing RDF-statements with reference information.

The situation is different, when no identifier is available from the extracted reference. In this case, an identifier must be found by other means.

Large bibliographic indexes exists, which are able to resolve a large amount of bibliographic references to an URL. Google Scholar is a prominent example and probably the largest index of academic material, but does not allow machines to work with the index. METABLOCK currently utilizes the BASE search system (Summann and Lossau, 2004) operated by the University Library of Bielefeld to resolve references and uses the Crossref system if BASE can not resolve the record.

Bibliographical records which can not be resolved to a http-based URL may in fact be unresolvable, that is, the reference is about printed material and no digital copy is available. In this case, the bibliographic citation from the reference list itself can be published and used by humans to identify and locate the resource.

Publishing references

A wide range of semantic web tools for the management, processing, visualization and analysis of semantic data is available today, and the existing tools can be used to publish reference information together with traditional bibliographic metadata like title, author and publishing date.

As stated earlier in this paper, the reference extraction process is a starting point for reference publishing. The METABLOCK-program uses the Apache Jena library to produce RDF data from the extracted references and inserts the statements into a triple store, currently supporting Jena's own database format (TDB) and Virtuoso.

Moreover, it is possible to write the RDF data to a Solr search index as defined by the VuFind resource discovery system (Hatop, 2013). This way, the RDF data are available from a SPARQL service endpoint and also as RDF resource description from VuFind's record view page by explicit user request or with content negotiation.

An example for a RDF statement obtained from an extracted reference is shown in Figure 1 (Namespace available at: http://purl.org/dc/terms/).

The Dublin Core references term is intended to be used with non-literal values as domain and range, so a valid linked data statement requires both the source URI (subject) and the target URI (object) of a RDF statement to serve RDF data if required.

Reference propagation

It is worth considering the inverse relation of the Dublin Core references term as a separate matter. The Dublin Core vocabulary reserves the term "isReferencedBy" for this type of statement. The additional complexity of stating that an academic work is referenced by another document stems from the fact that the referenced paper is usually published by a foreign organization. In a linked data scenario, this organization would care for the knowledge base about the paper and maintain RDF-statements about the resource. Thus, an update of a remote knowledge base is required, and the remote side would then know that the work has been cited by someone.

Nevertheless, the linked open data cloud is currently essentially read-only (Ibáñez *et al.*, 2014). As a consequence, writable linked data, and thus the possibilities of collaborative knowledge construction are rather limited at the moment. Remote SPARQL updates are the most obvious choice for adding triples to a foreign repository, but are usually not allowed for data security reasons. Thus, an infrastructure with a notion of trust between repositories would be helpful to overcome this limitation.

In the example use case above a SPARQL service endpoint can be derived from the data by using the void vocabulary. Due to the repository-internal structure of the

```
@prefix dcterms: <http://purl.org/dc/terms/>.
  <http://archiv.ub.uni-marburg.de/diss/z2014/0633>
      dcterms:references
          <http://archiv.ub.uni-marburg.de/diss/z2010/0551>;
```

Figure 1.
RDF statement to express a reference between two articles

statement, trust is not an issue and the inverse relation can thus be published by using the SPARQL update protocol without authentification as shown in Figure 2.

The triple store now contains the information, that the paper was cited (see Figure 3). This information is not only available as RDF-statement for machines, but also shown for human readers in the user interface in a "Cited By"-section of the work.

Although the scenario sketched above allows to transmit reference information between resources, it is restricted to a single repository because of the difficulties with remote update procedures.

Existing research on deploying semantic web technologies has tended to focus on data modeling, and software architecture and engineering issues have been comparatively neglected. However, toward the goal of linking scholarly work based on references, the following challenges can be identified:

- repositories should publish bibliographical metadata as linked data;
- users should be encouraged to use persistent identifiers like URL, DOI or URN in bibliographical references;
- persistent identifiers should be resolvable to a linked data URL; and
- repositories should allow update requests from other repositories.

Perhaps the most difficult requirement for collaborative authoring of reference information in a distributed environment is the last point above, the possibility to update a remote triple store. From an engineering standpoint, remote updates could be replaced by regular harvests of reference information, which in turn would require that reference publishing becomes more common.

Limitations and related work

The focus of this paper is on the overall workflow for reference publishing. This results in some limitations: only a limited number of metadata extraction programs have been considered for performance evaluation and reference extraction was tested for journal articles only, whereas Institutional Repositories usually do contain a large number of other material like monographs. Also, citation analysis is in an experimental state and citation sentiment is currently not published at all.

For future work, the problem of distributing reference information between repositories is an important problem that needs to be tackled. Although the SPARQL protocol does provide the mechanisms for updating remote knowledge bases, this functionality usually requires some sort of authentification and thus can not be done in a

```
PREFIX dcterms: <http://purl.org/dc/terms/>
INSERT DATA { GRAPH <http://archiv.ub.uni-marburg.de> {
  <http://archiv.ub.uni-marburg.de/diss/z2010/0551>
    dcterms:isReferencedBy
        <http://archiv.ub.uni-marburg.de/diss/z2014/0633> }
}
```

Figure 2.
SPARQL update for referenced papers

```
@prefix dcterms: <http://purl.org/dc/terms/>.
  <http://archiv.ub.uni-marburg.de/diss/z2010/0551>;
      dcterms:isReferencedBy
          <http://archiv.ub.uni-marburg.de/diss/z2014/0633>;
```

Figure 3.
RDF statement to express that a paper was referenced

simple way. The work presented here does not feature this aspect, but it can be regarded as the next key step toward a more collaborative academic publishing infrastructure.

Other work in the field of metadata extraction extends the scope of data that can be extracted from scientific papers, e.g. the extraction of author affiliations as reported by Tkaczyk *et al.* (2015). These data can be semantically published as well, and could eventually provide further links between academic work based on author affiliation.

The DCTerms ontology 1 used throughout this work provides a rather broad vocabulary for reference publishing. Although this ontology seems to be appropriate at the moment, this would change if within subsequent work on this topic the extraction of opinions about a referenced paper becomes more successful. The CiTO citation ontology developed by Peroni and Shotton (2012) is already prepared to express different language utterances and could be used to express fine grained statements about a referenced paper that state, whether a work is criticized, confirmed or merely summarized in a discursive context. In this sense, publishing references as described here can be seen as a step toward a more common and practical usage of data publishing technologies in the academic publishing domain and as a preparation to support the deployment of more semantic web technologies.

Conclusion

Although reference extraction from scholarly work is hard, it is not impossible. This paper examines the prospects and limitations of reference extraction and publishing with open source software components. Perfection is currently not possible for automatic extraction, but the achieved results are encouraging and can be used to establish a reference publishing workflow for Institutional Repositories.

A publication infrastructure where references are available for software agents would enable additional benefits like citation analysis, e.g. the collection of citations of a known paper and the investigation of citation sentiment.

The publication of reference information as demonstrated in this article is possible with existing semantic web technologies based on established ontologies and open source software components.

References

Athar, A. (2014), "Sentiment analysis of scientific citations", Technical Report No. UCAM-CL-TR-856, Computer Laboratory, University of Cambridge, Cambridge.

Athar, A. and Teufel, S. (2012), "Context-enhanced citation sentiment detection", *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics 2012*, pp. 597-601.

Groza, T., Grimnes, A. and Handschuh, S. (2012), "Reference information extraction and processing using random conditional fields", *Information Technology and Libraries*, Vol. 31 No. 2, pp. 6-20.

Hatop, G. (2013), "Integrating linked data into discovery", *Code4Lib Journal*, No. 21, available at: http://journal.code4lib.org/articles/8526 (accessed December 12, 2015).

Heath, T. and Bizer, C. (2011), *Linked Data: Evolving the Web into a Global Data Space – Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool, San Rafael, CA.

Ibáñez, L., Skaf-Molli, H., Molli, P. and Corby, O. (2014), "Making linked open data writable with provenance semirings", research report, LINA-University of Nantes, Nantes.

Kern, R. (2013), "Extraction of references using layout and formatting information from scientific articles", *D-Lib Magazine*, Vol. 19 Nos 9/10, available at: www.dlib.org/dlib/september13/kern/09kern.html (accessed May 21, 2015).

Lopez, P. (2010), "Automatic extraction and resolution of bibliographical references in patent documents", in Cunningham, H., Hanbury, A.H. and Rüger, S. (Eds), *Advances in Multidisciplinary Retrieval*, Springer, Berlin and Heidelberg, pp. 120-135.

Murray, J. (2015), "Improved citation context detection methods", available at: https://github.com/JonathanMurray/citation-context-thesis

Peroni, S. and Shotton, D. (2012), "FaBiO and CiTO: ontologies for describing bibliographic resources and citations", *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 17, pp. 33-43, available at: http://dx.doi.org/10.1016/j.websem.2012.08.001 (accessed November 23, 2014).

Qazvinian, V. and Radev, D.R. (2010), "Identifying non-explicit citing sentences for citation based summarization", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 555-564.

Sula, C.A. and Miller, M. (2014), "Citations, contexts, and humanistic discourse: toward automatic extraction and classification", *Literary and Linguistic Computing*, Vol. 29 No. 3, available at: http://doi.org/10.1093/llc/fqu019 (accessed July 22, 2015).

Summann, F. and Lossau, N. (2004), "Search engine technology and digital libraries: moving from theory to practice", *D-Lib Magazine*, Vol.10 No. 9, available at: http://dx.doi.org/10.1045/september2004-lossau (accessed February 8, 2015).

Sutton, C. and McCallum, A. (2011), "An introduction to Conditional Random Fields", available at: http://arxiv.org/abs/1011.4088 (accessed August 27, 2014).

Tandon, N. and Jain, A. (2012), "Citation context sentiment analysis for structured summarization of research papers", *Proceedings of the 35th German Conference on Artificial Intelligence (KI-12), Saarbruecken, September 24*, pp. 98-102.

Tkaczyk, D., Tarnawski, B. and Bolikowski, L. (2015), "Structured affiliations extraction from scientific literature", *D-Lib Magazine*, Vol. 21 Nos 11/12, available at: http://doi.org/10.1045/november2015-tkaczyk (accessed December 1, 2015).

Tkaczyk, D., Bolikowski, Ł., Czeczko, A. and Rusek, K. (2012), "A modular metadata extraction system for born-digital articles", 2012 10th IAPR International Workshop on Document Analysis Systems (DAS), pp. 11-16.

Tkaczyk, D., Szostek, P., Dendek, P.J., Fedoryszak, M. and Bolikowski, L. (2014), "CERMINE – automatic extraction of metadata and references from scientific literature", 2014 11th IAPR International Workshop on Document Analysis Systems (DAS), pp. 217-221.

Further reading

Auer, S., Dietzold, S. and Riechert, T. (2006), "OntoWiki – a tool for social, semantic collaboration", *Lecture Notes in Computer Science*, Vol. 4273, pp. 736-749.

Hatop, G. (2015), "Metablock: semantic publishing tools", available at: http://cloud8.github.io/Metablock