PaperHunter: A System for Exploring Papers and Citation Contexts

Michael Färber[1], Ashwath Sampath[1], and Adam Jatowt[2]

[1]Department of Computer Science, University of Freiburg, Germany
michael.faerber@cs.uni-freiburg.de
ashwath92@gmail.com
[2]Department of Social Informatics, Kyoto University, Japan
adam@kuis.db.kyoto-u.ac.jp

Abstract. In this paper, we present a system that allows researchers to search for papers and in-text citations in a novel way. Specifically, our system allows users to search for the textual contexts in which publications are cited (so-called *citation contexts*), given either the cited paper's title or the cited paper's author name. To better assess the citations qualitatively, our system displays indications about the so-called *citation polarity*, i.e., whether the authors wrote about the cited publication in a positive, neutral, or negative way. Our system is based on all computer science papers from arXiv.org and can be used by computer science researchers to reflect on their appearance within the scientific community as well as by researchers studying citations.

Keywords: scientific papers, bibliometrics, citation context, citation polarity

1 Motivation

Researchers in all scientific fields are nowadays confronted with vast amounts of scientific papers published within a short period [1,2,3]. As a consequence, scientists are increasingly dependent on publication search engines [3], such as Google Scholar, to search – typically via keywords – for the metadata of relevant papers and the papers themselves. However, to our knowledge, no system lets users search the citation relationships between papers directly.

We present such a system here. Besides the usual ability to search for papers' full metadata and papers themselves, it provides the following exclusive features:

1. Given a paper's title, the system lets users search for the text passages in which the paper is cited (i.e., citation contexts). Thus, this search functionality can be used to analyze how the paper is cited in other papers.
2. Given an author's name, the system lets users search for the text passages in which the author's publications are cited. Also, the system provides the metadata for these citing papers. This search functionality allows users to analyze how the community perceives an author.

3. To allow the user to quickly recognize how the citations are used within the text passages, our system presents indications about the so-called *citation polarity*, i.e., whether the authors of the text passage wrote about the cited publication in a positive, neutral, or negative way.

Several user groups can benefit from using our system:

1. *Ordinary researchers* might be interested to know *in which papers, in which contexts, and in which ways they or people in their environment (e.g., colleagues, competitors) are cited*. Having this information will allow researchers to react accordingly. For instance, aspects concerning the cited papers that other researchers have mentioned incorrectly or imprecisely, or that are missing, can be clarified in future papers and in communication between authors. Thus, our search functionalities will allow for more nuanced scientific exchanges between researchers.
2. *Bibliometrics, scientometrics, and social analysis researchers* who focus on analyzing citations and measure the impact of citations can use our search system to gain new insights concerning the usage of citations.
3. *Practitioners*, such as software developers, can use our system to determine how methods and data sets (published via papers) have been used.

Our demonstration system is available online at `http://paperhunter.net`. Also, the source code is available online as open source code.[1]

2 System Design

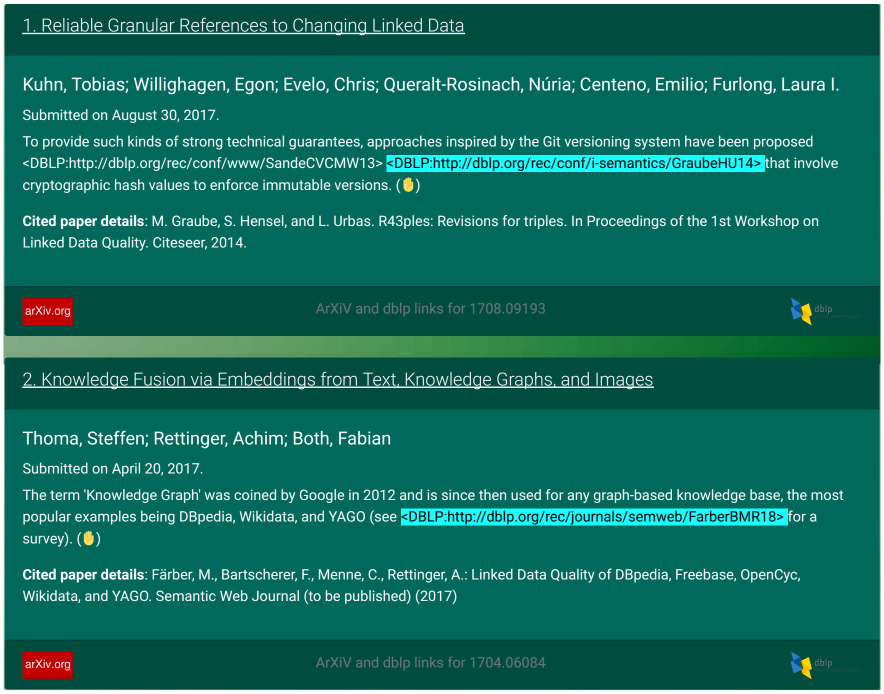
We use the *arXiv CS data set* [4] as our database. This data set contains metadata about all arXiv.org papers in the field of computer science published before the end of 2017 as well as the contents of these publications in plaintext. Overall, this data set contains about 16 million sentences from about 90,000 papers and is said to be one of the few data sets containing the full texts of papers of such a size and cleanliness [4]. Note that in the data set the formulas have been replaced by placeholders and in-text citation markers have been replaced by identifiers. Separate files contain the mappings of these identifiers to the cited papers' metadata (including the authors' names, title, venue, and year).

We index the arXiv CS data set using Apache Solr.[2] When a user searches for papers or citation contexts, our system retrieves all the result items according to the default TF-IDF scoring and then returns the top n results based on the papers' publication date. This is done due to users typically being interested in more recent papers. n is provided by the user (by default, $n = 100$). We use Python to process the data and Django to create the user interface.

In the following, we present the different search capabilities provided by our system. Note that the first three search functionalities can be considered basic functions and are prevalent in existing related works. The last two search possibilities are novel, exclusive search functionalities proposed in this paper.

[1] `https://github.com/michaelfaerber/paperhunter`.

[2] `https://lucene.apache.org/solr/`.



1. Reliable Granular References to Changing Linked Data

Kuhn, Tobias; Willighagen, Egon; Evelo, Chris; Queralt-Rosinach, Núria; Centeno, Emilio; Furlong, Laura I.

Submitted on August 30, 2017.

To provide such kinds of strong technical guarantees, approaches inspired by the Git versioning system have been proposed <DBLP:http://dblp.org/rec/conf/www/SandeCVCMW13> <DBLP:http://dblp.org/rec/conf/i-semantics/GraubeHU14> that involve cryptographic hash values to enforce immutable versions. (✋)

Cited paper details: M. Graube, S. Hensel, and L. Urbas. R43ples: Revisions for triples. In Proceedings of the 1st Workshop on Linked Data Quality. Citeseer, 2014.

arXiv.org ArXiV and dblp links for 1708.09193 dblp

2. Knowledge Fusion via Embeddings from Text, Knowledge Graphs, and Images

Thoma, Steffen; Rettinger, Achim; Both, Fabian

Submitted on April 20, 2017.

The term 'Knowledge Graph' was coined by Google in 2012 and is since then used for any graph-based knowledge base, the most popular examples being DBpedia, Wikidata, and YAGO (see <DBLP:http://dblp.org/rec/journals/semweb/FarberBMR18> for a survey). (✋)

Cited paper details: Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web Journal (to be published) (2017)

arXiv.org ArXiV and dblp links for 1704.06084 dblp

Fig. 1: Example of searching for citation contexts given a cited paper's title (here, the partial title "linked data quality").

1. Search for papers given a phrase. Given a phrase (here within the field of computer science), this search functionality allows users to search for all papers on arXiv that contain this phrase in the body text. Thus, this functionality is particularly suitable when papers with very specific keywords (e.g., "semantic cognition," "knowledge base completion," and "stochastic pooling") and not only abstract research topics need to be retrieved.

2. Search for papers given a paper's (incomplete) title. Given the title of a paper, this search functionality allows users to search for its full metadata. Also, only parts of the paper's title can be provided as inputs. For instance, searching for "linked data quality" allows users to search for all papers on that topic.

3. Search for papers given an author's name. Given an author's name, this functionality allows users to retrieve the full metadata of all papers written by this author.

4. Search for citation contexts given the cited paper's title. Given any paper's title, this search functionality allows users to retrieve all sentences from the bodies of arXiv papers in which the specified paper is cited (see Fig. 1). If a publication is cited several times within a paper, then all citation contexts of this paper are grouped together. To allow a quick assessment of the retrieved citation contexts by the user, an icon is presented next to each citation context. This icon indicates the citation context's polarity [5,6] and can be positive (i.e., the cited paper is praised), neutral, or negative. The citation polarity values were determined offline by using Athar et al.'s approach [7] and data set for training.

5. Search for citation contexts given the cited paper's author. Here, users can search for the citation contexts (plus the papers' metadata and the links to arXiv.org) in which papers written by the given author were cited. For instance, by searching for "Tim Berners-Lee," our search engine retrieves all contexts in which papers written by Tim Berners-Lee have been cited. We also provide the citation polarity indication for this search functionality.

3 Related Work

Search engines for papers and citation contexts. Paper search systems, such as *Google Scholar*, can be used to retrieve papers and obtain papers' metadata. The keywords provided by the user are thereby matched with the papers' metadata and with the papers' contents. Consequently, such systems essentially cover the first three functionalities. However, to the best of our knowledge, no system has been presented that allows users to search specifically for citation contexts, as enabled in our system. Note that citation contexts have already been analyzed with respect to various aspects [8,9,10,11,12]. However, we are not aware of available systems that provide the citation polarity information [7,5] of citation contexts.

Data sets with papers' contents and citation contexts. Besides the arXiv CS data set, the CiteSeerX [13] and the Microsoft Academic Graph (MAG) [14] are also considerably large data sets that contain already extracted citation contexts. However, the version of CiteSeerX available online is not up-to-date. MAG's citation contexts are partially noisy, and the exact positions of citations within the citation contexts are not provided.

4 Conclusion

In this paper, we presented a system that allows searching for papers and citation contexts in a novel way. As our system is based on the large collection of computer science papers on arXiv.org, it can be used by any computer science researcher. Also, scientometrics researchers can use the system to explore how citations are embedded in papers. In terms of future work, we plan on improving the ranking of papers and citation contexts and to include a component for recommending papers that are related to the retrieved citation contexts.

Acknowledgements. This research has been supported by the Research Innovation Fund of the University of Freiburg (#2100189801), by JSPS KAKENHI grants (#17H01828, #18K19841), and by the MIC/SCOPE (#171507010) grant.

References

1. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11) (2015) 2215–2222
2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379) (2018)
3. Ware, M., Mabe, M.: The STM Report: An overview of scientific and scholarly journal publishing. (2015)
4. Färber, M., Thiemann, A., Jatowt, A.: A High-Quality Gold Standard for Citation-based Tasks. In: Proceedings of the International Conference on Language Resources and Evaluation. LREC'18 (2018)
5. Abu-Jbara, A., Ezra, J., Radev, D.R.: Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In: Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics. NAACL'13 (2013) 596–606
6. Ghosh, S., Das, D., Chakraborty, T.: Determining Sentiment in Citation Text and Analyzing Its Impact on the Proposed Ranking Index. In: Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing'16 (2016) 292–306
7. Athar, A.: Sentiment Analysis of Citations using Sentence Structure-Based Features. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. NAACL'11 (2011) 81–87
8. Alvarez, M.H., Gómez, J.M.: Survey about citation context analysis: Tasks, techniques, and resources. Natural Lang. Eng. **22**(3) (2016) 327–349
9. Tahamtan, I., Bornmann, L.: Core elements in the process of citing publications: Conceptual overview of the literature. J. Informetrics **12**(1) (2018) 203–216
10. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: The next generation of citation analysis. JASIST **65**(9) (2014) 1820–1833
11. Bornmann, L., Daniel, H.: What do citation counts measure? A review of studies on citing behavior. Journal of Documentation **64**(1) (2008) 45–80
12. Todeschini, R., Baccini, A.: Handbook of Bibliometric Indicators: Quantitative Tools for Studying and Evaluating Research. Wiley (2016)
13. Caragea, C., Wu, J., Ciobanu, A.M., Williams, K., Ramírez, J.P.F., Chen, H., Wu, Z., Giles, C.L.: CiteSeer x : A Scholarly Big Dataset. In: Proceedings of the 36th European Conference on IR Research. ECIR'14 (2014) 311–322
14. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings of the 24th International Conference on World Wide Web. WWW'15 (2015) 243–246