Analysing author name mentions in citation contexts of highly cited publications

Rajesh Piryani[1,2], Wolfgang Otto[2], Philipp Mayr[2], and Vivek Kumar Singh[3]

[1] South Asian University, New Delhi, India
rajesh.piryani@gmail.com
[2] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
firstname.lastname@gesis.org
[3] Banaras Hindu University, Varanasi, India
vivekks12@gmail.com

Abstract. In this paper, we are analysing author name mentions in citation contexts of highly cited articles in a PLOS ONE corpus. First, we have identified author mentions in our corpus of citation contexts. Then, we examined frequent nouns and verbs in the neighbourhood of the identified author mentions using n-grams and utilized these top nouns and verbs to identify the most frequent patterns. We observed that most frequent patterns are associated with the methods which are proposed in the corresponding highly cited references.

Keywords: Citation Context · Citation Behavior · Author Mention · PLOS ONE · Method Papers

1 Introduction

Scientific research publications are structured texts which integrate specific characteristics associated to their references. The accessibility of full-text of research publications and available natural language processing techniques have largely extended the possibilities to analyse the citation behavior and utilization of the cited articles in bibliometrics. In this article, our main objective is to identify and analyse author name mentions (in the following author mentions) in citation contexts of PLOS ONE articles. For example in the citation context "Western immunoblotting was performed according to Towbin method [60][1].", the word *Towbin* is an author mention. In the following, we define an author mention in a citation context as a conscious mention of a concrete author by name. An example of a citation context *without* an author mention is: "This pathogen causes a wide spectrum of clinical illnesses, including skin and soft tissue lesions, and lethal infections such as osteomyelitis, endocarditis, pneumonia and septicemia [1]." Our main objective in this study is to explore some characteristics which could help to identify specific paper types in a set of highly

[1] [60] Towbin, H., Staehelin, T., and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proceedings of the National Academy of Sciences, 76(9), 4350-4354.

cited publications. Our first approach is to explore surface patterns in the corresponding contexts with author mentions. This study is inspired by a recent paper from Small [12]. In this and future work, we plan to use the results of this exploration to introduce methods to better understand the genesis and reasons of an influential work.

For the exploration of the relevant author name contexts we have counted verbs and nouns surrounding the author mentions using n-grams. We are utilizing the top frequent verbs and nouns to identify the frequent n-grams patterns involving author mentions. We have observed that most frequent patterns are associated with the methods which are proposed in highly cited publications in our corpus. In this paper, highly cited publications are those which are cited more than 100 times in our PLOS ONE corpus.

2 Related Work

Citations are an important parameter of connectivity of related research works. A lot of studies have focused on analysing citation for different purposes ranging from assessment of the quality of the article to tracing the flow of ideas on a topic. Sugiyama et al. [13] have suggested that there could be two kinds of citation analysis: (1) Citation counts and (2) Citation context analysis. They argue that citation context analysis could be a better technique to identify the influence of a research article. A citation context is often defined as the sentence where a particular reference is cited. Unlike simple counts, citation context analysis identifies the contextual relationship between citing research articles and referenced articles by applying various Natural Language Processing (NLP) and Machine Learning (ML) approaches [9]. In this way, the text of articles, particularly that portion where it cites another article are processed.

Earlier, some researchers have incorporated the citation contexts with the opinion mining of citations [5,1,4]. Hyland has analysed the self-mention in research articles [10]. He has explored forms and function of self-mention in a dataset of 240 research articles. Abu-Jbara and Radev [2] have developed various techniques on the identification of sentences that are associated with the targeted reference. They have used word classification, sequence labelling and segment classification techniques for detecting the fragments of a citing sentence. Yeh et al. [15] have proposed classification approach that differentiates cited and non-cited pairs and sentences references. In a recent work, Small [12] has investigated the highly cited publications based on citation contexts. An et al. [3] have used NLP techniques and citation contexts to find the characteristics of top-cited authors; they have used the ACL Anthology dataset [8]. Atanassova & Bertin [6] have explored the locations of citation context in IMRaD structure regardless of the age of the cited references. Bertin et al. [7] have shown the most frequent linguistic patterns identified in the citation contexts of articles varies according to their occurrence in the IMRaD structure. In a recent paper, we analyse citation contexts of highly cited publications in a PLOS ONE corpus [11]. In particular, we study the position of the contexts based on the IMRaD structure over time.

The work in this study is different from existing work insofar that we are identifying the author mentions in citation contexts and finding the frequent verbs and nouns surrounding the author mention using n-grams. This study is an exploratory analysis in the extension of our previous paper [11]. The identification and analysis of author mentions in citation contexts is a novel approach and to the best of our knowledge no work on this has been proposed till now.

3 Methodology

Our research object is a corpus of citation contexts of highly cited publications (see Appendix table A1 for more examples of citation contents with author mentions) introduced in [11]. In this corpus, the citation context is defined as *the sentence* in which the citation occurs. For all citation contexts, we have included information about the citing articles. As metadata for the cited articles, the extracted information which has been made available along the contexts comes from the reference sections of the citing articles (it comprises publisher-id, article title, year, abstract, etc..). We have created this corpus from 176,856 PLOS ONE full-text articles published from 2006 to 2017. Selected parameters of the corpus can be found in Table 1.

We have selected all references which are cited in more then 100 PLOS ONE articles. In the context of this article, we call those referenced objects highly cited publications. To remove the problem of deduplication, we have selected only those references whose PubMed IDs listed in the reference part of the citing publications. With the help of the PubMed IDs, we retrieved all citation contexts related to these articles. Due to errors in the annotation of citations in our data basis (PubMed XML documents), not every reference citation context for each

Table 1: Overview of the corpus by selected parameters.

Description	value
No. of citing PLOS ONE papers	176,856
No. of extracted reference strings	8,473,649
No. of extracted citation contexts	31,746,769
Publication period of citing papers	2006-2017
Publication period of the cited papers	1951-2015
No. of the cited publications published after 2010	69
No. of relevant citing PLOS ONE publications	62,127
No. of relevant references	666
Max no. of contexts of relevant references	3363
Min no. of contexts of relevant references	75
Mean no. of contexts of relevant references	261
Median no. of contexts of relevant references	184
No. of relevant extracted citation contexts	173,630
No. of referenced objects per citation context (mean)	1.78
No. of referenced objects per citation context (median)	1

highly cited publication is available. This leads to the fact that the smallest number of citation contexts per reference does not exceed 75.

Table 1 shows an example of a citation context with its associated metadata. In total, we have 666 references which are cited in more than 100 publications. We call those top-666[2]. Further, we have excluded all the citation contexts which do not cite any of the top-666 articles. With this procedure, we reduce the absolute number of citing papers used in our study to 62,127.

Table 2: Example of extracted data for one citation context

DOI of citing paper	10.1371/journal.pone.0028665
Citation context	Supernatant was collected, its volume recorded, and total protein concentration of each sample was measured using the Bradford protein assay and a BSA protein standard (both from BioRad, CA, USA) according to the Bradford method [50].
PubMed ID of reference	942051[3]
Author list of reference (surname)	[Bradford]
Section title	Materials and methods
Pub. year of the citing paper	2011

For our analysis, we have 173,630 relevant citations contexts. This number reflects the fact that only 0.5 percent of citation contexts (i.e. total 31,746,769) have at least one top reference in the context as cited. The top-666 highly cited publications are published between 1951 to 2015. The distribution of the number of citation contexts per top-666 follows a power-law-like shape. The most cited reference is mentioned in 3,363 citation contexts and the lowest number citation contexts is 75, where the median reference is mentioned in 184 contexts.

4 Results

The citation contexts of our interest are the ones with author mentions. One example is the context: “Western immunoblotting was performed according to Towbin method [60].” To identify matches we search for exact equivalents of author surnames based on our metadata of the cited publications in the text of each citation contexts. In our example Towbin, Staehelin, and Gordon are the authors of [60]. The first author Towbin has an exact match in the text of our citation context. We have found 11,977 contexts (i.e. 6.9% of 173,630 relevant citation contexts) whose author mention matches to at least one author of a cited

[2] The underlying dataset of our study contains all relevant contexts citing the top 666 referenced publications and is available under `https://github.com/Scientotext/PLOS-ONE-Dataset`

[3] https://www.ncbi.nlm.nih.gov/pubmed/?term=942051

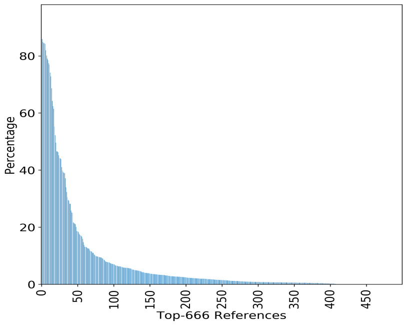

Fig. 1. Top-666 references with author mentions in citation contexts.

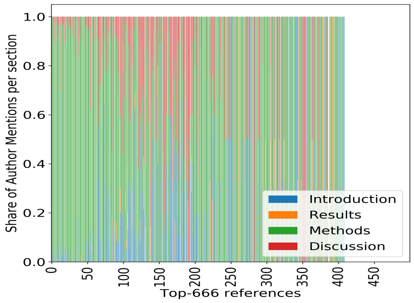

Fig. 2. Share of IMRaD sections of the contexts with author mentions.

reference. In the majority of these contexts (11,687 contexts) the first author was mentioned. In rare cases the second (194 contexts), the third (52 contexts), the fourth (24 contexts), the fifth (19 contexts) or the sixth (1 context) member of the author list is the first matching one. One of the possible reasons is a spelling error of the author name in a citing article. For our analysis, we have considered 11,687 citation contexts with a match of the first author surname of the author list found for the cited publication. Figure 1 shows the top-666 references whose author mentions found in relevant citation contexts. It is observed from Figure 1 that 409 top cited references from top-666 are found in citation context with author mentions. Figure 2 shows for each of our top-666 referenced publications on the x-axis the proportion of IMRaD sections in which the author mentions appear in the citation contexts. In more than half of the citation contexts author mentions were found in the method sections. The author mention identification method can also work for other citation styles such as APA, MLA. But, in our current corpus, IEEE citation style is used.

For the next analysis we parsed citation contexts with Parts of Speech (POS) tagger and removed stop words. Then we extracted all n-grams containing author mentions on the one hand, and verbs or nouns on the other. For Figure 3 and 4 we extracted all nouns and verbs of the selected n-grams. The figures are showing the frequency and the number of top-666 publications referenced at least one time in a context containing the shown verb or noun. Figure 4 gives a hint, that the author name matches are connected mostly to specific methods, protocols, tests or models.

To exemplify this hypothesis we selected all trigrams containing author match and verbs or noun. The frequencies of these trigrams are shown in Figure A1. Here we can figure out, that for specific highly cited publications specific formulation patterns can be identified. For example "using Bradford method" is used more than 250 times. We found this pattern for other author names, too. In future work we plan to cluster the highly cited publications by the occurrence of commonly used patterns identifiable by this statistic.

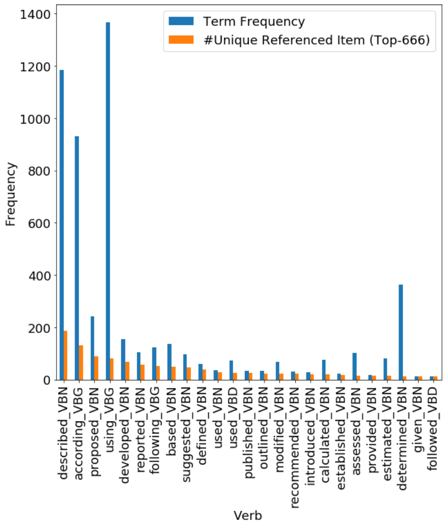

Fig. 3. Top 25 verbs next to author mentions in citation context ordered by the number of different referenced Top-666 publications.

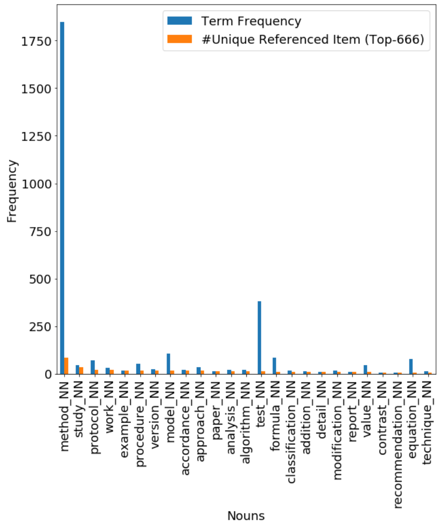

Fig. 4. Top 25 nouns next to author mentions in citation context ordered by the number of different referenced Top-666 publications.

5 Future work

As the results show, the search for patterns connecting metadata of highly cited publications with citation contexts seems to be promising. Based on that, we try to detect more patterns reflecting functional usage in citation contexts [14]. One approach is to search for abbreviations (e.g. LDA as abbreviation for Latent Dirichlet Allocation or MEGA6 for "Molecular Evolutionary Genetics Analysis version 6.0") in titles of referenced publications which are often used to describe tools and methods. By understanding which citation contexts use these abbreviations, we try to get a more complete picture of the functional use of citations for highly cited publications. Another promising perspective is to introduce overtime usage of the author and the abbreviation matching pattern [11]. To be able to do a time analysis, we need to switch to a corpus which reflects a larger time period of publications. Otherwise the supporting occurrences of the matching patterns based on time slices are to low. We will extend our work to design the algorithm to identify important methods used in cited article.

Acknowledgement

The authors acknowledge the enabling support provided by the Indo-German Joint Research Project titled Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles (Grant No. DST/INT/FRG/DAAD/P-28/2017) for this work.

References

1. Abu-Jbara, A., Ezra, J., Radev, D.: Purpose and polarity of citation: Towards nlp-based bibliometrics. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 596–606 (2013)
2. Abu-Jbara, A., Radev, D.: Reference scope identification in citing sentences. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 80–90. Association for Computational Linguistics (2012)
3. An, J., Kim, N., Kan, M.Y., Chandrasekaran, M.K., Song, M.: Exploring characteristics of highly cited authors according to citation location and content. Journal of the Association for Information Science and Technology **68**(8), 1975–1988 (2017)
4. Athar, A.: Sentiment analysis of scientific citations. Tech. rep., University of Cambridge, Computer Laboratory (2014)
5. Athar, A., Teufel, S.: Context-enhanced citation sentiment detection. In: Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies. pp. 597–601. Association for Computational Linguistics (2012)
6. Bertin, M., Atanassova, I.: Temporal properties of recurring in-text references. D-Lib Magazine **22**(9/10) (2016)
7. Bertin, M., Atanassova, I., Sugimoto, C.R., Lariviere, V.: The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. Scientometrics **109**(3), 1417–1434 (2016)
8. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M., Lee, D., Powley, B., Radev, D., Tan, Y.: A reference dataset for bibliographic research in computational linguistics. Proceedings of the Sixth International Language Resources and Evaluation. European Language Resources Association, Paris (2008)
9. Hernández-Alvarez, M., Gómez, J.M.: Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering **22**(3), 327–349 (2016)
10. Hyland, K.: Humble servants of the discipline? self-mention in research articles. English for specific purposes **20**(3), 207–226 (2001)
11. Otto, W., Ghavimi, B., Mayr, P., Piryani, R., Singh, V.K.: Highly cited references in PLOS ONE and their in-text usage over time. In: Proceedings of the 17th International Conference on Scientometrics & Informetrics (ISSI 2019) (2019), `https://arxiv.org/abs/1903.11693`
12. Small, H.: Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. Journal of Informetrics **12**(2), 461–480 (2018). https://doi.org/10.1016/j.joi.2018.03.007
13. Sugiyama, K., Kumar, T., Kan, M.Y., Tripathi, R.C.: Identifying citing sentences in research papers using supervised learning. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). pp. 67–72. IEEE (2010)
14. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 conference on empirical methods in natural language processing. pp. 103–110. Association for Computational Linguistics (2006)
15. Yeh, J.Y., Hsu, T.Y., Tsai, C.J., Cheng, P.C.: Reference scope identification for citances by classification with text similarity measures. In: proceedings of the 6th international conference on software and computer applications. pp. 87–91. ACM (2017)

Appendix

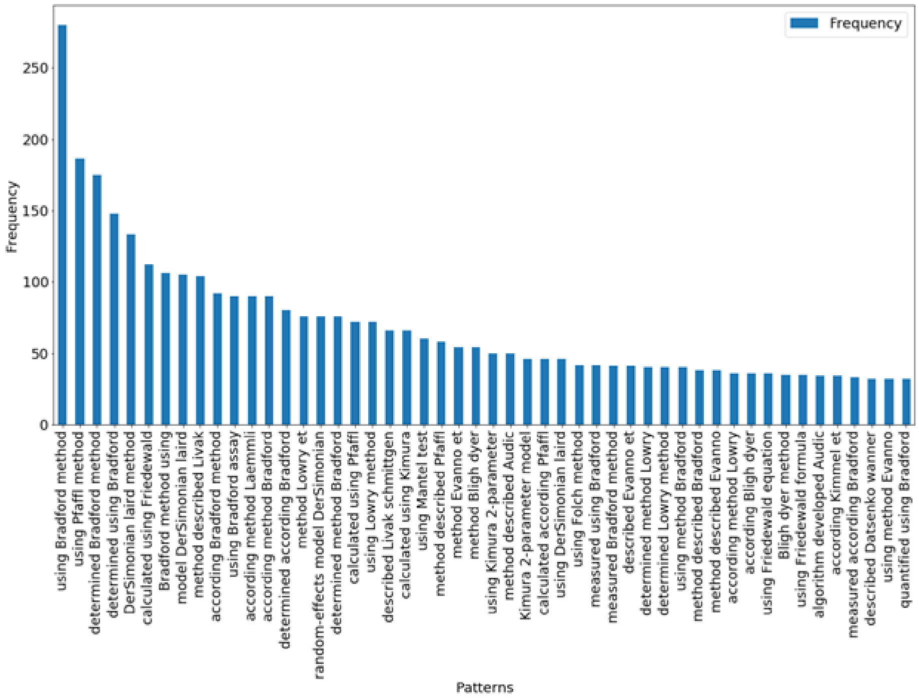

Fig. A1. Trigram patterns containing author mentions and top 15 frequent verbs and top 15 frequent nouns.

Table A1: Examples of Citation Contexts with Author Mentions

	Citation Context
1	Standards proposed by *Lanidis* and *Koch* [34] were used to interpret resulting kappa values, where perfect agreement equates to a kappa of 1 and chance agreement equates to 0.
2	To analyse the historical demography, the mismatch distribution [43] of the sums of squares deviation (SSD) [44] and *Harpending*'s raggedness index (HRI) were used, which allowed for testing of the model of sudden population expansion [45].
3	The laboratory strains of P. falciparum were grown and maintained in culture using the method of *Trager* and *Jensen* with some modifications [15,16].