An Authoritative Approach to Citation Classification

David Pride*
david.pride@open.ac.uk
KMi, The Open University
Milton Keynes, United Kingdom

Petr Knoth*
petr.knoth@open.ac.uk
KMi, The Open University
Milton Keynes, United Kingdom

ABSTRACT

The ability to understand not only that a piece of research has been cited, but why it has been cited has wide-ranging applications in the areas of research evaluation, in tracking the dissemination of new ideas and in better understanding research impact. There have been several studies that have collated datasets of citations annotated according to type using a class schema. These have favoured annotation by independent annotators and the datasets produced have been fairly small. We argue that authors themselves are in a primary position to answer the question of why something was cited. No previous study has, to our knowledge, undertaken such a large-scale survey of authors to ascertain their own personal reasons for citation. In this work, we introduce a new methodology for annotating citations and a significant new dataset of 11,233 citations annotated by 883 authors. This is the largest dataset of its type compiled to date, the first truly multi-disciplinary dataset and the only dataset annotated by authors. We also demonstrate the scalability of our data collection approach and perform a comparison between this new dataset and those gathered by two previous studies.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Digital libraries and archives**; • **Applied computing** → *Publishing*.

KEYWORDS

Citation Typing, Citation classification, Data Mining, Open Access, Scholarly Data, Research Evaluation

ACM Reference Format:
David Pride and Petr Knoth. 2020. An Authoritative Approach to Citation Classification. In *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3383583.3398617

1 INTRODUCTION

Citation behaviour has been the focus of a wide number of research studies over the last fifty years. The study of citations can effectively be grouped into two over-arching classes. Firstly there is the study of *why* authors cite each other. As long ago as 1957, Merton [8] wrote that citations were acknowledgment of credit for new ideas as well as a form of social recognition. Eugene Garfield himself suggested 15 reasons why authors may cite one another [3]. In 1977 Ina Speigel-Rosing defined 13 citation types that have been used in many studies in this domain [10]. In 1998, Leydesdorff [6] notes that alongside the sociological interpretation of citations there lies an information-theoretical one. This leads to the second area of citation usage which attempts to quantify citations and use these to produce bibliometric measures. These two areas are most often viewed independently, however, we suggest that these two areas are inextricably linked. Without an understanding of why citations are occurring, we argue that bibliometric measures that rely on citation counts alone are potentially missing a large amount of information. It is therefore more interesting and valuable to know not only that a piece of work was cited, but also the reason for this citation. Our study shows, at scale, that authors are well placed and, in our opinion, the best informed to identify why a particular piece of work was cited. The availability of much more detailed citation information, beyond that of a single digit, opens up a range of possibilities in terms of not only research evaluation but also in tracking scientific discourse and mapping the spread of new ideas and the adoption of new tools and methodologies. Our Academic Citation Typing (ACT) platform is a fully scalable online citation annotation tool that is compatible with all PDF files. Using this platform we engaged 883 authors as annotators and used this tool to collate the largest dataset to date of annotated citations. We also show that annotations by authors are closely aligned with that of domain experts and independent annotators. This tool and methodology is now made available to others. Additionally, there are strong use cases for adoption of this technology at the point of publication / deposit with publishers, repositories, journals and conference systems.

2 RELATED WORK

Numerous previous studies have introduced citation classification schemes as a way to identify the meaning or purpose behind a particular citation. They have, however, produced fairly small datasets overall (in the region of 2-3,000 annotations). Collating accurately labelled datasets of this type requires a significant amount of human effort. These studies have mostly used domain experts or the authors of the particular study themselves for the annotation process. In 2006, Teufel et al. [11] introduced the then largest dataset of 2,829 annotated citations. These citations were drawn from computer science papers from the ACL anthology reference corpus. Their citation classification schema used 12 distinct types for annotation and the annotations were created by the study's authors. This dataset and the classification schema applied has become the basis for much of the following work in this domain. In 2016 Jurgens et al. [4] introduced a new dataset of 1,969 annotated citations and

*Both authors contributed equally to this research.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06.
https://doi.org/10.1145/3383583.3398617

simplified the 12 types first suggested by [11] into six types as follows:

- Background
- Uses
- Compare / Contrast
- Motivation
- Extension
- Future Work

The authors in Jurgens et al. [4] show that the six categories provide sufficient granularity in terms of identifying citation purpose whilst remaining broad enough for the classification scheme to be usable. In this study, the annotations were completed by two independent domain experts.

Most recently Cohan et al. [2] took the classification schema first defined by Jurgens et al. [4] and reduced these six classes to just three; METHOD, RESULTCOMPARISON and BACKGROUND. This study used crowd-sourced volunteers as annotators who were trained using instructions from domain experts.

3 METHODOLOGY

Collection of accurately annotated citation data has previously been slow, expensive and relied on the subjective opinion of independent annotators. Our novel methodology for collecting annotated citations differs from previous works as we employ a large number of authors as annotators. As discussed in Case [1], asking authors may potentially introduce problems of both recall (will the citing author remember their reasons for citing?) and the 'social desirability' of answers (will they answer honestly?) This is of particular relevance in the case of negative or contradictory citations. However, Case concludes "One must start somewhere if we are to achieve a better understanding of citation behavior. If the motivations of authors are to be understood, then asking authors directly about their motivations, despite the methodological pitfalls of self-reporting, is a logical place to approach the issue." We suggest that despite these caveats, the author is likely best placed to annotate citations in their own paper as this immediately removes a layer of interpretation and overcomes any limitations in expression of language. Further, by removing the requirement for domain experts this immediately increases the scope and scalability of this type of study. As Teufel et al. [11] noted; Citation function is hard to annotate because it, in principle, requires interpretation of author intentions. These observations were highly influential to our approach. Whilst authors will undoubtedly have their own biases, we agree with Case [1] that asking authors to annotate their own papers is both logical and effective. We also compare our completed dataset to two earlier ones to assess overall levels of agreement between authors and independent annotators in terms of citation classification and distribution.

3.1 Selection of a citation classification scheme

Careful consideration must be given when selecting both the class labels and the most effective number of classes for the specific task of citation classification. Both the method of collection and the classification schema utilised will have an impact on the utility of the final dataset. As a starting point, we used the following statements:

- All citations must have a class
- No citation may have more than one class.

The intention here is there should be no 'null' class. The six types outlined above provide broad enough definitions to allow all citations to comfortably fit within a single class. The second point is somewhat more complex as an author may cite a source several times in one paper. In one sentence the author may reference a particular piece of work as motivation for their own study and in another also be comparing their results to those of the cited work. However, distinctly different language can be found within the citing sentence itself for each particular class. Our work builds primarily on the studies of [4], [11] and [12] as we choose to collect annotations according to both purpose and influence. To allow for future cross-study comparison of results, we keep our classification schema compatible with those of [4] and [11], however we add an additional layer to the compare/contrast category; show similarities, show differences or show disagreement. The addition of a sub-class here presents a new way to demonstrate that not all citations are necessarily created equal. It can be argued that citations that disagree with or refute earlier work deserve far more weight and visibility than current citation metrics allow for. Our final classification schema can be seen in Table 1.

Class Label	Description
BACKGROUND	The cited paper provides relevant Background information or is part of the body of literature.
USES	The citing paper uses the methodology or tools created by the cited paper.
COMPARE_CONTRAST - similarities - differences - disagreement	The citing paper expresses similarities or differences to, or disagrees with, the cited paper.
MOTIVATION	The citing paper is directly motivated by the cited paper.
EXTENSION	The citing paper extends the methods, tools or data etc. of the cited paper.
FUTURE	The cited paper may be a potential avenue for future work.

Table 1: The citation classification schema

3.2 Annotating Citations

For the annotation process we employ our own online annotation tool we call the Academic Citation Typing (ACT) platform [9] which displays the full text of the authors' research paper alongside a point-and-click style classification interface. In-text citation markers are automatically highlighted as a visual prompt for the annotator and displayed alongside the cited paper's title, author name, publication date and the full sentence containing the citation. This allows authors to rapidly and accurately assign one of the six class labels to each citation in their paper. Additionally, the ACT

platform also records the time taken by each author to complete the annotation process, which was an average of nine minutes, around 22s per citation.

The ACT platform was, prior to launch, tested for user experience and reliability with 6 internal evaluators. To ensure the tests were as realistic as possible, the evaluators were the first authors of the sample papers used in the annotation process. Our observations and feedback from the evaluators who completed the annotation process is that first authors, in almost all cases, remember their own reasons for citing a particular paper without prompting and can therefore complete the process quickly and with confidence.

We conducted post-annotation interviews with the six evaluators who tested the ACT platform. These interviews demonstrated two key points; Authors rarely need the *contents* of the citing sentence to annotate the citation. Provision of the title and author name is largely sufficient for the author to remember the reason for the citation. This is particularly true for papers published most recently. A live demonstration of the platform in use can be viewed here: https://youtu.be/8l7frJ-fde8

4 DATASET COLLATION AND RESULTS

Using full-text research papers drawn from CORE [1] we collated an initial dataset of 26,652 papers from across multiple disciplines. We extracted first author names, email addresses and approximately 407,907 citing sentences using Grobid [7]. This dataset was then uploaded to the ACT platform which automatically generates a unique URL token for each paper. Invites to take part in the annotation process were then sent via email to authors. 883 authors responded and completed a total of 11,233 annotations. This is the largest dataset in existence of citations annotated according to type and the only author-annotated dataset. The full composition of this dataset can be seen in Table 3.

BACKGROUND	54.61%
USES	15.51%
COM/COM	12.05%
MOTIVATION	9.92%
EXTENSION	6.22%
FUTURE	1.70%

Table 2: Breakdown of ACT dataset by class

The full dataset of annotated citations contains the citing paper title, author and publication date; the extracted citing sentence and the author-annotated class label. Also included is the CORE paper ID [2] which is the source of the full-text paper and the Microsoft Academic Graph (MAG) ID [3] which can be used to retrieve further bibliographic and bibliometric data for each paper. The full ACT dataset will be available for download after an initial embargo period. [4]

[1] http://core.ac.uk
[2] https://core.ac.uk
[3] https://academic.microsoft.com/home
[4] This is due to a portion of the dataset being used for the 8th WOSP Workshop being held in conjunction with JCDL2020.

4.1 Comparison of datasets and collection methodologies

In this section, we look at how previous studies have collected and annotated data and compare this to our collection method. Table 4 shows the total number of papers, annotators and citations from our study and from three seminal studies in this domain.

Our work is most closely aligned with that of Jurgens et al. [5]. We retain the six-way classification schema as previously noted. Figure 1 shows the comparative breakdown of the datasets from these two studies. Whilst we do observe some small differences in the distribution of citation types when comparing the datasets that have been annotated by domain experts and those that have been annotated by authors, overall there is a strong positive correlation, r=.93, p≤.04, n=6, between the two.

It is interesting to note that authors are much more likely than independent annotators to regard a piece of work as motivational or an extension to previous work rather than simply background information. It is felt that this level of insight can only be provided by the author. Whilst the studies of Teufel et. al and Jurgens et. al used domain experts for the annotation process, the study by Cohan et al. [3] employed crowd-sourced volunteers. Volunteers were trained in the annotation process, guided by a domain expert. Several steps were undertaken to ensure the accuracy of the crowd-sourced an- notations. The final Sci-Cite dataset produced by Cohan et al. [3] contains 11,020 annotated citations taken from papers in two domains, bio-medicine and computer science.

Both the chosen citation class labels and the method of annotation will have profound effects on the composition of the final dataset. This can be most clearly seen if we compare the dataset from Cohan et al. [3] with those from Jurgens et al. [5] and this study. To allow for a direct cross-study comparison we collapsed the six classes from our dataset and that of Jurgens et al. [5] to the three classes used by Cohan et al. [3]. This simplification entails re-labelling citations from the EXTENSION, MOTIVATION and FUTURE classes as BACKGROUND. This is the same process as used in the original study.

When the class granularity is reduced to three overall classes it is reasonable to expect some small changes in class distribution. The final breakdown across all three studies can be seen in Figure 2. The most notable difference in the dataset produced by Cohan et

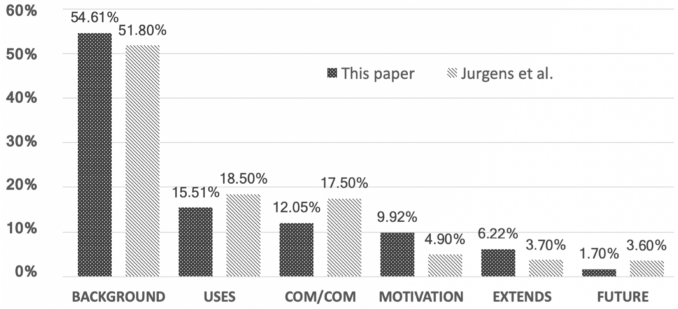

Figure 1: Comparison of dataset breakdown by class from Jurgens et al. and Pride and Knoth.

Study	Papers	Annotators	Anno. by	Citations	Discipline(s)
Teufel et al.	360	3	Study Authors	2,829	Comp. Limngustics
Jurgens et al.	185	3	Domain Experts	1,989	Comp. Science
Cohan et al.	?	880	Volunteers	11,020	Comp. Sci / Bio Sci.
Pride & Knoth	883	883	Paper Authors	11,233	Multi-disciplinary

Table 3: Cross-study comparison of dataset collation and annotation type.

al. [3], when compared to that of the other two studies, is that the METHOD class contains 29% of all citations, with a resulting drop in the total number of annotations in the BACKGROUND class. This is significantly different from Jurgens et al. [4] where this figure is 18.5%, and our dataset at 15.5%. It is known from correspondence with the authors that this difference was produced by oversampling of citations from the lesser represented METHOD class. Although over-sampling techniques are popular and easy to use, there are many pitfalls to avoid when they are applied. Over-sampling of an imbalanced dataset can produce potentially optimistic results when the resulting dataset is then used to train machine learning models. As Vandewiele et al. 2020 [13] notes in a recent study on the effects of over-sampling, *'The results may more reflect the model's capability to memorize samples seen during training, rather than its predictive performance if it were applied in a real-world setting on unseen data.'*

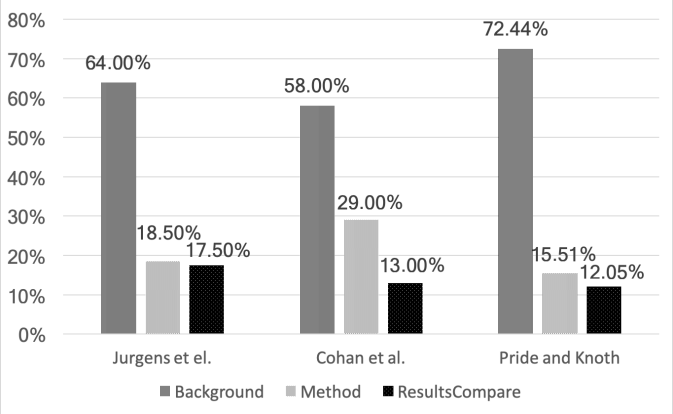

Figure 2: Dataset breakdown by citation type.

5 CONCLUSION

If reason for citation can become a part of the standard metadata accompanying a scientific publication this opens many further avenues for study and also provides opportunities for enhancing current bibliometrics. In this study we show that authors are best placed as the source of this valuable additional citation information. Further we demonstrate that whilst citations annotated by authors show close correlation with experts there are important and significant differences. Authors themselves were approximately twice as likely to regard a citation as motivational or as extending previous work. Our citation classification scheme also adds the ability to recognise citations that explicitly disagree with or contradict previous works. None of this information is available with current citation metrics. The importance of having the ability to classify citations according to type has often been overlooked. If citations can be shown to demonstrate actual utility, it can be argued this is a better reflection of the impact of a piece of research. The output from this study is two-fold. Firstly the new ACT dataset can be utilised in improving models for the automatic identification of citation type. Further, the ACT platform itself can be adopted by a range of stakeholders for use at the point of publication. It is easy to envisage a scenario where a publisher or content provider is able to provide enhanced bibliographic information using author annotated citations.

6 ACKNOWLEDGEMENTS

This work has been funded by Jisc and has also received support from the scholarly communications use case of the EU OpenMinTeD project under the H2020-EINFRA-2014-2 call, Project ID: 654021

REFERENCES

[1] Donald O Case and Georgeann M Higgins. 2000. How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science* 51, 7 (2000), 635–645.
[2] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. *arXiv preprint arXiv:1904.01608* (2019).
[3] Eugene Garfield et al. 1972. Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science.
[4] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *arXiv preprint arXiv:1609.00435* (2016).
[5] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6 (2018), 391–406.
[6] Loet Leydesdorff. 1998. Theories of citation? *Scientometrics* 43, 1 (1998), 5–25.
[7] Emilio Delgado López-Cózar, Nicolas Robinson-Garcia, and Daniel Torres-Salinas. 2012. Manipulating Google Scholar citations and Google Scholar metrics: Simple, easy and tempting. *arXiv preprint arXiv:1212.0638* (2012).
[8] Robert K Merton. 1957. Priorities in scientific discovery: a chapter in the sociology of science. *American sociological review* 22, 6 (1957), 635–659.
[9] David Pride, Petr Knoth, and Jozef Harag. 2019. ACT: An Annotation Platform for Citation Typing at Scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 329–330.
[10] Ina Spiegel-Rosing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science* 7, 1 (1977), 97–113.
[11] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 103–110.
[12] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations. In *AAAI Workshops*. http://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185
[13] Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, et al. 2020. Overly Optimistic Prediction Results on Imbalanced Data: Flaws and Benefits of Applying Over-sampling. *arXiv preprint arXiv:2001.06296* (2020).