Incidental or influential? - A decade of using text-mining for citation function classification.

David Pride and Petr Knoth

The Knowledge Media Institute, The Open University, Milton Keynes, UK.
{david.pride, petr.knoth}@open.ac.uk

Abstract. This work looks in depth at several studies that have attempted to automate the process of citation importance classification based on the publications' full text. We offer a comparison of their individual similarities, strengths and weaknesses. We analyse a range of features that have been previously used in this task. Our experimental results confirm that the number of in-text references are highly predictive of influence. Contrary to the work of Valenzuela et al. (2015) [1], we find abstract similarity one of the most predictive features. Overall, we show that many of the features previously described in literature have been either reported as not particularly predictive, cannot be reproduced based on their existing descriptions or should not be used due to their reliance on external changing evidence. Additionally we find significant variance in the results provided by the PDF extraction tools used in the pre-processing stages of citation extraction. This has a direct and significant impact on the classification features that rely on this extraction process. Consequently, we discuss challenges and potential improvements in the classification pipeline, provide a critical review of the performance of individual features and address the importance of constructing a large scale gold-standard reference dataset.

1 Introduction

Citation analysis and bibliometrics are being increasingly used as a tool in assessing the impact of research. The three largest citation databases; Google Scholar, Web of Science (WoS) and Scopus all give prominence to citation counts to provide information regarding the number of times a paper has been cited. Most measures widely used to measure performance of research, such as the controversial Journal Impact Factor (JIF) [2], h-index [3] and Eigenfactor [4], rely on citation counts. All of the above methodologies suffer from the same base limitation in treating all citations equally.

It has been long established that treating all citations with equal weight is counterintuitive. Garfield, the original proponent of the JIF [2], proposed a range of 15 different reasons a paper may be cited [5]. These can include such reasons as: paying homage to pioneers, substantiating or refuting the earlier work of others, identifying methodologies used or simply giving background information regarding previous work. It can be seen from Garfields original list that simply counting citations cannot paint an entire picture of a papers impact.

Therefore, there is an increasing need in the automatic identification of the nature of a particular citation. Additionally, the growing availability of publication full texts is now making it possible to extend bibliometric studies further than those previously attempted with analysis of abstracts and citation networks alone. Open Access repositories such as that provided by CORE[1] [6] are allowing researchers to utilise the full text of research papers and articles in ways not possible with the meta-data offered by bibliographic databases alone. This has given rise to new areas of study including Semantometrics [7] which attest that the full text of a publication is required to effectively ascertain its impact.

In this paper, we address the problem of identifying influential citations based on publications' full text. The rest of the paper is organised as follows. In Section 2, we introduce key studies on which our work is based. We then discuss the approach for detecting influential citation, providing a critical analysis of features previously applied in this task in Section 3, selecting a set of three key features for further analysis. We present a comparative study of the identified features in Section 4, together with the challenges inherent in this task.

2 Related Work

There have been several different methodologies applied to this task, Teufel (2006) [8] focuses on semantic similarity, identifying cue-phrases in the citing paper such as "we used" or "further to the work of". Citations are classified into 11 types, which are then grouped into the higher grain categories of weak, positive and neutral sentiment. Later studies expand this and, rather than classifying only according to sentiment, attempt to classify citations as either *influential* or *non-influential*.

Hou et al. (2011) [9] first suggest the idea of using an internal citation count based on the full text of a research paper rather than just the bibliography to determine influence. They demonstrate a positive correlation between the number of times a citation occurs and its overall influence on the citing paper. Zhu et al. [10] combine these earlier approaches and suggest a range of 40 classification features including both semantic and metric features to determine influence. Most recently, Valenzuela et al. (2015) [1] made significant efforts to construct a reference set which was publicly released and which this study relies heavily on. They suggest a range of 12 features, many of which show similarity with those of [10].

The features in these studies can be divided into having *internal reliance* or having *external reliance*. The former requires only the full-text of the citing paper whereas the latter relies on additional, external information being available. Furthermore these studies identify three essential feature types. They are semantic based features, similarity based features and metrics / count based features. All of the studies under consideration use a range of different features and test them on different datasets. Consequently, getting a deeper understanding

[1] (http://www.core.ac.uk)

of which of the previously suggested features are most effective at this task is needed.

3 Methodology

The typical workflow for classifying citation types involves the following steps:

- Extracting the full text of the manuscript.
- Parsing the full text to detect the document structure, such as the document metadata, references, citation markers and sections.
- Extracting the features from the document structure, possibly with an enrichment step for features based on external evidence.
- Applying a classifier trained using supervised machine learning approaches.

In the rest of this section, we describe this workflow concentrating on the selection of features used in the citation type classification task.

3.1 Extracting the full-text and parsing

Unless a paper is available in a structure format, such as an XML, there is a requirement for converting the original PDF file into full text prior to analysis. There are numerous tools available for the conversion of PDF to text files. However, automatic text extraction from PDF is known to be problematic [11]. Some tools for inferring the document structure, such as ParsCit [12], require initial conversion to plain text. Others, such as GROBID [13], operate directly on the PDF file.

3.2 Features used by prior studies

One of the overriding aims of this work is to establish which of the previously identified classification features perform most strongly as predictors of citation importance and to use this as a baseline from which to build future work.

We consider the features presented in the two most recent studies. In Zhu et al. (2015) [10] we first see an expansion of the features into a rich range that move beyond simple counting of in-text citations;

- 1. Count-based features
- 2. Similarity-based features
- 3. Context-based features
- 4. Position-based features
- 5. Miscellaneous features

Valenzuela et al. (2015)[1] take a similar approach to the construction of the features list. The 12 features used in this study are;

F1	Total number of direct citations
F2	Number of direct citations per section
F3	Total number of indirect citations and number of indirect citations per section
F4	Author overlap (boolean)
F5	Citation Is considered helpful (boolean)
F6	Citation appears in table or caption
F7	1 / Number of references
F8	Number of paper citations / all citations
F9	Similarity between abstracts
F10	PageRank
F11	Number of citing papers after transitive closure
F12	Field of cited paper.

Table 1. Valenzuela et al. (2015) Feature List

3.3 Selection of features for experiments and comparison

We analysed the 40 features presented by Zhu et al. [10] and 12 features presented in the study of Valenzuela et al. [1]. Of the 40 features, a combination of just 4 features resulted in the best performance of Zhu's model. Adding features beyond this actually lowered the performance. Out of these 4 features, we could not reliably replicate one feature (countsInPaperSecNum). Out of the 12 features of Valenzuela (Table 1), we found three features irreproducible (F3, F5[2], F12), we were unable to reliably replicate two features due to PDF extraction issues (F2, F6) and we elected not to use two features as they rely on external and potentially changing evidence (F10, F11). Two features we tested (F7, F8) did not produce any significant correlation with the gold standard.

Of the three remaining features of Valenzuela, we found a complete overlap of two features (F1-countsInPaperWhole, F4-aux_SelfCite) and a close match on the third (F9-simTitleCore). These three selected features correspond to the best (F1-countsInPaperWhole) feature of Zhu, the worst feature of Valenzuela (F9-simTitleCore) and a third where the opinion regarding the usefulness of this feature was divided between the two studies (F4-aux_SelfCite). In the rest of the paper, we will provide a direct cross-comparison of these features on a single dataset:

- *Number of direct citations* (Integer): This feature is labeled by [1] as 'F1 - Direct Citations' and by [10] as 'countsInPaperwhole. Both of these studies, and the earlier study by [9] find the total number of times a paper is cited to be a strong indicator of academic influence on the citing paper.
- *Abstract Similarity* (Real): This is feature F9 in the [1] study. Whilst [10] tested various similarity based features, none performed better than their randomly assigned baseline (equivalent to the prior distribution of the influential label in their dataset). Valenzuela et al. [1] also listed this as the

[2] We attempted to reproduce this feature, but failed due to Valenzuela's dictionary of cue words not being available.

weakest feature. This feature is calculated as the $tf - idf$ cosine similarity between citing paper abstract and cited paper abstract.
- *Author Overlap / Self-Citation* (Boolean): This feature is labeled F4 by [1] and as 'auxselfCite' by [10]. The two studies differ markedly in their opinion of the value of this feature. While [10] found little correlation between author overlap and influence, [1] listed author overlap as their third best performing feature. It was therefore selected for further investigation.

3.4 Classification

Using the identified features, we perform a binary incidental / influential classification. WEKA 3 [14] was selected as the machine learning toolset in our study.

4 Results

4.1 Dataset

The dataset released by [1] contains incidental/influential human judgments on 465 citing-cited paper pairs for articles drawn from the 2013 ACL anthology, the full texts of which are publicly available. The judgment for each citation was determined by two expert human annotators and each citation was assigned a label. Both a fine-grained (4-Way) label and a binary (incidental / important) label were provided. Using the authors binary classification, 396 citation pairs were ranked as incidental citations and 69 (14.3%) were ranked as influential (important) citations.

It is extremely interesting to note that all studies which employed human annotators to judge citation influence [8, 10, 15, 1] reported a broadly similar ratio of positive examples. This ranged from 10.3% [10], through 14.3% [1], to 17.9% [8]. This is an important finding as it gives a clear indication that only a relatively small percentage of all citations are actually influential at all. All of the studies find that the majority of citations are perfunctory at best. Negative citations are extremely rare and this in itself further increases the difficulties in constructing a balanced reference set. Automatic identification of those influential citations is therefore both a more important and less straightforward task than may be first imagined.

To obtain a clean dataset for our experiments, we first collected the PDF files of the citing and cited papers used by Valenzuela et al. [1] from the ACL Anthology. We processed these papers using pdf2txt [16] to extract metadata, citations, the full text and other document structure information. Any papers where the extraction was not possible or the abstract was not available were then removed. This left us with a dataset of 415 pairs with 355 citation pairs marked as incidental and 60 citation pairs (14.45%) marked as influential citations. As this corresponds to only a relatively small reduction in the number of examples from the original dataset and reflects the original ratio between incidental and

influential citation classes, we consider this dataset to be sufficiently representative for our experiments. We then process the XML files using ParsCit and applied calculations to extract features for each example.

4.2 Analysis and comparison of selected features.

Our experiments tested a range of features and their efficacy as predictors of citation influence. We achieved the best results using the Random Forests Classifier. We tested the model using bagging with 100 iterations and a base learner, using a 10-fold cross-validation methodology. The WEKA toolset was used to generate P/R curves for each of the individual features as well as the combination of all the features (Table 2).

Feature	P@R=0.05	P@R=0.1	P@R=0.3	P@R=0.5	P@R=0.7	P@R=0.9
F1	0.4	0.34	0.33	0.3	0.26	0.21
F4	0.27	0.35	0.14	0.15	0.14	0.14
F9	0.46	0.49	0.21	0.2	0.18	0.16
All	0.5	0.38	0.37	0.37	0.29	0.23

Table 2. Interpolated precision at different recall levels for all features for the random forest classifier.

We also measured the correlation between each of the individual features and the classification given by the human annotators. Valenzuela et al. [1] present their results in terms of P/R values for each feature whereas [10] shows the Pearson correlation with their gold standard. We therefore present the results of our experiments in both formats to allow for accurate comparison. Our work

Table 3. Comparison of results by feature

	Precision@Recall=0.9		Pearson r	
Feature	Valenzuela et al. [1]	Our results	Zhu et al. [10]	Our results
Direct Citations	0.30	0.21	0.330	0.281
Abstract Similarity	0.14	0.14	N/A	0.373
Author Overlap	0.22	0.16	0.020	0.132

confirms the earlier findings reported in [10] and [1] that the number of direct instances of a citation within a paper is a clear indicator of citation influence. We also find that author overlap, or self-citation, does have value as a classification feature. Contrary to the work of [1] we find that the similarity between abstracts is more predictive of citation influence than previously shown.

The correlation of this feature with the reference set (r=0.373, $p < 0.01$, 2-tailed) was the highest of all the features we tested. It is our contention that testing all features using P/R values, at R0.90 masks some of the predictive value of those features when the dataset contains only a small number of instances of

the influential class. Table 3 shows the precision of the random forests classifier at various recall levels. It can been seen from these results that the classifier initially performs quite well and identifies many of the influential cases, however it has difficulty identifying the last few instances which substantially decreases the classifier's performance at $R0.90$. Using Mean Average Precision (MAP) or a similar metric that provides a single-figure measure of quality across recall levels would be a better choice in this case.

Results for Individual Features

Number of Direct Citations - F1: This feature is rated as the highest value in terms of predictive ability by [10] and the second highest by [1]. The latter shows P0.30 at R0.90, however our results demonstrate a slightly lower P value, P0.21 at R0.90.

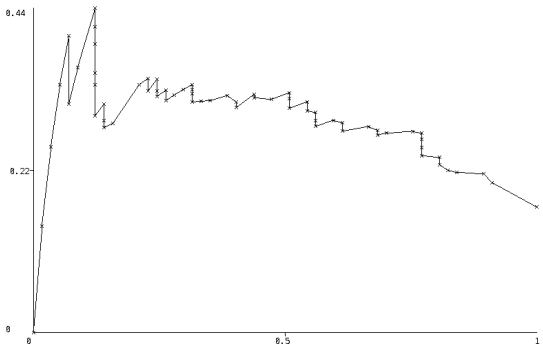

Fig. 1. *P/R curve for feature F1 - Direct Citations.*

[10] lists the equivalent 'countsinPaper_Whole' as the most significant feature of their classifier, with a Pearson correlation coefficient of P0.35. We find a Pearson correlation of P0.28 (significant at the 0.01 level, 2-tailed) for this feature with our dataset. The small difference in this result is likely caused by the differences in the two datasets. Our results therefore confirm that the number of times a citation appears is a strong indicator of that citation's influence.

Author Overlap - F4: The results from the two earlier studies for this feature vary considerably. In the results for [1] this is the third ranked 'most significant feature with P0.22 for R0.90. We find slightly less precision than [1] for this feature; P0.16 at R0.90. [10]'s results show little correlation with their gold standard for the similar feature aux_selfCite (Pearson 0.02). Interestingly, despite the low correlation, this feature was the fourth one selected by their model and

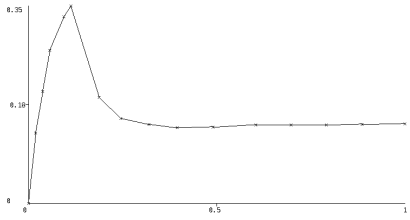

Fig. 2. *P/R curve for feature F4 - Author Overlap.*

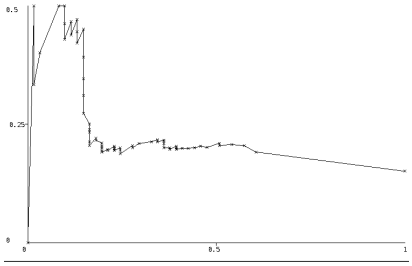

Fig. 3. *P/R curve for feature F9 - Abstract Similarity.*

did indeed improve the performance of the classifier, albeit only slightly. The experiments with our dataset show a far stronger positive correlation, P0.132 (significant at the 0.01 level, 2-tailed), than that found by [10].

Abstract Similarity - F9: Whilst [10] generated many similarity-based features, they did not compare citing abstract and cited abstract. This is somewhat surprising as we consider it to be an interesting feature and one that also seems innately logical. The abstract similarity is calculated as the cosine similarity of the tf-idf scores of the two abstracts. By ensuring that the dataset only contains valid data, i.e. the abstract is available for both citing and cited paper, a direct comparison can be made for this feature with [1] who rank this as the lowest of their twelve features, P0.14 at R0.90.

Here our results are the same as [1],with P0.14 at R0.90. However, the Pearson correlation with the gold standard dataset for this feature is the highest of the three features tested in our experiments. We find a Pearson correlation of 0.373 (significant at the 0.01 level, 2-tailed). This feature was not tested by any of the other earlier studies covered in this work. Our results demonstrate that abstract similarity between citing and cited paper is more predictive of citation influence that previously shown.

4.3 The value of complex features.

It is now over a decade since Teufel [8] first attempted to automate the classification of citation function. This original study and several subsequent ones have suggested classification features ranging from the extremely simple to the extremely complex. Many of these complex features have been shown to have little predictive ability in regards to classifying citation function or importance. Some of the most basic features have been shown to offer the strongest potential in identifying important or influential citations. Our research confirms that one of the most simplistic features, i.e. the number of times a citation appears in a paper, is highly predictive of influence.

Replicating complex features is a non-trivial task unless exact details of how the values for these features were calculated or source code are provided by

the original study. We believe that it is essential that the types and values of all features should be provided as part of the research dataset (as opposed to providing just source prior to feature extraction) to serve as a roadmap in replicating them. Furthermore, features that rely on external datasets, changing evidence (e.g. citations, downloads, etc.), or utilise sets of rules, that are not available to other researchers cannot be replicated by this or any other future study. There has now been a decade of research in this area and the predictive ability of many complex features is still uncertain. This is sorely detrimental to the overall value of the original studies.

4.4 Analysis of PDF extraction.

Both [1] and [10] use ParsCit - the citation parsing tool, based on Conditional Random Fields (CRF). As this is a critical pre-processing stage we conducted experiments to determine the efficacy and accuracy of this tool. GrobID [13] is a similar CRF based tool and was chosen to provide a comparison.

There are several types of errors that can be introduced during the PDF conversion process:

- PDF is a scan and would require OCR.
- Custom encoding instead of Unicode or ASCII.
- Readable XML file not created at all due to failed PDF conversion process.
- References not identified or counted correctly.
- Citations not identified or counted correctly.
- Abstract not extracted correctly or not present in cited paper.
- Title names / Author names misspelled in different parts of the paper.
- Elements being mis-tagged.

We argue that these errors unavoidably impact on the validity of any classification features that are reliant on this process. Of particular concern is the likelihood of citations being either under-counted or over-counted. The results of our experiments demonstrate that this is indeed the case in many instances. To understand the impact of this, we conducted the following experiment. Ten papers were randomly chosen from the Valenzuela dataset and the citation counts for each citation were extracted using both ParsCit and GrobID. The results of both tools were then compared to a manual check / count. [1] and [10] demonstrate that the number of times a citation appears in the body of the text is a significant indicator of influence. There is however a difference in the number of citations identified, depending on the chosen method of parsing. The reference count for five of the chosen example papers is shown in table 5.

These results show that ParsCit correctly identified the exact number of citations in only 40% of cases. GrobID was even less successful. It was exactly correct in only one case and missed a significant number of citations in many others. We argue that this demonstrates a potentially serious failing in current methodologies that rely on PDF extraction for calculation of number of citations.

Paper ID	ParsCit count	GrobID count	Actual Number
C00-2140	33	21	35
W06-0202	17	14	18
W09-1118	25	25	25
E12-1072	31	21	30
P02-1058	13	8	13

Table 4. Comparison of in-text citations counts by extraction method

5 Discussion

One of the major limitations of this and previous studies is the size of the publicly available, annotated, datasets. The study by [1] uses 465 citing / cited paper pairs. The study by [10] uses just 100 papers by 40 authors. Due to the unbalanced split between the incidental and influential classes, our complete dataset contained only 61 examples of the positive (influential) class. We argue that due to the relative sparsity of influential citations a much larger reference set is required. This is equally true for negative citations, which have been shown to be even rarer. Training a classifier when the dataset contains so few instances of the non-neutral classes is problematic and we will address this in future work. The construction of a gold standard dataset containing many thousands of annotated citations, rather than a few hundred, is a significant undertaking but we believe this is a vital step in improving the abilities of the classification models.

There is a noticeable difference between the datasets used by [10] and [1] which warrants further study. The [1] dataset annotation was undertaken by two independent annotators and finds significant value in using author overlap as a classification feature. However, the [10] reference set is annotated by the authors themselves and this study ranks author overlap / self-citation as being of very low importance. It may be that is demonstrates shyness or reticence on behalf of authors to regard their own, earlier, work as being a significant influence. A large scale author-annotated reference set would be extremely helpful in ascertaining the level of this bias when compared to an anonymously-annotated dataset such as that of [8] or [1] Finally we argue that if a citation is considered influential, this original influence remains regardless of external factors or the environment. Therefore, classification features which rely on external and potentially fluid information should be used somewhat cautiously. In future work we will address this issue in greater detail.

6 Conclusions

Of the features we tested, we find the feature *Abstract Similarity* shows the strongest positive correlation for predicting citation influence. We find *Number of Direct Citations* to also be highly predictive and we find *Author Overlap / Self-Citation* to be less predictive but still valuable as a classification feature. It is important to note that many of the features suggested by earlier studies have been shown to have little predictive ability. Additionally, despite significant

efforts ,we were not able to reproduce or validate several of the features used by [1] or by [10]

Furthermore, our results demonstrate that any automatic classification model that relies on PDF extraction in the pre-processing stage is unlikely to capture all of the relevant data which is fundamental to the calculation of the value of some features. We argue that this introduces a level of potential inaccuracy that has not been fully addressed. There is scope for further work surrounding the efficacy and in particular the reproducibility of some of the previously tested classification features. Many of the earlier studies in this domain present results based on sometimes complex and irreproducible features. We contest that this is detrimental to this area of study as a whole and, whilst earlier studies have identified several effective features, having the ability to reproduce them is fundamental to further development in the area of citation classification.

Whilst it may be a relatively easy task for a human being to identify important or influential citations, building a model to automatically classify these citations with any degree of accuracy is a non-trivial task. A larger scale reference set than those used in this and previous studies is essential, particularly due to the inevitably skewed nature of any dataset of citations annotated according to influence or importance.

7 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. Valenzuela, M., Ha, V., Etzioni, O.: Identifying meaningful citations. In: AAAI Workshops. (2015)
2. Garfield, E., et al.: Citation analysis as a tool in journal evaluation, American Association for the Advancement of Science (1972)
3. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America (2005) 16569–16572
4. Bergstrom, C.: Scholarly communication eigenfactor: Measuring the value and prestige of scholarly journals
5. Garfield, E., et al.: Can citation indexing be automated. In: Statistical association methods for mechanized documentation, symposium proceedings. Volume 269., National Bureau of Standards, Miscellaneous Publication 269, Washington, DC (1965) 189–192
6. Knoth, P., Zdráhal, Z.: CORE: three access levels to underpin open access. D-Lib Magazine **18**(11/12) (2012)
7. Knoth, P., Herrmannova, D.: Towards semantometrics: A new semantic similarity based measure for assessing a research publications contribution. D-Lib Magazine **20**(11) (2014) 8

8. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP '06, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 103–110
9. Hou, W.R., Li, M., Niu, D.K.: Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. BioEssays **33**(10) (2011) 724–727
10. Zhu, X., Turney, P., Lemire, D., Vellino, A.: Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology **66**(2) (2015) 408–427
11. Lawrence, S., Giles, C.L., Bollacker, K.: Digital libraries and autonomous citation indexing. Computer **32**(6) (1999) 67–71
12. Councill, I.G., Giles, C.L., Kan, M.Y.: Parscit: an open-source crf reference string parsing package. In: LREC. Volume 2008. (2008)
13. Lopez, P.: Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In: International Conference on Theory and Practice of Digital Libraries, Springer (2009) 473–474
14. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)
15. Di Iorio, A., Nuzzolese, A.G., Peroni, S.: Identifying functions of citations with citalo. In: Extended Semantic Web Conference, Springer (2013) 231–235
16. Ma, J.: pdf2text 1.0.0. Available from: https://pypi.python.org/pypi/pdf2text/