Identifying and Capturing the Semantic Aspects of Citations

Thesis

How to cite:


For guidance on citations see FAQs.

© 2020 David Matthew Pride

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.21954/ou.ro.000146ff

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
Identifying and Capturing the Semantic Aspects of Citations

David Pride

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Knowledge Media Institute
The Open University
Milton Keynes
United Kingdom
Abstract

This dissertation presents new work in understanding the nature of citations, their potential different purposes and how they are used. This research addresses a fundamental question: ‘Should all citations be treated equally?’ It explores the concept of influential citations and examines key features for their identification. It then investigates the key challenges in capturing these semantic aspects of citations and then presents solutions that overcome current limitations in the domain.

The dissertation provides an overview of current bibliometrics that use raw citation counts and then details the problems associated with these methodologies. It then presents an overview of two areas in which bibliometrics and citation data are already being applied: information retrieval and research evaluation. The first
study (Chapter 3) focuses on the second of these areas, the use of
citation data in research evaluation; it is the largest investigation
to date into the correlation between peer review and bibliometrics
at the institutional / discipline level, using data from the UK’s
Research Excellence Framework (REF2014). This study was the
first to identify strong correlations between simple citation-based
indicators and aggregate peer review results in approximately one
third of domains covered by the REF2014 process, notably those
domains in which the peer review panels used citation data to
inform their peer review decisions.

There is already wide-scale usage of bibliometrics and cita-
tion data in research evaluation, not only in exercises such as the
REF2014, but also in other Performance Related Funding Exer-
cises (PRFS) globally. Furthermore, citation-based metrics, such
as the h-index (Hirsch [2005]) or Journal Impact Factor (JIF)
(Garfield [1983]), are being used to measure individual academics,
despite the aforementioned limitations and risks associated with
these methodologies.

Critically, all of these metrics, without exception, treat all cita-
tions equally, even citations which refute or negate previous work. This seems not only illogical, but also limiting. There is far greater opportunity in understanding not only that a particular piece of research was cited, but why it was cited.

The study presented in Chapter 4 then examines the key features in identifying influential citations and replicates a range of features tested in prior works. The work then evaluates the current state of the art in the automatic identification of citation purpose using machine learning and natural language processing techniques and addresses the key challenges in this domain. It demonstrates that datasets compiled by earlier works are of limited size, largely due to the selected annotation methods, and are drawn from one or two domains at most. Experimental results show that this affects the accuracy of classification models built using these datasets.

The dissertation then presents a novel methodology, using first authors as citation annotators, and a new platform for the collection of citations annotated according to both purpose and influence. These tools were used to produce the largest dataset of
annotated citations, covering 19 different disciplines, which can be used to foster new research in this domain and improve the performance of classification models.
I owe an enormous debt of gratitude to a long list of people that have guided me to the point where I am able to write these words. The final words of a four-year, multi-continental journey that ended with the submission of this thesis. In the first instance, to my supervision team; Dr Petr Knoth, Professor Zdenek Zdrahal and Dr Miriam Fernandez. In particular to Petr for five years of unwavering support, advice and opportunities to grow and develop my craft as a researcher. The countless hours of help with writing and with understanding new theories will always be a hugely important part of my life and I cannot thank him enough. It was incredibly tough in parts, but the support I have had has enabled me to complete something that I would never have thought within my reach.
I also owe huge thanks to Emeritus Professor Marian Petre, erstwhile champion of the Post Graduate Forum group and incredible pillar of support throughout this process, including not only proof-reading every word of this thesis, but probing and questioning what was written to ensure I understood every part of it. This was such a tremendous help in improving the work and finally seeing me to the end.

To everyone in the PG Forum group, and its current facilitator Dr Daniel Gooch, thank you. The regular sessions and opportunities to learn, laugh and sometimes cry together were a central part of my journey and undoubtedly helped me cross the finish line.

To Professor John Domingue, lab director at KMi, for also believing in me and for making me feel a part of the KMi family. And to my colleagues in the CORE team, Matteo Cancellieri, Dr Nancy Pontika, Lucas Anastasiou and Samuel Pearce – you folks are awesome. Amazing thanks too to my Obi-Wan Kenobi, Dr Dasha Herrmannova, who completed this same journey a couple of years before me. You were a true source of inspiration and I
can’t thank you enough. I’ll always remember the British Library.

And to those who believed in me from before the start of this journey, including my wonderful advisers from my Masters program, Professor Maria Schilstra and Dr Parivash Ashravi who encouraged me to apply for the PhD. studentship at KMi.

And finally to my darling, my incredible wife, who has put up with five long years of excitement and elation as well as the darkest times along this very long and winding path. Without her I undoubtedly would not be where I am today. Her encouragement saw me enter academia to complete my Masters and her continued faith in me has been a rock when all else around me has often seemed tenuous and ethereal. As I end this journey, she begins hers. I hope I can now do for her that which she has done for me.
Contents

Abstract 1

Acknowledgements 5

Publications 20

Invited Talks 22

Abbreviations used 24

Key Findings and Contributions 25

1 Introduction and Background 26
   1.1 Introduction ........................................ 27
   1.2 Motivation ......................................... 37
       1.2.1 The role of citations in information retrieval 38
1.2.2 The role of citations in research funding and evaluation .......................... 42

1.3 Problem statement ............................................. 48

1.3.1 Peer review and citation data in research evaluation ................................. 48

1.4 Research Gap ...................................................... 53

2 State of the Art .................................................. 58

2.1 Introduction: terminology .................................... 59

2.2 From Biblio- to Sciento- to Alt-: the evolution of metrics .............................. 63

2.3 A brief history of metrics ...................................... 67

2.4 The foundations of the Science Citation Index ..................................... 70

2.5 Counts and controversy: the introduction of the Journal Impact Factor .............. 73

2.6 Imbalance of journal citations ................................ 76

2.7 Technical limitations of the Journal Impact Factor ............................... 77

2.8 Alternative citation-based methodologies ............................................ 80

2.8.1 The $h$-index and its contemporaries ........................................ 80
2.8.2 h-index values in Scopus and Web of Science 84
2.8.3 SCImago Journal & Country Rank . . . . . . 87
2.8.4 Citation network metrics: PageRank and
Eigenfactor . . . . . . . . . . . . . . . . . . . . . . 87
2.9 Altmetrics and the age of the social academic . . . 89
2.10 The use of bibliometrics in research evaluation . . . 92
2.11 DORA, Leiden and Responsible Research Metrics . 97
2.12 Defining and identifying reasons for citation . . . . 100
2.13 Text and structure extraction with natural language
processing . . . . . . . . . . . . . . . . . . . . . . 104
2.14 Deep learning models for natural language processing 108
2.15 Citation analysis and classification . . . . . . . . 109
2.16 Conclusions . . . . . . . . . . . . . . . . . . . . 113

3 Peer Review or Citation Data? Predicting Institutional Rankings in the UK’s Research Excellence Framework 116
3.1 Introduction . . . . . . . . . . . . . . . . . . . . 117
3.1.1 Peer review and citation data in the U.K.’s Research Evaluation Framework (REF)  .  118

3.1.2 Use of citation data in the REF peer review process  .  121

3.1.3 Destruction of individual rankings for outputs  .  123

3.2 Dataset  .  124

3.3 How well do peer review judgments correlate with citation data at the institutional level?  .  128

3.4 How well can citation data predict peer review based institutional rankings?  .  133

3.5 Discussion  .  137

3.6 Summary of findings and conclusion  .  141

4 Effective Features for Identifying Citation Influence  .  145

4.1 Introduction  .  146

4.2 Methodology  .  150

4.2.1 Extracting the full text and parsing  .  151

4.2.2 Features used by prior studies  .  151
4.2.3 Selection of features for experiments and comparison .......................... 152
4.2.4 Classification ......................................................... 156
4.3 Analysis and results ..................................................... 156
  4.3.1 Dataset ............................................................. 156
  4.3.2 Analysis and comparison of selected features 159
  4.3.3 Results for Individual Features ................................. 161
  4.3.4 Replication of earlier studies ................................. 166
  4.3.5 Analysis of PDF extraction ................................. 167
4.4 Discussion .......................................................... 170
4.5 Summary of findings and conclusion ................................. 172

5 Overcoming the Key Challenges in Capturing the Semantic Aspects of Citations 176
  5.1 Introduction .......................................................... 177
  5.2 Selection of a citation classification schema .......................... 179
  5.3 Annotation methodology ........................................... 181
  5.4 A new platform for annotating citations .......................... 184
  5.5 ACT dataset - collation and comparison .......................... 186
5.5.1 Comparison of datasets and collection methodologies ............................................. 189
5.5.2 Authors or third-party annotators ................................................................. 190
5.5.3 Comparison of class distribution .............................................................. 192
5.5.4 The effects of changes in class schema ..................................................... 194
5.6 Citation classification experiments ............................................................... 198
5.6.1 Citation classification using Jurgens et al. and ACT datasets ....................... 199
5.6.2 Domain coverage of the ACT dataset ....................................................... 203
5.6.3 Does language change across domains? .................................................. 204
5.6.4 Citation classification using single and multi domain datasets .................... 206
5.7 Summary of findings and conclusion ........................................................... 210

6 Findings, Implications and Future Work ....................................................... 213
6.1 Findings and contributions ............................................................................ 214
6.2 Implications and Limitations ........................................................................ 220
6.3 Future work .................................................................................................... 225
6.4 Conclusion ...................................................................................................... 229
List of Figures

1  Mapping of research questions, findings and chapters  25

1.1  Monthly uploads to ArXiv 1992-2016  . . . . . .  39
1.2  The growth of scientific articles in Crossref  . . .  40
1.3  U.S. Spending on research as a proportion of GDP
   1962-2017  . . . . . . . . . . . . . . . . . . . . . . .  42
1.4  The linear model of innovation, redrawn from Sutherland et al. [2011]  . . . . . . . . . . . . . . . . . . . . . . . .  44

2.1  Tag cloud scaled according to results on Google Scholar  . . . . . . . . . . . . . . . . . . . . . . . .  60
2.2  The rise of metrics - Google Scholar ‘in-title’ search
   1975-2016  . . . . . . . . . . . . . . . . . . . . . . .  63
2.3 From ISI to MAG - a timeline of citation networks and metrics 68
2.4 Hirsch’s h-index and its contemporaries 83
2.5 A typical citation classification workflow 111
3.1 Citation enrichment workflow used in dataset creation 124
3.2 Plot of REF GPA / citation data correlation against MAG coverage for each UoA 132
4.1 P/R curve for feature F1 - Direct Citations 162
4.2 P/R curve for feature F4 - Author Overlap 164
4.3 P/R curve for feature F9 - Abstract Similarity 164
5.1 The ACT Platform in use. A. Details of current citation for annotation. B. Author’s PDF manuscript. C. Hover-over pop-out for current citation details. D. In-text highlighted citation marker. E. Annotation section. (A live demo of the ACT Platform is available at: https://youtu.be/qQz_gB0Yjx4) 185
5.2 Comparison of dataset breakdown by class of citation from Jurgens et al. [2016] and Pride et al. [2019] ............................................. 193

5.3 Comparison of datasets when class granularity is reduced. ............................................. 195

5.4 Domain breakdown of the Academic Citation Typing (ACT) dataset ............................................. 204

5.5 How language changes across domains ............. 206

5.6 F1 performance of NLP models on single-domain and multi-domain datasets ......................... 208
List of Tables

2.1 Underlying problems with the JIF. Recreated from Seglen [1997] . . . . . . . . . . . . . . . . . . . . . 75
2.2 Average citation rates per paper by discipline (Source: Clarivate [2016] . . . . . . . . . . . . . . . . . . . 79
2.3 Classification schema used in previous studies . . . 103
3.1 Classification of outputs submitted to REF2014 . . 119
3.2 REF result metrics . . . . . . . . . . . . . . . . . . 120
3.3 UoAs with the highest mean-citations-per-paper. . 127
3.4 Dataset statistics . . . . . . . . . . . . . . . . . . 127
3.5 Pearson Correlation Coefficient between REF GPA output rankings and citation data. For all results $p < 0.001$ . . . . . . . . . . . . . . . . . . . . . . 130
3.6 Rankings by GPA and predictions produced using
\( med_{2017} \) and \( med_{2014} \) respectively for the three most
highly correlated UoAs

3.7 Rank prediction quality for the top 10 UoAs with
the highest mean-citations-per-paper

3.8 Comparison of the prediction performance of this
study with Mryglod et al. [2015]

3.9 REF UoA Sub-panels using citation data in REF2021

4.1 Valenzuela et al. (2015) Feature List

4.2 Description of features used for comparison with
earlier studies

4.3 Interpolated precision at different recall levels for
all features for the Random Forest Classifier

4.4 Comparison of results by feature

4.5 Comparison of in-text citations counts by extrac-
tion method

5.1 The citation classification schema

5.2 Breakdown of ACT dataset by class of citation
5.3 Cross-study comparison of dataset collation and annotation type .......................... 189
5.4 Examples of citations and labels from ACT dataset. 192
5.5 Comparison of micro F1 results from Cohan et al. [2019] ................................. 197
5.6 Classification models used in experiments ............................................................. 200
5.7 Comparison of results (micro F1 score) of NLP models on Jurgens et al. [2016] and ACT datasets ................................................................. 201
5.8 Composition of single- and multi-domain datasets ............................................. 207
5.9 Comparison of results (F1 score) for multiple models on single- and multi-domain datasets ................................................................. 209
Publications

The chapters of this thesis are based on the following publications.

**Chapter 3**

- Pride D, Knoth P. Peer review and citation data in predicting university rankings, a large-scale analysis. In International Conference on Theory and Practice of Digital Libraries 2018 Sep 10 (pp. 195-207). Springer, Cham. *Best Paper Award*

**Chapter 4**


Chapter 5


• Pride D, Knoth P. An Authoritative Approach to Citation Classification. In 2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL ’20), August 1–5, 2020, Virtual Event, China. IEEE
Invited Talks and Presentations

In addition to the presentations given at the above conferences, sections of this body of work were presented at the following invited talks and presentations.


- Pride D. *Beyond the count. Enhancing citation value with contextual typing* German Centre for Higher Education Research and Science Studies (DZHW) Opening speaker for two day workshop on Text Mining in Science Studies. Humboldt-University, Berlin. 2019

- Pride D. *All citations are created equal. Only some are
more equal than others Presentation at The Open University 3MT (Three Minute Thesis) competition, 2019. Highly Commended.
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DORA</td>
<td>San Francisco Declaration on Research Assessment. A worldwide initiative covering all scholarly disciplines that recognises the need to improve the ways in which research assessment is conducted, particularly in regards to the use of quantitative metrics.</td>
</tr>
<tr>
<td>GPA (in REF)</td>
<td>Grade Point Average. The average score for all outputs submitted by an institute to the U.K.’s Research Excellence Framework in 2014. GPA data is available aggregated to the institute / discipline level.</td>
</tr>
<tr>
<td>JIF</td>
<td>Journal Impact Factor. A purely citation based metric for measuring the overall impact of a journal. Originally conceived at the ISI, then sold to Thompson-Reuters, now provided by Clarivate Analytics as a part of their Journal Citation Reports.</td>
</tr>
<tr>
<td>MAG</td>
<td>Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study.</td>
</tr>
<tr>
<td>MAS</td>
<td>Microsoft Academic Search - the forerunner to MAG</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing is a sub-field at the boundary between computer science and linguistics. NLP’s aim is to develop tools and process to help analyze large text corpora.</td>
</tr>
<tr>
<td>PRFS</td>
<td>Performance-Related Funding Systems. National scale exercises for distribution of research funding based on academic performance, measured using peer review, bibliometrics or a combination of both. Examples include the U.K.’s Research Excellence Framework and the Italian VQR (Valutazione della Qualità della Ricerca)</td>
</tr>
<tr>
<td>REF</td>
<td>Research Excellence Framework. The U.K.’s largest single component of research funding. £1.6b allocated annually to U.K. HEIs based on the results. Last conducted in 2021 (Postponed from 2020 due to Covid-19)</td>
</tr>
<tr>
<td>SCI</td>
<td>Science Citation Index. Published the first citation based index for journals. (Published quarterly from 1963.)</td>
</tr>
<tr>
<td>UoA (in REF)</td>
<td>Unit of Assessment (in the U.K.’s REF process) There were 36 UoAs covering all scientific domains. Each UoA appointed a separate panel of expert peer-reviewers.</td>
</tr>
</tbody>
</table>
Key Findings and Contributions

Figure 1 shows the mapping for each of the research questions to each of the relevant chapters. Also summarised are the key contributions and findings from each chapter.

Figure 1: Mapping of research questions, findings and chapters
Chapter 1

Introduction and Background
1.1 Introduction

Citation Data and Bibliometrics

Citation analysis [Moed, 2006] and bibliometrics Broadus [1987] are used widely as tools for assessing the impact of research and for the distribution of research funding, and are the basis for the hiring and retention policies at many institutions. Bibliometrics are purely quantitative, but are often used as a proxy for qualitative measures. They are highly attractive due to their simplicity in using them at scale. However, there is currently insufficient evidence to demonstrate the connection between citation rates and research quality. As noted by Wallin [2005], this is paradoxical in light of the widespread use of bibliometrics such as the Journal Impact Factor (JIF) or h-index in research assessment exercises.

Bibliometric indicators, such as the JIF [Garfield, 1972] and h-index [Hirsch, 2005], rely solely on citation counts. Whilst these measures have gained wide acceptance and popularity, there are warnings from some quarters regarding their validity and efficacy. These deficiencies will be reviewed in detail in Chapter 2.
There are other metrics that use methodologies based on network and graph theory that have been used to rank and quantify academic papers; these include the PageRank algorithm, developed by Page et al. [1999], which was originally used to rank web pages according to the network of links that connects them. More recently Bergman [2012] introduced the EigenFactor, which similarly uses the network of citation connections between papers to assess the contribution of each. Somewhat indicative of the domain, the wryly titled ‘YetRank’ (from ‘Yet another ranking algorithm’) is another example of this type of network ranking algorithm from Hwang et al. [2010].

Altmetrics

Newer types of quantitative analysis, grouped together under the banner of Altmetrics, have arisen in recent times, leveraging the reach of the internet and the growth in social media. Rather than counting citations, these methods count other usage, such as paper views or downloads, mentions on social media (Twitter, Facebook, etc.), ‘bookmarking’ using services such as Mendelay.
Reader, and mentions on other blogs and websites. These efforts are an attempt to overcome some of the limitations of citation-based metrics. Immediacy is a potentially large advantage of Altmetric measures. Citations, as discussed, take time to accrue, whilst much of the data used in calculating Altmetric scores starts to become available within a very short time span after publication. There is, however, wide-ranging debate as to whether these new metrics measure what they attempt to measure, and whether they are subject to the same limitations and caveats as other bibliometric methods [Thelwall and Kousha, 2015], [Bornmann and Leydesdorff, 2015].

**Citation Data and Cumulative Advantage**

It can be seen that there are many uses of citation data, and of metrics built on citation data. However there are many factors that can determine why a particular paper is cited over another, and this often has little to do with the quality of the paper in question. Whilst it is logical that authors will cite prior works that are important or relevant to their own work, there are many
other factors to consider when this information is used at scale.

De Solla Price [1965] coined the term ‘cumulative advantage’ to refer to the effects of ‘the rich getting richer’, i.e., the more citations a paper has, the more likely it is to receive citations in the future. This was demonstrated also to hold true for highly-regarded scientists, institutions, and even journals. This phenomenon was more famously termed ‘the Matthew effect’ by Merton [1968] in *The Matthew effect in science: the reward and communication systems of science are considered*. His original inspiration for the term is somewhat archaic, coming from the New Testament and the Gospel according to St. Matthew: “For to all those who have, more will be given” (Matthew 25:29). Indeed, these two papers themselves can be seen as an example of the Matthew effect in action. De Solla Price’s work preceded Merton’s by three years, yet the latter, by the arguably more famous and influential scientist, has become the established terminology. Notably, Merton’s paper has also been cited three times more than de Solla Price’s.

There are many other factors which influence the likelihood of a paper receiving citations. In an in-depth review of citation
analysis and methodologies, MacRoberts and MacRoberts [1989] defined seven multi-layered problems with the use of citation data as the basis for any form of measurement of quality. These include the non-citing of both formal and informal influences, biased citation practices, self-citation, variations in citation rates across domains, publication type, and even the nationality of the researchers, and when the paper was written.

As recently as 2015, Bornmann and Leydesdorff [2015] demonstrated the clear correlation between the number of authors a paper has and the number of citations it will receive. Domain also plays a key role in determining the number of authors [Patience et al., 2017], which then, as shown, has a direct influence on the number of citations.

Biases, unconscious and otherwise, have long been a factor in academia. These also appear in citation practices. Female researchers are less likely to be published in journals with a high Journal Impact Factor, and less likely to be cited than their male counterparts [Cislak et al., 2018].
A further potential problem is the temporal nature of citations. There is an unavoidable delay between publication and citation that is dependent on numerous factors. Burton and Kebler [1960], using the analogy of radioactive material, described the rate of obsolescence of scientific literature as its half-life. Clarivate Analytics (formerly Thomson Reuters) used a similar approach to define a metric entitled the cited half-life of a journal in their Journal Citation Reports. This is the average age of citations of a paper, the half-life life age determining when a paper will have received, on average, half the citations it will receive in total. In many fields this will be several years; in Physics, for example, the cited half-life is 4.9 years. As citations take time to accrue, bibliometrics based on citation counts can therefore disadvantage early-career researchers.
Reasons for Citation

Citation behaviour has been extremely well investigated over the last fifty years. The study of citations can effectively be grouped into two over-arching classes. There is the informational / theoretical interpretation of citations [Leydesdorff, 1998], which encompasses bibliometrics and other studies attempting to use citation data to measure, analyse, and track research outputs. Alongside this is the arguably more complex sociological interpretation, that is: why do authors cite each other?

As long ago as 1957, Merton [1957] wrote that citations were acknowledgment of credit for new ideas, as well as a form of social recognition. Eugene Garfield, founder of the Science Citation Index and viewed by many as the founder of bibliometrics, suggested 15 reasons why authors may cite one another [Garfield, 1972]. These are: paying homage to pioneers; giving credit for related work; identifying methods, equipment, etc.; providing background reading; correcting one’s own work; correcting the work of others; criticizing previous work; substantiating claims; alert-
ing researchers to forthcoming work; providing leads to poorly disseminated or uncited work; authenticating data and classes of fact; identifying original publications in which an idea was discussed; identifying the original publication describing a concept or term; disclaiming the work or ideas of others (negative claims); or disputing priority claims of others (negative homage). In 1977, Ina Speigel-Rosing defined 13 citation types that have been used in many studies in this domain [Spiegel-Rosing, 1977].

These two areas, informational and sociological, have most often been viewed independently, and there are no bibliometric measures that encompass citation intent. However, we suggest that these two areas are inextricably linked. Without an understanding of why citations are occurring, bibliometric measures that rely on citation counts alone ignore potentially valuable information.

Methodology Overview

In chapter 3, this dissertation examines global Performance-based Research Funding Systems (PRFS) that use peer review, bibliometrics, or a combination in their evaluation
processes. The results of many of these exercises are the basis for research funding allocation in the respective country, and this analysis looked in-depth at the U.K.’s most recent research evaluation exercise, the Research Evaluation Framework (REF). The REF used both peer-review and citation data to evaluate research outputs from U.K. institutions, and £1.6b in research funding was distributed based solely on those results. The study presented in Chapter 3 was the first to identify that peer review outcomes in the REF and simple bibliometric indicators are highly correlated at the institutional level in certain domains. This leads directly to two further questions: first, what additional value did the peer review process add to the REF review process (this is of particular relevance due to the considerable costs involved), and second, if basic citation data correlate with peer review in many instances, to what extent can measures that give a more detailed view of citation intent or purpose provide better information to peer reviewers? As mentioned, in addition the quantitative aspects of citations, there have been many attempts to define and categorise the qualitative aspects too. Numerous previous studies have sug-
gested differing reasons for, and types of, citation. Further, several previous studies have introduced the notion of classifying citations according to type, following a citation classification schema. Chapter 5 presents an analysis of the prior studies in this domain, the results of which were then used to inform the development of a new methodology and platform for collecting annotated citations.

This work investigated whether improvements to current purely-quantitative, citation-based metrics can be made by leveraging the semantic, qualitative aspects of citations. These two paradigms, referred to by Leydesdorff [1998] as informational / technological and sociological, have existed almost independently of one another. There have been numerous studies regarding the use of informational / technological citation data and the creation of bibliometrics, and many further studies debating the efficacy of these metrics. There have been fewer studies into understanding the sociological and semantics aspects of citation and their applications. This work is positioned at the interface between these two domains.
1.2 Motivation

This section presents an overview of the motivation behind this work and reviews some of the current applications that leverage citation data.

Citation data and bibliometrics are used in a wide range of applications. It is common to see results in academic search engines ranked according to number of citations as well as recency and relevance to the initial search term. This is the case for Scopus, Web of Science and Google Scholar. Many other ranking systems, for institutions, journals, and individuals, also use citation data as a key indicator.

Further, in part due to the continually expanding scientific literature and the need for institutions, funders, and government to demonstrate return-on-investment and value-for-money, data – notably citation data – is often used as a proxy for the quality and impact of research. There are, however, numerous pitfalls to this approach, which are discussed in depth in the following sections.

Garfield himself noted: “Evaluation studies using citation data
must be very sensitive to all divisions, both subtle and gross, between areas of research; and when they are found, the study must properly compensate for disparities in citation potential” [Garfield, 1963], (Pages 289-291).

Bibliometric measures are often used as a shortcut for quality or impact. However, evidence that these metrics actually measure what they are intended to is conflicting, and this has wide ramifications in many areas. This is highly relevant to individual researchers, to institutions, to funding agencies, and to Government. A brief overview of this landscape is presented here.

1.2.1 The role of citations in information retrieval

As of January 2018, there were some 28,100 peer-reviewed scientific journals, and new scientific papers are now being published at a rate approaching 2.5m papers per year [International Association of Scientific, Technical and Medical Publishers, 2016]. The sheer volume of new research now being produced on an annual basis is far beyond the capacity of a single researcher to investigate even the narrowest of domains without effective search tools.

38
Figure 1.1 shows the monthly uploads (to 2016) to the Arxiv pre-print server and the dramatic increase in the volume of papers being submitted to just this one repository.

![Figure 1.1: Monthly uploads to ArXiv 1992-2016](image)

Formed in 1999, and incorporated in 2000, the Publishers International Linking Association (PILA), an independent not-for-profit organisation, created the Crossref system for collaborative reference linking. From the original 12 founding members, the organisation has grown to 16,870 members (2019 figure). The total amount of scientific literature to which Crossref links grew from
6m publications in 2002 to over 100m in 2018. As of September 2020, 109,440,179 articles are now referenced by Crossref. This again indicates the enormous growth in the amount of scientific literature available to researchers (Figure 1.2)

![Figure 1.2: The growth of scientific articles in Crossref](image)

It is clear that tools are required to help scholars locate relevant literature amongst the burgeoning wealth of published articles. However tools that rank publications based on citation counts alone may not be the most effective in returning relevant search results. Indeed, academic search results based on citation counts are likely to simply increase the incidence of the previously-mentioned Matthew effect. If the top results returned by academic
search engines are merely the most-cited articles, these articles will continue to accrue citations at a greater rate than lower-ranked articles, due purely to visibility. Search engine users, both academic and mainstream, are human and will almost inevitably follow the most-visible links. Studies indicate that 91% of searchers never go beyond the first page of results [van Deursen and van Dijk, 2009].

One recent development is an attempt to move away from this methodology. Microsoft Academic Search’s (MAS) results are based on the semantic similarity of the documents and the search term, and far less weight is given to the citation count of the articles. Another organisation using a similar model is Semantic Scholar ¹, from the Allen Institute for Artificial Intelligence.

Attempts are now being made to enhance current purely-quantitative bibliometrics with some of the semantic aspects of citation; however, there are serious limitations with prior studies in this area which will be covered in-depth in Chapter 5.

¹https://semanticscholar.org
1.2.2 The role of citations in research funding and evaluation

To use the United States as an example, research funding as a proportion of GDP fell from 9-12% in the 1960s to less than 4% in 2016 (see Figure 1.2.2). As research budgets have been squeezed over the last five decades, there has been a fundamental shift in attitudes toward the end goals of research.

Figure 1.3: U.S. Spending on research as a proportion of GDP 1962-2017

Today, funding bodies are held to account for not only the scientific impact of funded research, but also the societal impact. In some domains, this is more easily monitored, for example: medical...
ical research that leads directly to new drugs or treatments, or aerospace research that delivers advances in technology. However, in many research areas, the impact of academic outputs is far more difficult to measure. Further, as noted by Nightingale and Scott [2007]: “Research that is highly cited or published in top journals may be good for the academic discipline but not for society.” (Pages 543-553).

Sutherland et al. [2011] identified the three main benefits that research brings to the society: improved life quality or sustainability, economic benefits, and contribution to knowledge. Figure 1.4, also suggested by Sutherland et al. [2011], shows the Linear Model of Innovation, in which research moves from ‘pure’ research to ‘applied’ research, which is then developed and disseminated, resulting finally in societal benefits. There is, however, a strong argument that not all research can be so clearly categorised, nor does all research necessarily lead to immediate and quantifiable societal benefits. There are many notable cases in which the eventual outcomes or benefits of the research were not foreseen at the time the research was conducted. For example, Theodore
Maiman, the developer of the first laser, failed to have his original paper accepted for publication. The laser was seen as “a solution seeking a problem.” according to Maiman’s colleagues [Hughes, 2010]. It took almost two years and many revisions before the paper was accepted by Nature. Nevertheless, this one article ignited an entirely new scientific domain and eventually resulted in a Nobel Prize for its author.

![Figure 1.4: The linear model of innovation, redrawn from Sutherland et al. (2011)](image)

Government, institutions, and individuals are these days typi-
cally required to justify the funds allocated to research, particularly when public funds are at stake, and there is a wide range of policies and practices in place globally for the tracking of research outputs.

Again, bibliometrics are used as a proxy for high-quality research. Whilst many metrics can be used to show the quantity of research, few, if any, capture the essence of research quality. Many research funding systems, however, do rely in part or in whole on bibliometric indicators, which again raises the question as to the accuracy and veracity of these metrics and their suitability for the purposes for which they are being applied.

In the view of Martin and Irvine [1983] (Pages 61-90), “indicators based on citation counts are seen as reflecting the utility, rather than the quality or importance, of the research work”. Further, Nieminen et al. [2006] demonstrated that “quality is not necessarily associated with the number of citations [a paper receives]”.

Directly tied to research funding is the need for research evaluation. In addition to traditional grant or patronage funding, there is growing use of Performance-based Research Funding Systems
(PRFS) in many countries. The PRFS that first linked research performance directly to funding was introduced in 1986 in the UK in the guise of the Research Selectivity Exercise, later re-labelled the Research Assessment Exercise (RAE). This exercise was repeated every five or six years until its replacement by the Research Excellence Framework (REF) in 2008. The REF was next undertaken in 2014, and it is due to be conducted again in 2021. These research evaluation systems fall largely into two categories: those that focus on peer review judgements for evaluation, and those that use a bibliometric approach. The UK and New Zealand both have systems weighted heavily toward peer review. Northern European countries other than the UK tend to favour bibliometric methodologies, whereas Italy and Spain consider both peer review judgments and bibliometrics. Research Evaluation Systems overall have dual and potentially dichotomous ends: first identifying the best quality research, and second, in many cases, the distribution of research funds. There is, however, a large variance in the level of institutional funding granted, based on the results of these exercises. The UK’s Research Councils distribute
£1.6 billion annually entirely on the basis of the results of the Research Excellence Framework (REF) which is the largest single component of research funding. At the other end of the scale, the distribution of funds based on the results of the Finnish PRFS is just 3% of the total research budget. Furthermore, the PRFS in Norway and Australia are both used for research evaluation but are not used for funding distribution. Those PRFS that rely solely or more heavily on bibliometrics tend to be those that use the outputs more for tracking research progress and distribute less funding based on the results Hicks [2012].

Peer-review-based PRFS are hugely time consuming and costly to conduct. For example, the cost of the 2014 U.K. REF was in the region of £250m. The use of bibliometrics in many of these exercises is therefore viewed as a necessary time- and cost-saving measure.
1.3 Problem statement

The previous section defined two areas in which citation data are currently used despite their limitations: information retrieval, and research evaluation and funding.

The use of citation data is very much established in each of these two areas. The results of this work are therefore applicable across a wide range of use cases.

1.3.1 Peer review and citation data in research evaluation

This section briefly introduces how citation data is already influencing peer review and research evaluation. Since the birth of academia, peer review has been the standard framework for qualitative analysis of academic output. Academics wishing to publish must first have other academics in their field review, evaluate, and comment on their work. The author must also accept recommendations for changes to the work from the peer reviewers. In this way, it is argued, only ‘good’ science is published, ‘bad’ science is
rejected, and each published article or paper becomes a building block for other researchers and academics. In an ideal world, this would be the case on every occasion. We do not live in an ideal world. There is currently an urgent discourse regarding the reproducibility of scientific results, with Ioannidis [2005] and Flier [2017] noting that many articles that have passed the peer review process and been published in highly prestigious journals suffer from a lack of reproducibility. Additionally, is has been shown that peer review can fail to detect work that is simply incorrect or badly conducted [Ioannidis, 2015], and there are many well-known examples of highly-cited articles that were later retracted [Davis, 2012].

Additionally, there are many known biases in the peer review process, as noted by Priem and Hemminger [2010]. There are also the time and cost issues associated with the peer review process.

Notwithstanding all of the above, peer review is still the current ‘gold standard’ for evaluating good science [Nicholas et al., 2015]. However, papers held in high regard by peer reviewers do not necessarily go on to become highly-cited papers. One longitudinal
study by Sen and Patel [2012] that reviewed citation rates for award-winning papers over a 25-year period found that 25% of those papers were never subsequently cited, and an additional 30% were cited just once.

Peer review also has dual applications. First, it is used to assess submissions prior to publication, whether that be a journal paper, a conference article, or a full manuscript. Second, peer review is used during research evaluation exercises that can occur at any unspecified time after initial publication. Many of these evaluation exercises use bibliometrics to complement and inform the peer review process, yet it has been shown in HEFCE [2015] that many bibliometric indicators show little correlation with peer review judgements at the article level.

Furthermore, many prior studies have demonstrated that the myriad current metrics based on simply counting citations exhibit a range of biases such as institutional bias Vaccario et al. [2017], gender bias [Gallagher and Barnaby, 1998], and bias against early-career researchers [Jannot et al., 2013]; are open to gaming and manipulation [López-Cózar et al., 2012]; and do not necessarily
reflect the true impact of a particular study or piece of work [Borchardt et al., 2018]. In the view of Martin and Irvine [1983], (Page 61): “Indicators based on citation counts are seen as reflecting the utility, rather than the quality or importance, of the research work.” and demonstrate that The number of citations to a publication is not a direct reflection (a “measure”) of its quality or importance, nor even of its impact.”.

The UK’s research evaluation exercise, the REF, is used to quantify the academic outputs of all higher education institutions. Each submission to the REF is considered by a domain-specific panel of expert peer reviewers. Submissions are then rated as U (unclassified) or given a rating of one to four stars, where four stars indicate that the work is world-leading in terms of originality, significance, and rigour. Of the 36 subject areas in the REF, 11 of the subject peer review panels chose to use citation rates to inform their decision making process. HEFCE [2014e].

This is despite the conclusion by a £1.8m study funded by the Higher Education Funding Council for England (HEFCE) that: “Citation information is not sufficiently robust to be used formu-
laically or as a primary indicator of quality in the REF” HEFCE [2014d].

Chapter 2 presents a systematic review of current bibliometrics; their use in research evaluation and the failings of these metrics.

Any metric based solely on citation counts cannot consider either the type or sentiment of the original citation. Furthermore, whilst many hundreds of bibliometric methods have been developed, to the best of our knowledge, none accounts for the type, polarity, or sentiment of the citations being counted.

The two areas discussed in the motivation section – information retrieval and research funding / research evaluation – have different goals, yet rely, at least in part, on citation data or bibliometrics. The work in this dissertation examines how to identify and capture the semantic aspects of citations, and how this additional information can then be leveraged. These qualitative aspects of citation are ignored in current ranking methodologies or bibliometric measures.
1.4 Research Gap

The previous sections have demonstrated that citation data, and bibliometric indicators built on these purely-quantitative measures, are already being used in a host of areas as a proxy for research quality. This is particularly problematic when these indicators are used in research evaluation exercises or in promotion and hiring policies enacted by institutes.

This research is centered on the hypothesis that citations contain far more information than can be expressed in a single digit, that the reasons behind citations are varied, both in terms of purpose and centrality to the citing paper. The Leiden Manifesto (Hicks et al.), produced in 2015 by a group of expert bibliometricians, gave 10 principles for evaluating research and commented *Reading and judging a researcher’s work is much more appropriate than relying on one number*. This statement is categoric that the actual full text content of a piece of research is what should be assessed when determining its value rather than a metric that is merely a proxy indicator of this value.
Further, whilst many different reasons for citation have been identified, and attempts to define a standard schema for citation types have been made, there has been little cross-pollination between quantitative applications of citation data and studies focused on the qualitative reasons for citation.

This leads to the overarching question at the centre of this work: **How can we effectively identify and capture the semantic aspects of citations at scale?**

Both of the areas discussed in the motivation section are worthy of detailed examination. Of particular interest is the use of citation data in research evaluation and funding. Chapter 3 addresses the following question:

**RQ1: To what extent do peer review and current bibliometric indicators correlate at the institutional / discipline level?**

Whilst Studies by Hicks et al. [2004] and Bornmann and Leydesdorff [2015] looked at the overall relationship between bibliometrics and research funding. The study by Ravenscroft et al. [2017] investigated the correlation between bibliometrics (using
citation data from CiteSeer, arXiv an PubMed Central) and REF Impact case studies and found that “citation-based metrics for impact measurement do not correlate well with REF impact results.”

We are unaware of any other previous work have found no previous work that has looked in depth at the correlation between overall REF GPA results, calculated as an averaged score for all outputs submitted to the UK’s REF2014 exercise, and citation data.

The results from the REF are available in the public domain, which presented an opportunity to study citation usage in a research evaluation scenario. Using this data, combined with the full text of articles (outputs) submitted to the REF and citation data retrieved from Microsoft Academic Graph, enabled a substantial study into how citation data can influence peer review panels in research evaluation exercises.

A strong correlation would indicate that metrics, where available, can lessen the burden of peer review on national Performance-based Research Funding Exercises, such as the REF, leading to considerable cost savings, while a weak correlation would suggest
each methodology provides different insights.

Chapter 4 begins to investigate the additional information contained in a citation in regards to identifying whether some citations are more influential than others, and whether this can be detected automatically. This chapter addresses the following question;

**RQ2: What are the key features that indicate a citation is influential, rather than just incidental?**

This is achieved by first recreating one of the earliest studies in the domain, that of Valenzuela et al. [2015], and then performing a comprehensive review of the features used for citation classification. We find only two of the previously tested features show predictive ability in identifying influential citations and also demonstrate some difficulties in recreating features from these earlier works.

The tools used for extracting a paper’s full text and citation data from its original PDF file are compared, as this is a critical stage in the process and is fundamental to the success of the classification process.
Chapter 5 demonstrates that the reasons for citation can be seen within the semantic structure of the citation itself, and then reviews prior work that has attempted to classify citations automatically according to type and influence, thereby addressing the question:

**RQ3: How can we identify and overcome the key challenges in capturing the semantic aspects of citations?**

The approach used in the earlier studies of Teufel et al. [2006a], Jurgens et al. [2016], and Valenzuela et al. [2015] mean they all suffer from limitations of scale. The datasets compiled by these studies are fairly small, with around 2-3,000 annotated citations. The annotation of citation data is slow and therefore costly; it may also be subjective. These studies used annotations provided either by domain experts or by the authors of the study. This placed a restriction on the size of the dataset that could be compiled. A fully-scalable method for citation annotation is therefore needed to go beyond what is possible with current collation methods. Chapter 5 introduces a novel solution for this.
Chapter 2

State of the Art
2.1 Introduction: terminology

This investigation began with wanting to understand the quantitative methodologies currently in use that attempt to measure the impact of scientific research. Many global performance-related funding systems (PRFS) use metrics as a proxy for research quality. As the usage of metrics is so endemic, the focus was on initially gaining a full understanding of their history and development. Comparative searches were undertaken on Scopus and Google Scholar. These searches encompassed six paradigms that were identified as the key points in the development of metrics overall, namely:

- Bibliometrics
- Scientometrics
- Informetrics
- Webometrics
- Altmetrics
• Cybermetrics

The tag cloud in Figure 2.1 is scaled to represent the number of titles returned by Google Scholar for each of the search terms. Whilst Bibliometrics still dominates the academic landscape, several of the other search terms now show rising prominence. Scientometrics has been growing since the 1970s, with large numbers of papers published in the domain. At the other end of the scale is Webometrics and the almost-invisible Cybermetrics.

Webometrics, proposed by Ingwersen and Almind [1997], uses Bibliometric and Informetric approaches, but its overarching focus is on mapping the network of connections that comprises the world wide web. A second definition of Webometrics was (some might say confusingly) suggested by Thelwall in 2009 with specific meaning for the Social Sciences: “...the study of web-based
content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study” [Thelwall, 2009] (Page 6). It should, however, be noted that there has been little adoption of this label. A Google Scholar search for articles with ‘Webometrics’ in the title returned scant results (and an average of approximately 30 papers per year being produced). There was an even starker result for the term ‘Cybermetrics’. This term did not gain traction amongst the wider academic community, and Google Scholar found just a handful of papers with Cybermetrics in the title since 2013.

Informetrics is another label that was designed to be overarching and to encompass the fields of both Bibliometrics and Scientometrics. Introduced in 1979 by Otto Nacke, Informetrics is defined as the “quantitative study of the production, dissemination and use of all forms of information” Nacke [1979] (Page 220). Again, however, a search on Google Scholar, Scopus, or Web of Science reveals few published papers that have adopted this term, in comparison to Scientometrics or Bibliometrics. Several papers that do mention Informetrics simply do so in a comparison to Bib-
liometrics and Scientometrics. It is therefore also interesting to note that, as of August 2017, The Journal of Informetrics had a noticeably higher Journal Impact Factor (2.92) than The Journal of Scientometrics (2.14).

In recent years, since the publication of The Altmetrics Manifesto in 2010 Piwowar [2012], Altmetrics (for ‘alternative metrics’) has shown a steady rise in terms of the volume of papers and articles being produced. Altmetrics leverages public, web-based data such as paper download statistics, social media mentions and bookmarks to infer the impact of a research article.

One simple way to visualise the prominence of any particular metric is to conduct a year-by-year search for papers mentioning the specific metric in the title of the paper. Figure 2.2 shows the normalised results for the number of academic papers with the three most prominent metrics named in the title from 1975 through 2016 on Google Scholar. Bibliometrics and Scientometrics show both exponential growth and very close correlation over time. Since the introduction of Altmetrics in 2010, there has been increasing interest in using alternative quantitative measures that
do not rely on citation data.

2.2 From Biblio- to Sciento- to Alt-: the evolution of metrics

Following the overall arc of the development of metrics, it can be seen that Altmetrics has not replaced standard bibliometrics in the way initially envisioned in the Altmetrics manifesto. Indeed,
the citation-based methodologies employed in Bibliometrics and Scientometrics seem now to be ingrained firmly within academic practices, despite efforts in many quarters to move beyond the current status quo.

Bibliometrics and Scientometrics are linked by both an evolutionary process and a crossover in application. Whereas de Bellis [2009] (Page 3) notes that Bibliometrics “stresses the material aspect of the undertaking: counting books, articles, publications and citations”, Scientometrics has a wider aim, to quantify scientific achievement.

In 1963, Derek de Solla Price published his seminal work, *Little Science, Big Science* [de Solla Price, 1963]. In the chapter ‘Transcience and Continuance’, Price first introduced *Price’s Law*, which notes that 25% of academics are responsible for 75% of academic output. The highly-productive authors also seem to be the highly-influential authors, and Price goes to some length in attempting to quantify this.

*Price’s Model* is introduced in another essay, *Networks of scientific papers* de Solla Price [1965]. This revolutionary model
represents the first attempt to map citations as a network graph, building on the earlier work of Herbert Simon. The model states that the number of new citations a paper receives is directly proportional to the number of citations it already has. This is referred to as cumulative advantage and is also a demonstration of the Matthew Effect [Merton, 1988]. Price also first noted the apparent shortening in timescales of citation, which he termed the Immediacy Factor: “This is demonstrated by the time distribution: much-cited papers are much more recent than less-cited ones.” [de Solla Price, 1986] (Page 514). This led to the creation of the Immediacy Index, which is provided as a part of Journal Citation Reports by Clarivate Analytics (formerly Thompson Reuters). The index is simply the number of times a paper is cited in its first year of publication. This metric is again an attempt to provide researchers and institutions with meaningful, citation-based statistics over and above the Journal Impact Factor, which is discussed in depth in the following section.

Following the work of de Solla Price, Scientometrics gained traction, aided by the ever-increasing number of publications and
the development of computing technology with the power to analyze this information. The *Journal of Scientometrics* was first published in 1978 and describes itself as covering the latest development in the fields of Scientometrics – and Informetrics. Much later, in 2007, the *Journal of Informetrics* was established by Leo Egghe, creator of the *g-index*, which is discussed in the following section. This journal describes itself as covering the latest developments in the field of Informetrics and Scientometrics. This again points to the somewhat confusing nature of the terminology used in the domain.

The number of papers, journals, and conferences covering the metrics domain grew steadily throughout the 70s, 80s, and 90s. In 1987, Egghe organised the first International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval. 1993 saw the formation of International Society for Scientometrics and Informetrics (ISSI) which took place at the International Conference on Bibliometrics, Informetrics and Scientometrics held in Berlin.

One can see that the terms Bibliometrics, Scientometrics an
Informetrics have been used variably and interchangeably over the years with distinct overlap in journals and conferences dedicated to the domain.

2.3 A brief history of metrics

Not everything that can be counted counts, and not everything that counts can be counted. [Cameron, 1963]

As a caveat for academics attempting the quantification of research impact, the above quotation should bring into sharp focus the underlying difficulties associated with the task at hand. Somewhat ironically, this quotation is often mistakenly attributed to Einstein. The original is actually from William Cameron’s *Informal Sociology, A Casual Introduction to Sociological Thinking* [Cameron, 1963]. Coincidentally, 1963 was the year of the first publication of the Science Citation Index.

In the early part of the twentieth century, and long before Eugene Garfield founded the Institute for Scientific Information, Bibliometric studies were being undertaken by a range of aca-
Figure 2.3: From ISI to MAG - a timeline of citation networks and metrics
demics and, notably, librarians. In an aside that could be from 2020, but is actually from 1927, two librarians, P. L. K. Gross and E. M. Gross, noted that it was “no longer sufficient to compile a list of journals one would feel was indispensable [to their college library]” [Gross and Gross, 1927]. Their stated reasons? The bias, intended or otherwise, of the compiler of the list, and the range of journals currently available. In what was one of the first such exercises, these forward-thinking academics tabulated the citations from the 1926 edition of The Journal of The American Chemical Society. The cited papers were ranked by the number of citations received “in such a way that the relative importance of any single periodical for any five year period can be seen.” (Page 386) This enlightened exercise produced a ranking for each journal paper and demonstrates remarkable parallels with Garfield’s Journal Impact Factor, which would not appear for almost another forty years.
2.4 The foundations of the Science Citation Index

As a PhD student in the 1950s, Eugene Garfield joined the Johns Hopkins University’s Welch Medical Library Indexing Project, for which his research initially focused on designing algorithms to transpose chemical nomenclature into chemical formulae. Whilst the Welch Project focused on indexing scientific literature, Garfield began to focus on the citations contained within said literature. Rather than indexing the documents, Garfield started to collect and index the citations within these documents, and the network of connections that arose from the citations.

An early corollary of the Scientific Citations Index can be found, not in the academic literature, but in legal history. In the United States, The United Kingdom, and other countries that model their legal systems on these countries, modern justice and democracy is founded on the principal of *stare decisis*, that is, ‘reasoning by precedent’. In passing judgment, courts and judges are bound by precedents set by earlier cases. These precedents
are collated and published as reference tools for the legal profession. In the U.S., the best-known of these was *Shepard’s Citations*, which began in 1873 [Ogden, 1993].

Academic citations perform a similar role, recognising the prior efforts of others and thus aiding the incremental improvement of science as a whole. At the core of citation extraction and analysis lies an intrinsic problem that warrants an entire area of study of its own, and that falls largely outside the scope of this review: the automatic extraction, retrieval, and analysis of textual information. The central task for the Welch project was examining how, and if, the new emergent computing technology could be used to automate the indexing of medical literature [de Belleris, 2009]. This built on the fundamental work of Harris [1951] in transposing and analysing textual content. The three methodologies that form the basis of this are lexical analysis, syntactic analysis, and semantic analysis.

In 1960, Garfield moved fully into the commercial arena and founded the Institute for Scientific Information. Its first publication, entitled *Current Contents*, was a weekly list of academic
journals’ tables of contents. This seemingly simple idea was pop-
ular amongst researchers as it offered up-to-date information on
up-and-coming research, often pre-print. The list of domains cov-
ered by this publication grew over the years to include informa-
tion science, library services, biomedicine, and physical sciences
[de Bellis, 2009].

The seed of what would become the Science Citations Index
(SCI) can be traced to 1963, when the U.S. National Institutes
of Health (NIH) was called to account for the impact of its own
research projects. The NIH believed that building a citation net-
work or index was well suited to this purpose, and Garfield’s ISI
was approached to carry out the task. Funding was acquired,
and the first Genetics Citation Index was published. Garfield of
course realised the cross-discipline relevance of such an index, and
the first edition of the SCI was published shortly thereafter. This
index also revealed a new and hitherto unforeseen application of
such an index, namely tracking and assessing the impact of a par-
ticular paper, author, or institution using the citation networks.

The publication of the print edition of the SCI continued quar-
terly, growing exponentially in terms of the size of the index, and the stature that a journal’s inclusion in the index conveyed. ISI was eventually sold to Thompson Reuters in 1992, and thus the index moved into the digital realm, re-branded as the *Web of Knowledge*. In 2002, the platform was again renamed and became the *Web of Science* (WOS).

### 2.5 Counts and controversy: the introduction of the Journal Impact Factor

Since 1975, the SCI has been supplemented by the annual publication of the *Journal Citation Reports*, which compare over 9,000 individual journals and rank them according to various metrics, the most widely-known and controversial of these being the Journal Impact Factor (JIF).

This metric was originally devised by Garfield to help select journals for the SCI. It is the number of citations received in that year by articles published in that journal during the two preceding years, divided by the total number of articles published in that
journal during the two preceding years. Thus all journals can be ‘ranked’ according to this measure.

There have been years of debate surrounding the validity of this method, particularly when the methodology is applied to rating the impact of individual researchers and institutions, rather than journals. Many previous studies have evaluated the JIF and its application. Both Browman [2015] and Seglen [1997] summarise the limitations and vagaries of the JIF and find more than 20 concerns with its usage as a general measure of research impact.
Problems associated with the use of the Journal Impact Factor

<table>
<thead>
<tr>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal impact factors are not statistically representative of individual journal articles</td>
</tr>
<tr>
<td>Journal impact factors correlate poorly with actual citations of individual articles</td>
</tr>
<tr>
<td>Authors use many criteria other than impact when submitting to journals</td>
</tr>
<tr>
<td>Citations to “non-citable” items are erroneously included in the database</td>
</tr>
<tr>
<td>Self citations are not corrected for</td>
</tr>
<tr>
<td>Review articles are heavily cited and inflate the impact factor of journals</td>
</tr>
<tr>
<td>Long articles collect many citations and give high journal impact factors</td>
</tr>
<tr>
<td>Short publication lag allows many short term journal self citations and gives a high journal impact factor</td>
</tr>
<tr>
<td>Citations in the national language of the journal are preferred by the journal’s authors</td>
</tr>
<tr>
<td>Selective journal self citation: articles tend to preferentially cite other articles in the same journal</td>
</tr>
<tr>
<td>Coverage of the database is not complete</td>
</tr>
<tr>
<td>Books are not included in the database as a source for citations</td>
</tr>
<tr>
<td>Database has an English language bias</td>
</tr>
<tr>
<td>Database is dominated by American publications</td>
</tr>
<tr>
<td>Journal set in database may vary from year to year</td>
</tr>
<tr>
<td>Impact factor is a function of the number of references per article in the research field</td>
</tr>
<tr>
<td>Research fields with literature that rapidly becomes obsolete are favoured</td>
</tr>
<tr>
<td>Impact factor depends on dynamics (expansion or contraction) of the research field</td>
</tr>
<tr>
<td>Small research fields tend to lack journals with high impact</td>
</tr>
<tr>
<td>Relations between fields (clinical v basic research, for example) strongly determine the journal impact factor</td>
</tr>
<tr>
<td>Citation rate of article determines journal impact, but not vice versa</td>
</tr>
</tbody>
</table>

Table 2.1: Underlying problems with the JIF. Recreated from Seglen [1997]
2.6 Imbalance of journal citations

Every journal consists of many articles and many citation to other articles. However, journals contain highly-cited articles and highly-impactful articles (whether these two are the same thing is also a matter of debate), alongside less-cited and less-impactful articles. There is long-standing evidence that a high proportion of academic papers is never cited at all [Larivière et al., 2009]. Citations are known to follow a power-law distribution [de Solla Price, 1986], with academic citations also largely following Pareto’s Law which in this example states that 20% of all academic papers will be responsible for 80% of all citations [Gupta et al., 2005].

The JIF is an average, and therefore the few highly-cited articles, and in some cases a single article, can have demonstrable effect in raising the impact factor of a particular journal. An example of this effect is the article ‘A short history of SHELX’, which stated: “This paper could serve as a general literature citation when one or more of the open-source SHELX programs (and the Bruker AXS version SHELXTL) are employed in the course of a
"crystal-structure determination" Sheldrick [2008] (Page 112). Subsequently, this one article received over 6,600 citations in 2006/7, and consequently the JIF of *Acta Crystallographica Section A*, the journal in which this article appeared, rose from 2.051 in 2008 to 49.926 in 2009. At the time, this represented a higher impact factor than either *Nature* or *Science*. Yet in subsequent years, once the effect of this single paper was removed from the JIF calculation for this publication, its impact factor immediately fell back to its former levels [Thompson Reuters, 2010].

### 2.7 Technical limitations of the Journal Impact Factor

There are myriad reasons why the JIF can be argued to be flawed, particularly when it is applied at researcher or institute level. Garfield himself repeatedly warned that citation data can be misinterpreted or inadvertently manipulated for improper purposes at the level of the individual researcher [Garfield, 1983]. In addition, there is the compounding factor as to what is considered a
citeable source within a particular journal. Letters, editorials, and the like are not considered to be citeable, even though these are indeed often cited. Furthermore, there is the distinct issue with the different citation rates (in terms of both volume and time scale) in different disciplines. Table 2.7 shows the average citation rates per article for a range of disciplines; there is significant variance in these averages.

Molecular biology and genetics, immunology, and neuroscience are the areas with the highest average citation rates, whereas computer science and mathematics exhibit the lowest averages.

The time scale of citation must be considered. Brody et al. [2006] (Page 1060) note that:

“In physics, the cited half-life of an article (the point at which it has received half of all the citations it will ever receive) is around 5 years.”

This is longer than the cited half-life of 24-36 months for the health sciences and approximately the same as mathematics at 48-60 months. All of these figures, however, fall outside the rolling
<table>
<thead>
<tr>
<th>Discipline</th>
<th>Average Citation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular biology</td>
<td>25.62</td>
</tr>
<tr>
<td>Immunology</td>
<td>21.81</td>
</tr>
<tr>
<td>Neuroscience</td>
<td>19.47</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>17.25</td>
</tr>
<tr>
<td>Microbiology</td>
<td>15.79</td>
</tr>
<tr>
<td>Space science</td>
<td>14.3</td>
</tr>
<tr>
<td>Clinical medicine</td>
<td>12.93</td>
</tr>
<tr>
<td>Pharmacology</td>
<td>12.2</td>
</tr>
<tr>
<td>Environment/ecology</td>
<td>11.35</td>
</tr>
<tr>
<td>Psychiatry/psychology</td>
<td>11.26</td>
</tr>
<tr>
<td>Chemistry</td>
<td>11.19</td>
</tr>
<tr>
<td>Geosciences</td>
<td>9.7</td>
</tr>
<tr>
<td>Physics</td>
<td>8.97</td>
</tr>
<tr>
<td>Plant/animal science</td>
<td>7.74</td>
</tr>
<tr>
<td>Agricultural sciences</td>
<td>7.05</td>
</tr>
<tr>
<td>Materials science</td>
<td>7.03</td>
</tr>
<tr>
<td>Economics/business</td>
<td>6.22</td>
</tr>
<tr>
<td>Engineering</td>
<td>4.76</td>
</tr>
<tr>
<td>Social sciences, other</td>
<td>4.67</td>
</tr>
<tr>
<td>Computer science</td>
<td>3.75</td>
</tr>
<tr>
<td>Mathematics</td>
<td>3.48</td>
</tr>
</tbody>
</table>

Table 2.2: Average citation rates per paper by discipline (Source: Clarivate [2016])

two-year window for the JIF calculation, and therefore it does not consider these time frames. In an effort to counter this imbalance, in 2007, Thompson Reuters introduced the *Five Year Journal Impact Factor*.
2.8 Alternative citation-based methodologies

2.8.1 The \textit{h-index} and its contemporaries

There have been many attempts over the years to quantify scientific impact at the level of the institution and the individual researcher. The intrinsic problems associated with the JIF when applied to institutions and researchers rather than journals has spurred the search for more effective and accurate measures. The European Association of Science Editors recommends:

\begin{quote}
\textit{“...that journal impact factors are used only – and cautiously – for measuring and comparing the influence of entire journals, but not for the assessment of single papers, and certainly not for the assessment of researchers or research programmes either directly or as a surrogate.”} Piwowar [2012]
\end{quote}

There have been many attempts, but few successes. One metric that has gained traction in recent years and is now used by many, is the \textit{h-index}, devised in 2005 by Jorge Hirsch. Further to the
work of de Solla Price in separating quantity from quality, the h-index was formulated to reflect both productivity and impact. The h-index of a researcher is a single figure, calculated thus: a researcher with an index of $h$ has published $h$ papers, each of which has been cited in other papers at least $h$ times [Li et al., 2010]. The logic is to reward consistently-cited researchers. A single paper with many citations, or lots of papers each with few citations, will not increase an author’s h-index. It is the cumulative effect of multiple citations over an extended time period that differentiates the h-index from a simple count of citations.

There are, however, still numerous caveats for the use of the h-index. Again, as with basic citation counts, it discriminates against early-career researchers, because their papers have not yet had the time to accumulate citations. Moreover, the h-index does not account for the differing citation rates across disciplines, as shown in Table 2.7. The h-index can be normalised for interdisciplinary comparisons, but this is not done as a matter of course when these figures are presented by Scopus, WOS, and the like, so one must still be wary when attempting to make cross-disciplinary
Hirsch’s original h-index became a catalyst for many other researchers who expanded the basic calculation for the h-index in numerous different ways. In 2007, Burrell proposed the *Hirsch Rate Burrell* [2007], which attempts to remove the shifting nature of the h-index over time; however, there are numerous limitations to this method. Radicchi et al. [2008] attempted to correct the cross-disciplinary problems by formulating a reference standard for each discipline, based on the average number of citations per domain. Sidiropoulos et al.’s ‘*contemporary*’ and ‘*trend*’ h-indices refine the basic index to consider the age of the publication being cited, and the age of the other citations that publication has received. Sidiropoulos et al. [2007]

Bornmann et al. [2010] proposed \( h^2_{\text{lower}} \), \( h^2_{\text{center}} \), and \( h^2_{\text{upper}} \), to enhance the original figure and to differentiate between publications in high-visibility journals (upper) and those in low-visibility journals (lower). The authors claim that, in this way, this metric adds a degree of qualification as well as quantification. There have been myriad other attempts by other researchers
to enhance the original h-index. Batista et al.’s $h^i$-index [2006], Schreiber’s $h^m$-index [2009], and Wan et al.’s $h^p$-index [2007] attempt to calculate scores based on the number of co-authors. As recently as 2014, efforts were continuing, with Harzing and Alakangas [2016] introducing the $h^{IA}$-index, another attempt to normalise for both differences in disciplines and the age of the researcher.

All of the above metrics attempt in one way or another to refine and improve Hirsch’s original index.

Figure 2.4: Hirsch’s h-index and its contemporaries.
2.8.2 h-index values in Scopus and Web of Science

None of the above suggested refinements to the original h-index have yet been widely adopted by either academia or the information providers. Both Scopus and WOS give the ability to calculate the h-index for any author; however, this comes with its own proviso. As is shown in the following section, each provider has differing coverage and citation counts. This affects any ranking or calculation based on those counts, and this has specific impact on the h-index of any author, depending on which data is used.

Google Scholar offers its own citation-based ranking at the individual author level, the $i_{10}$ index. This is a simplistic count that is one step removed from a simple count of citations. The $i_{10}$-index value for an author is the number of publications with at least 10 citations. Of course, those citations must be available to Google Scholar.

This again highlights the issues caused when differing institutions are using different sources for information. This becomes particularly critical when such metrics are being used to evaluate
researchers, to assign tenure and promotions, or to inform hiring practices. There is a further caveat that can be applied to metrics that are defined using either the Scopus or Web of Science datasets: both base the calculation of the h-index on the author’s name; this leaves the door open for contamination of the results by different authors with the same name.

Suggested in 2006 by Leo Egghe, the *g-index* is, like Hirsch’s h-index, primarily an author-level metric. For the complete set of papers by a particular researcher, the *g-index* is calculated as the (unique) largest number, such that the top g articles received (together) at least $g^2$ citations. Egghe [2006] reported his reasoning, that this calculation adds weight to the more highly-cited papers, thereby increasing the scores for highly ‘visible’ researchers. The h- and g-indexes were calculated for all Price Medal holders, and the values were compared. In his conclusions, Egghe states that ranking these academics using the g-index “resembles more the overall feeling of ‘visibility’ or ‘life time achievement’ ” Egghe [2006] (Page 143). For a deeply Scientometric study, this is a subjective conclusion. The g-index also suffers from one of the major
limitations of the h-index, in that it takes time to accrue citations and consequently disadvantages younger researchers. It has also been noted by Rousseau and Leuven [2008] that both of these metrics may encourage academics to ‘rest on their laurels’. Citation counts can only increase; an author therefore does not have to continue to produce research in order for the author’s h-index or g-index to increase. Wildgaard et al. [2014] undertook a comprehensive review of the current state of metrics and discovered and analysed 108 different types, including ‘h-based’ and ‘non h-based’ metrics. Of the 108 analysed, 79 were described as being ‘potentially’ useful. They state that the wide availability and ubiquity of data sources has markedly improved the ability of researchers to collate and use this data as a means of tracking impact. Their conclusions, however, noted an extremely important caveat that “superior author-level indicators...were too complicated for end-users to apply them in practice.” (Page 153) This is of course critical and presents a barrier to wide-spread adoption of these types of metrics.
2.8.3 SCImago Journal & Country Rank

In 2008, academics from three Spanish universities formed SCImago and partnered with Elsevier to launch an open-access citation database which draws its content directly from Scopus. This was an interesting development, undoubtedly designed to compete directly with the services and rankings being offered by Thompson Reuters in the form of the JIF and its variants. Its two primary outputs, the SCImago Journal Rank and the SCImago Country Rank, are again h-index-based metrics.

2.8.4 Citation network metrics: PageRank and Eigenfactor

There is one further branch of the bibliometric tree that uses citations as its base, but then treats them as a weighted graph network to assign importance or influence to each node in the graph. This method was first suggested by Pinski and Narin [1976] in relation to the publication of physics research in journals. They introduced three metrics, rather than a single one, to compensate for
the length of each of the papers, and also the number of papers published by each journal. This graph-based approach for ranking influence has now become a part of daily life; this is the method used by Page et al. [1999] in their eponymous and now widely-known PageRank algorithm, which forms the basis of how Google ranks web pages.

Another recent ranking metric, the EigenFactor, is based on a similar approach, but in this case the corpus for the graph can be the articles published in a journal, in a research field, by a conference, or from an institution. It was proposed by Bergstrom [2007] and has gained some traction within the research community. It has also been adopted by publishers, including Clarivate Analytics (formerly Thompson Reuters), which includes this metric as a part of its standard statistics.
2.9 Altmetrics and the age of the social academic

Interestingly, in its original format, Altmetrics did not consider citation counts and instead ranked the impact or value of an article according a range of online metrics. These descriptions were formalised in in 2014 by Piwowar [2012]; at this point, citation counts were included in the definition:

- Viewed - HTML views or PDF file downloads
- Discussed - Mentions on blog posts, social media posts, journal comments etc.
- Saved - Using book-marking services such as Mendelay or CiteULike.
- Cited - Standard citations.
- Recommended - Citations by editorials, press mentions, etc.

Several publishers, including BioMed, PLOSOne, and Elsevier, are now providing Altmetric data to their subscribers. Altmetrics
has also been commercialised in recent years, with Altmetrics.com becoming a proprietary provider and recently partnering with Elsevier. Other suppliers of similar services have also appeared. Of note are ImpactStory, responsible for the re-written definitions above, and also Plum Analytics, which provides a range of Altmetric analytical tools.

The latter half of the first decade of the 21st century saw a seismic shift in the global usage of the Internet, and the birth of the social networks that have come to encompass most of our lives. A 2014 survey by the London School of Economics [Tinkler, 2014] found that 99% of universities have a Facebook page, 97% use Twitter, 88% have a YouTube channel, and 55% are publishing online blogs in addition to publishing news articles and press releases via traditional media. With respect to individual academics, it is more difficult to gain an accurate picture of the scope of social media use for purely academic purposes. This is largely due to the somewhat amorphous nature of a person’s different social media profiles. A person may use Facebook for connecting with family and friends, and Twitter for business or
academic purposes. This separation is evident in the oft-quoted mantra that appears in many social media personal biographies; “..the views here are my own and do not represent those of my employer / institution / etc.” A 2015 survey conducted by Inside Higher Education which polled 2,175 academics about their attitudes toward social media found that 75% of faculty respondents said they don’t use social media to express their views on scholarship or politics [Straumsheim, 2015]. Academics, it seems, are still undecided about whether social media is worth their time. 35% of those surveyed said Facebook, Twitter, and other platforms are useful for communicating with the broader public, but another 36% disagreed with this statement. A more interesting question, in the context of this research, would have been how academics are using the platform to communicate with each other.

Another survey by Tinti-Kane et al. [2010] mirrored these figures. It found that, whilst 80% of academics do have a social media account of one sort or another, only approximately 30% of these are using these platforms for discourse with students, faculty, or other academics. In an interview for the Times Educational
Supplement; Adam Dinsmore, evaluation officer at the Wellcome Trust, said that the trust used Altmetrics to identify potentially interesting “narratives” about the influence of research it funded. However, he said, “a high Altmetric score did not necessarily imply that the research had made a crucial scientific impact”, and it was important to consider whose attention papers had caught Jump [2014].

2.10 The use of bibliometrics in research evaluation

There has long been wide ranging and often contentious discussion regarding the efficacy of both peer review and bibliometrics and whether one or other, or both should be used for Research Evaluation. In many Performance-based Research Funding Systems, the endemic indicators used largely reflect the models of success shown by traditional, established, wealthy and largely English-speaking research universities (Hazelkorn [2015]). These types of indicators are therefore not relevant or accurate for institutes on the global
scale. Several studies have specifically investigated the correlation between the results of different nations’ peer review focused Performance-based Research Funding Systems and bibliometric indicators. Anderson et al. [2013] finds only weak to moderate correlation with results from the New Zealand PRFS and a range of traditional journal rankings. The highest correlation is \( r = 0.48 \) with the Thomson Reuters Journal Citation Report. However Anderson states that this may be due to the much broader scope of research considered by PRFS processes and the additional quality-related information available to panels. Contrary to Anderson, Smith [2008] used citations from Google Scholar (GS) and correlated these against the results from the New Zealand PRFS in 2008. He found strong correlation, \( r = 0.85 \) for overall PRFS results against Google Scholar citation count.

A comprehensive global PRFS analysis was conducted by Hicks in 2012. Hicks states there is convincing evidence that when PRFS are used to define league tables this creates powerful incentives for institutions to attempt to 'game' the process, whether in regards to submission selection or staff retention and recruitment poli-
cies Hicks [2012]. A more recent study by Herrmannova [2018] investigated how quality was measured as a part of a number of PRFS exercises in several different countries and found wide-ranging adoption of bibliometric measures being used as a proxy for research quality in many of these exercises.

A UK government funded report, The Metric Tide, was published in 2015 and gave a range of recommendations for the use of metrics in research evaluation exercises. The Metric Tide study had access to the anonymised scores for the individual submissions to the REF and was therefore directly able to compare on a paper by paper basis the accuracy of a range of bibliometric indicators. This study tested correlations with a range of different bibliometric measures and found correlation with rankings for REF 4* and 3* outputs for some UoAs. Metrics found to have moderately strong correlations with REF scores for a wide range of UoAs included: number of tweets; number of Google Scholar citations; source normalised impact per paper; SCImago journal rank and citation count HEFCE [2015]. Yet none of these measures have been adopted in the latest round of the exercise, REF
In a related study, Mryglod et al. [2015] used departmental h-index aggregation to predict REF rankings based on outputs. Their work was completed before December 2014 when the REF results were published and contained ranking predictions based on their model with some degree of success. They also experimented by normalising the h-index for each year between 2008 and 2014 but surprisingly found little evidence that timescale played a part in the strength of the correlations they found. An ad hoc study by Bishop [2013] also found a moderate to strong correlation between departmental research funding based on the results of the UK’s Research Assessment and Evaluation (RAE) exercise conducted in 2008, and departmental h-index.

Mingers et al. [2017] recently completed an investigation that collected total citation counts from Google Scholar (GS) for the top 50 academics\(^1\) from each UK institute and he found strong correlations with overall REF rankings. To our knowledge, the study

\(^{1}\)If there were not 50 academics then the total number of academics on GS for that institute was used.
in chapter 3 is the first large-scale in-depth study that investigates the correlation between citation data and peer review rankings by discipline at the institutional level, taking into account all papers submitted to REF.

The above studies and the study covered in chapter 3 focus on the outputs submitted to the REF in 2014. A comprehensive study by Ravenscroft et al. [2017] examined the correlations between REF impact studies and a range of metrics including citation based metrics using data from CiteSeerX, arXiv and PubMed Central. They additionally tested correlations with altmetric scores obtained using the Altmetric API. Two experiments with REF impact scores and citation data were conducted. The first examined the correlation between the average h-index of authors whose work was cited in the impact case study and the REF impact score. The authors found no clear correlation (Pearson’s $r=-0.005$). The second experiment treated each case study as a 'single author' and calculated the average h-index for all papers cited in a particular study and used this for comparison. Again the results showed little correlation with REF impact data (Pearson’s $r=0.141$). In
their conclusions, the authors note; “We believe that REF impact is more of a proxy for comprehensive impact whereas citation-based metrics and altmetrics mostly measure academic impact.”

This is an interesting point to note as the studies by Mingers et al. [2017], Traag and Waltman [2019], and the study reported in Chapter 3 later showed clear correlations between citation data and REF GPA scores based on academic outputs.

2.11 DORA, Leiden and Responsible Research Metrics

There have been a number of efforts in recent years to move away from purely citation-based metrics. The Declaration on Research Assessment (DORA) was developed in 2012 during the Annual Meeting of the American Society for Cell Biology in San Francisco. It is now signed by 1,962 organisations and over 16,000 individuals and is explicit in its calls for institutions to abandon the practice of correlating a scientist’s contributions with bibliometric indicators when assessing research output. Three key themes encompass the
approach outlined by the declaration;

- The need to eliminate the use of journal-based metrics, such as Journal Impact Factors, in funding, appointment, and promotion considerations.

- The need to assess research on its own merits rather than on the basis of the journal in which the research is published.

- The need to capitalize on the opportunities provided by online publication (such as relaxing unnecessary limits on the number of words, figures, and references in articles, and exploring new indicators of significance and impact).

In 2015, the Leiden Manifesto for research metrics (Hicks et al. [2015]) stated; “Research evaluation has become routine and often relies on metrics. But it is increasingly driven by data and not by expert judgement. As a result, the procedures that were designed to increase the quality of research are now threatening to damage the scientific system.” The Leiden Manifesto includes ten principles for the responsible use of bibliometrics that can be applied across all disciplines and settings.
Both DORA and the Leiden manifest are examples of the push to establish *responsible research metrics*. In 2015 The Metric Tide Report ([Wilsdon 2016](#)) recommended that the UK research community establish a UK Forum for Responsible Research Metrics (FFRRM). The independent sector group was established in September 2016, chaired by Professor Max Lu, Vice-Chancellor at The University of Surrey.

However, frameworks such as these are the beginning of the process. Senior management at HEIs must engage with these processes if they are to be implemented and are to succeed.

The UK FFRRM Guidance on the use of metrics states; "Where metrics are used as part of assessment processes, clear guidance needs to be available which outlines: how metrics will be considered by peer reviewers and/or promotion panels; which metrics will be used for what purpose; how applicants and/or researchers should present metrics (or not) when applying for new positions or funding; and how reviewers should use/interpret any metrics presented."

For wide-scale adoption of responsible research metrics, a wide-
range of stakeholders need to work together including funders, researchers and HEI management, guided by an established framework such as DORA or The Leiden Manifesto.

2.12 Defining and identifying reasons for citation

In line with efforts such as those covered in the previous section to move beyond using citations as a singular data point there have been many efforts to define and capture the differing nature or type of citations. One of the critical components for the wide-scale adoption of schema for citation typing is the availability of a standardised set of classes or a usable ontology for classifying citations according to type or purpose. The work of Garfield [1963], Spiegel-Rosing [1977], and others in identifying and defining different reasons for citation, has been followed by efforts to standardise the class descriptors.

In 2013, Shotton [2010] introduced CiTo, an ontology for describing the type and nature of citations. They stated that the
first purpose of CiTo is: “to enable the citations within a citing work to be recorded and published in machine-readable form” (Page 3). The second purpose is to enable the characterisation of citations, that is, classification according to a range of types. CiTo uses 23 classes divided into factual, for example ‘isCitedBy’, and rhetorical, for example ‘confirms’ or ‘refutes’.

Other more recent work has also begun to combine the production of a class schema for citations with an automatic, machine-learning-based model for the automatic classification of citations according to types defined in the schema. These studies have consisted of two parts: the initial collation phase, to gather a dataset of annotated citations, followed by the classification phase. In 2006, Teufel et al. [2006b] introduced the then-largest dataset of 2,829 annotated citations. These citations were drawn from computer science papers from the ACL anthology reference corpus [Bird et al., 2008]. Teufel et al.’s citation classification schema used 12 distinct types for annotation. 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. [2016] intro-
duced a new dataset of 1,969 annotated citations, and simplified the 12 types first suggested by Teufel et al. [2006b] into 6 types. Most recently, Cohan et al. [2019] took the classification schema first defined by Jurgens et al. [2016] and reduced their 6 classes to just 3. The breakdown of these classification schema is presented in Table 2.3. There is still large variance in both the number and precise types of classes used in these studies. This adds an additional layer of complexity when trying to compare and benchmark results.

Crucially, studies investigating reasons for citation also indicate that a certain number of citations may well refute or disagree with the work being cited. A clear-cut example of this, one of the most highly-cited works in its domain with over 3,500 citations, is the infamous and now retracted 1998 paper by Dr. Andrew Wakefield that linked Autism to the Measles, Mumps, and Rubella (MMR) vaccination Wakefield [1998]. This is despite the paper having been retracted and the author being charged with dishonesty and struck off the medical register in the UK. One cannot therefore imagine that many authors are citing this paper in a positive
<table>
<thead>
<tr>
<th>Teufel et al. 2006</th>
<th>Weak</th>
<th>Weakness of cited approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoCoGM</td>
<td>Contrast/Comparison in Goals or Methods (neutral)</td>
<td></td>
</tr>
<tr>
<td>CoCo-</td>
<td>Author’s work is stated to be superior to cited work</td>
<td></td>
</tr>
<tr>
<td>CoCoR0</td>
<td>Contrast/Comparison in Results (neutral)</td>
<td></td>
</tr>
<tr>
<td>CoCoXY</td>
<td>Contrast between 2 cited methods</td>
<td></td>
</tr>
<tr>
<td>PBas</td>
<td>Author uses cited work as basis or starting point</td>
<td></td>
</tr>
<tr>
<td>PUse</td>
<td>Author uses tools/algorithms/data/definitions</td>
<td></td>
</tr>
<tr>
<td>PModi</td>
<td>Author adapts or tools/algorithms/data</td>
<td></td>
</tr>
<tr>
<td>PMot</td>
<td>This citation is positive about approach used or problem addressed</td>
<td></td>
</tr>
<tr>
<td>PSim PSup</td>
<td>Author’s work and cited work are similar</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Author’s work and cited work are compatible/provide support for each other</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jurgens et al. 2018</th>
<th>Background</th>
<th>The cited paper provides relevant background information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Compare/Contrast</td>
<td>The citing paper compares results or approach with cited paper</td>
</tr>
<tr>
<td></td>
<td>Extension</td>
<td>The citing paper is a direct extension of the cited paper</td>
</tr>
<tr>
<td></td>
<td>Future</td>
<td>The cited paper is a direction for future work from the citing paper</td>
</tr>
<tr>
<td></td>
<td>Motivation</td>
<td>The citing paper is motivated by the cited paper</td>
</tr>
<tr>
<td></td>
<td>Uses</td>
<td>The citing paper uses the tools/methods of the cited paper</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cohan et al. 2019</th>
<th>Background</th>
<th>The citation states, mentions, or points to the background information.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method</td>
<td>Making use of a method, tool, approach or dataset</td>
</tr>
<tr>
<td></td>
<td>Results Compare</td>
<td>Comparison of the paper’s results/findings with the results/findings of other work</td>
</tr>
</tbody>
</table>

Table 2.3: Classification schema used in previous studies
manner. This is an extreme example, but it is used to illustrate that not all citations should be viewed in the same light.

2.13 Text and structure extraction with natural language processing

The growing availability of Open Access academic literature which enables the analysis of the full text content of a paper or article has created the opportunity for developments of new methods and techniques which can leverage this content. If one wishes to automatically assess the semantic and syntactic content of a document it must first be analysed. According to the Oxford English Dictionary, Natural Language Processing is the *application of computational techniques to the analysis and synthesis of natural language and speech*. The key steps for the processing of a full text document using Natural Language Processing (NLP) were summarised by Nelson [2017] as follows:

[i] **Structure extraction**

Identifying fields and blocks of content.
[ii] Identify and mark sentence, phrase, and paragraph boundaries

These markers are important when doing entity extraction and NLP since they serve as useful breaks within which analysis occurs.

[iii] Tokenisation

Divides up character streams into tokens which can be used for further processing and understanding. Tokens can be words, numbers, identifiers or punctuation.

(iv) Lemmatisation / stemming

This reduces word variations to simpler (root) forms that may help increase the coverage of NLP utilities.

This simplification of a document into a structured representation allows for further techniques to be applied. The transformation of a document into a vector space model allows for mathematical and algebraic calculations and comparisons between documents. A common methodology applied to these structured documents is Term frequency - Inverse document frequency (tf-idf) analysis suggested by Sparck Jones [1972] which can identify
which are the significant and important terms in a document. This type of processing can then be used for undertaking comparisons of documents in terms of similarity.

The deconstruction and analysis of documents using the above methodology allows for summarisation and the classification of documents and also for classification of individual parts of a document. Several frameworks for text classification using NLP models have become available in recent years including Stanford’s Core NLP Suite (Manning et al. [2014]) and the Natural Language Toolkit (NLTK) for the Python programming language (Bird et al. [2009]). These toolkits support a range of supervised machine learning models for text classification such as Linear Regression, Support Vector Machines (Cortes and Vapnik [1995]) and Random Forest classifiers (Breiman [2001]). These toolkits simplify many of the processes required for NLP analysis of large text corpora and in some cases simplify the replication of earlier studies in the domain. This is of key importance as the availability of both code and data can be vital for reconstructing experiments and for developing works created by earlier efforts.
In 2014, Knoth and Herrmannova introduced the concept of *Semantometrics* which states that the full text of a research document is required to fully understand its contribution. They further introduce the idea of a research contribution measure which, rather than semantic similarity, uses the semantic distance between documents to assess a paper’s contribution Knoth and Herrmannova [2014].

The methods discussed above fall into the category of *similarity models*. There are other methods that also utilise the full text of a document in order to understand it, notably *probabilistic models*; Blei et al. 2003, introduced Latent Dirichlet Allocation (LDA), a probabilistic model for categorising datasets such as text corpora. They demonstrate applications for this model in the areas of in topic modeling, text classification and collaborative filtering.

Distributed vectors, also called *word embeddings*, are based on the distributional hypothesis that states that words appearing within similar context are likely to possess similar meaning. In 2013 Mikolov et al. [2013b] introduced word2vec which uses a shallow neural network to produce word embeddings that retain
the semantic and syntactic information of the original words. This enriched information is then used in the training of models for a range of NLP tasks.

2.14 Deep learning models for natural language processing

The majority of methods used to study NLP problems employed time-consuming, hand-crafted features. More recently, there has been significant interest in the increasing potential of deep learning models for NLP applications including text extraction and summarisation, machine translation and question answering. In 2011, Collobert et al. first applied a Convolutional Neural Network (CNN) to NLP tasks such as part-of-speech tagging, chunking, named entity recognition, and semantic role labeling and showed a semi-supervised model could achieve SoTA for these tasks. Another deep-learning model BERT (Bidirectional Encoder Representations from Transformers) was introduced by Devlin et al. in 2018. BERT offers improvement on previous models by applying
bi-directional training to the language model. That is, the text sequence is parsed both left to right and right to left. The results of the original study demonstrates that this offers improved performance over earlier models that only consider textual information in one direction. An update to BERT was released in 2019. Trained specifically on a corpora of scientific texts, SciBERT Beltagy et al. [2019] is designed for NLP tasks specifically related to scientific texts. The SciBERT model was used by Cohan et al. [2019] as the basis for their classification experiments and is discussed in Chapter 5.

2.15 Citation analysis and classification

The inherent problems in the application of simple citation counts are widely known and have been extensively discussed in the previous section. Citation analysis, which may offer the opportunity for new bibliometrics based on the semantic content of citations themselves, can take several different forms, Gipp and Beel [2009] introduced Citation Proximity Analysis (CPA) in 2009 which posits
that the closer together two citations appear in a document, the
more likely they are related. This technique allowed the develop-
ment of improved paper / article recommender systems based on
this observation. In 2013 Bertin et al. [2013] analysed the distri-
bution of citations in a research paper, noting that most papers
follow the standard IMRAD structure (Introduction, Methods,
Results and Discussion). There is therefore an opportunity to
evaluate whether the position of the citation within the document
can provide additional information.

Automatic classification according to citation type was first in-
troduced by Teufel et al. [2006a]. The authors typed citations into
one of twelve categories; Explicit statement of weakness, a con-
trast or comparison with other work (4 categories), showing agree-
ment/usage/compatibility with other work (6 categories), and a
neutral category. 26 papers containing 548 citations were classi-
fied according to type.

The general workflow used for classifying citations is shown in
Figure 2.5. The extraction of the citations is performed using Nat-
ural Language Processing (NLP) techniques. The exact method
employed and the processing steps required are variant depending on the format of the original document. The study by Bertin et al. [2013] used papers taken from PlosOne which are available in XML format, removing the need for any textual preprocessing. Documents that are only available in PDF format can present considerable difficulties in terms of text mining the full content of the document. These difficulties are discussed in chapter 4.

Valenzuela et al. (2015) Valenzuela et al. [2015] introduced classification of citations according to influence on the citing pa-
per. This is an extremely interesting approach and demonstrated that only approximately 14-15% of citations are truly influential to the citing paper, the rest are incidental or perfunctory. Citation based metrics that consider the type or sentiment of the citation present a new research opportunity. Recently Pride and Knoth [2017b] attempted to replicate the studies of Valenzuela et al. [2015] and Zhu et al. [2015] and found many of the selected features offered little in the way of predictive ability or were irreproducible due to the lack of information provided by the original authors. This study is covered in full in the chapter 4.

The study by Teufel et al. [2006b] used an instance-based learning algorithm (IBk, Aha and Kibler [1991]) to classify citations according to type. Jurgens et al. employed an SVM with a random forest classifier for their experiments using the citation classification labels shown in Table 2.3. In 2019 Cohan et al. applied a deep learning model (SciBeRT, Beltagy et al. [2019]) to the citation classification problem. Their study also employed word embeddings as discussed in the previous section. Their use of a bi-directional LSTM (long-short-term memory) model combined
with word embeddings offered improvement in F1 scores using the same classification labels used by Jurgens et al. [2016]. There are, however, limitations in these prior studies, most notably the lack of sizable datasets for training machine learning models. Annotating citations is a labour-intensive and therefore expensive task; this has limited the size of the datasets previously compiled. Further, these datasets are drawn only from one or two domains; none of the datasets is truly multi-disciplinary.

2.16 Conclusions

The initial over-arching review of current metrics demonstrated that, despite their wide-spread adoption, there are inherent limitations and caution must be exercised when these metrics are used as a shortcut for research quality. Citations have been demonstrated to show the utility of research, but not the quality. There are also biases in citations, one example being the Matthew Effect as discussed. Furthermore, differing citation rates across different domains also make direct comparison more complex. Crucially,
this review highlights that no current citation-based metric considers the semantic aspects of a citation, i.e. the language contained within the citing paper itself.

Many of the studies reviewed here have defined different reasons for citation and others have then developed these initial ideas into classification schema for citations. There has however been little agreement on both the number of classes required, and the specific descriptors for these classes. Additionally, the datasets produced by studies investigating citation classification have been of limited size and have been drawn from a single domain or two at most. These factors present key challenges in the automatic classification of citations.

If citations can be accurately and automatically classified according to type, with particular regard to citations that either support or refute the cited work, there is then the potential to incorporate this qualitative aspect of citations in new bibliometric measures.

In addition, citation data is already being used globally in PRFS exercises. Providing enhanced citation data that leverages
the semantic aspect of citations may prove more informative than current metrics for peer reviewers in these research evaluation exercises.
Chapter 3

Peer Review or Citation Data?
Predicting Institutional Rankings in the UK’s Research Excellence Framework
3.1 Introduction

Addressing one of the two main areas previously identified as deserving further research, this chapter describes a study that reviewed the current use and impact of citation data in a Performance-based Research Funding System (PRFS), using data from the UK’s Research Excellence Framework (REF). This chapter is based on a paper published as part of the proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL2018) where it won the Best Paper Award [Pride and Knoth, 2018].

The study focused on a specific national PRFS for which submission and results data were available, in order to address Research Question 1:

**RQ1. To what extent do peer review and current bibliometric indicators correlate at the institutional / discipline level?**

The experiments focused on the academic outputs (research papers) component of the REF, as the metadata for all outputs
for all institutions were publicly available for download from the REF website.

3.1.1 Peer review and citation data in the U.K.’s Research Evaluation Framework (REF)

The 2014 UK REF exercise was the largest overall assessment of university research output ever undertaken globally [Cressey and Gibney, 2014]. In total, 154 UK Higher Education Institutions made a 3-part submission to the REF, including: the outputs - the research products, including publications, patents, physical and digital artefacts, and exhibitions and performances; Impact cases that described the social and economic impact of the research that institution undertook; and the research environment, including the places and the practices of research. The REF process convened a panel of expert peer reviewers for each unit of assessment (UoA), whose role was to conduct peer review, to international standards, for each submission. In-depth detail regarding this process is given in the discussion section of this chapter.

REF 2014 graded 191,000 research outputs from 154 institu-
<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4*</td>
<td>Quality that is world-leading in terms of originality, significance and rigour.</td>
</tr>
<tr>
<td>3*</td>
<td>Quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence.</td>
</tr>
<tr>
<td>2*</td>
<td>Quality that is recognised internationally in terms of originality, significance and rigour.</td>
</tr>
<tr>
<td>1*</td>
<td>Quality that is recognised nationally in terms of originality, significance and rigour.</td>
</tr>
<tr>
<td>U</td>
<td>Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment.</td>
</tr>
</tbody>
</table>

Table 3.1: Classification of outputs submitted to REF2014

tions as U (unclassified) or from one- to four-stars, as shown in Table 3.1. The grading for each output was determined on the basis of originality, significance, and rigour.

Institutions were then ranked in each discipline according to the mean score of all their outputs. This was presented as a Grade Point Average (GPA) score. Note that the Universities were not provided feedback as to the ranking for each individual output. They were however told what percentage of their submissions were
in each of the categories (U-4*) for each discipline.

The full REF results league table shows the dominance of the traditional centres of research: Imperial College London, Cambridge University, Oxford University, and The London School of Economics were four of the top five institutions; the overall first-ranked institution was The Institute for Cancer Research, which was the only non-university institution to take part in the exercise [Times Higher Education, 2014a].

Each institution’s results are given as three separate metrics, shown in Table 3.2. The GPA (Grade Point Average) is the mean ranking for all outputs from one- to four-star. The two further metrics are designed to normalise for the number of academics who were submitted to the REF (Power) and the percentage of academics submitted as a percentage of total staff (Intensity).

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>Cumulative REF quality score (1-4*)</td>
</tr>
<tr>
<td>Power</td>
<td>GPA x number of academics submitted</td>
</tr>
<tr>
<td>Intensity</td>
<td>GPA x % of academics submitted</td>
</tr>
</tbody>
</table>

Table 3.2: REF result metrics

The rankings for the institutions change markedly dependent
on which of the 3 metrics listed above are used. The league tables produced by *Times Higher Education* showed all three results, but the final ranking was determined solely by the institutions’ GPA. This led to some interesting anomalies in the final rankings, with evidence that some institutions had clearly ‘gamed’ the REF process in terms of their selection of total number of academics submitted and the selection of areas of submissions [Times Higher Education, 2014b]. Perhaps the most striking example is Cardiff University, which had a very selective submission. This resulted in Cardiff coming 6th out of all UK universities based on GPA alone. However, it was 18th for power, and 50th for intensity [Mingers et al., 2017].

### 3.1.2 Use of citation data in the REF peer review process

For assessing the research outputs, each of the REF peer review panels chose whether or not to use citation data to inform its decisions. 11 of 36 selected panels chose to do so and were provided with citation data from Elsevier Scopus to assist their decision
making. Panels were given the number of citations required to put the paper in the top 1%, 5%, 10%, or 25% of papers for each domain and each year of publication. This gave REF reviewers a subject-level benchmark against which to consider the citation data [HEFCE, 2014a].

Additionally, each journal in the Scopus database is assigned to one or more subject classifications, using Scopus’s ‘All Science Journal Classification’ (ASJC) codes. Panels were told the mean number of times that journal articles and conference proceedings were published worldwide in the years covered by the REF process (2008-2014), for each ASJC code.

The REF organisers noted that:

“Citation counts depend partly on the field of research and a publication’s age. Therefore, where sub-panels are making use of citation data the REF team will provide contextual data to assist in the sub-panel’s interpretation of the citation counts. This will consist of information about the citation behaviour of groups of papers
worldwide, published in a similar subject area and of a similar age.” [HEFCE, 2014b]

3.1.3 Destruction of individual rankings for outputs

Once all submissions for a Unit of Assessment (UoA) had been graded by the peer review panel, the percentages of unclassified and 0* to 4* submissions for each institution were recorded. A Grade Point Average was assigned to each Higher Education Institution (HEI) for each discipline. This was calculated as the mean ranking for each institution for each UoA and was the final figure used to calculate the overall ranking for an institution’s submissions.

Whereas the aggregate GPA ranking for all UoAs and all institutions was publicly available, it is now not possible to obtain a direct comparison between citation data and the individual rankings for each research output, as HEFCE states that these data were destroyed. The rationale behind this was to preempt any requests for this data under the Freedom of Information Act.
3.2 Dataset

For this study, the first complete dataset of outputs submitted to the REF, along with their corresponding bibliometric data and aggregate GPA rankings, was compiled.

The dataset creation procedure is depicted in Figure 3.1. First, the REF 2014 submission list HEFCE [2014a] was downloaded.
For each output, the list contains; publication title, publication year, publication venue, name of institution, and UoA. These fields were fully populated for 190,628 out of 190,963 submissions to the ‘outputs’ category of the REF process.

In order to test the hypothesis with respect to REF rankings and citation-based metrics at the discipline level, it was necessary to first collect historic and contemporary citation data for all papers contained within the REF dataset.

The Microsoft Academic Graph (MAG) was used to enrich the REF outputs list with citation information. At the time of the experiment, MAG contained approximately 168 million individual papers and 1.15 billion citation pairs. The decision to use MAG instead of Scopus was motivated by a constraint in the Scopus API service: although Scopus (operated by Elsevier) was used to provide citation data to the REF process, the free version of the Scopus API service is limited to 20,000 requests per week. It would therefore have taken almost 2 months to gather the required data, more than 10 times slower than using MAG. Additionally, studies by Herrmannova and Knoth [2016] and Hug and Brändle
have recently confirmed how comprehensive the MAG citation data are. Google Scholar could not be used, as it does not offer an API and prohibits ‘scraping’ of data.

The MAG Evaluate API was queried systematically for each submitted output using a normalised version of the publication’s title (lower case, diacritics removed). This returned a set of MAG IDs which were potential matches to the article. The MAG Graph Search API was subsequently queried to validate each of the potential matches. The most similar publication title that had at least 0.95 cosine similarity was accepted as a match. This threshold was set by manually observing about one hundred matches. This process successfully matched 145,415 REF research outputs with 6.95 million citations, corresponding to a recall of 76% of the total initial REF submitted-outputs list.

Table 3.3 is ordered by the mean-citations-per-paper (MCPP) and shows total number of outputs, percentage of these outputs available in MAG, and the total citations of these outputs.

In addition, as described in Figure 3.1, the Assessment Data were downloaded from the REF 2014 website. These data con-
Table 3.3: UoAs with the highest mean-citations-per-paper.

<table>
<thead>
<tr>
<th>UoA / Subject</th>
<th>Outputs</th>
<th>% in MAG</th>
<th>Citations</th>
<th>MCPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Health</td>
<td>4,881</td>
<td>94.61%</td>
<td>505,950</td>
<td>109.56</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>13,394</td>
<td>90.78%</td>
<td>1,278,810</td>
<td>105.17</td>
</tr>
<tr>
<td>Physics</td>
<td>6,446</td>
<td>84.51%</td>
<td>491,151</td>
<td>90.15</td>
</tr>
<tr>
<td>Biological Sciences</td>
<td>8,608</td>
<td>92.20%</td>
<td>620,009</td>
<td>78.12</td>
</tr>
<tr>
<td>Earth Systems / Environment</td>
<td>5,249</td>
<td>91.64%</td>
<td>315,429</td>
<td>65.58</td>
</tr>
<tr>
<td>Chemistry</td>
<td>4,698</td>
<td>87.71%</td>
<td>246,361</td>
<td>59.78</td>
</tr>
<tr>
<td>Allied Health Professions</td>
<td>10,358</td>
<td>89.35%</td>
<td>402,033</td>
<td>43.43</td>
</tr>
<tr>
<td>Ag. Vet. and Food Science</td>
<td>3,919</td>
<td>90.76%</td>
<td>150,959</td>
<td>42.44</td>
</tr>
<tr>
<td>Comp. Science and Informatics</td>
<td>7,645</td>
<td>89.22%</td>
<td>284,815</td>
<td>41.76</td>
</tr>
<tr>
<td>Economics and Econometrics</td>
<td>2,600</td>
<td>88.81%</td>
<td>95,591</td>
<td>41.4</td>
</tr>
</tbody>
</table>

Table 3.4: Dataset statistics

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Units of Assessment (UoAs)</td>
<td>36</td>
</tr>
<tr>
<td>Number of institutions</td>
<td>154</td>
</tr>
<tr>
<td>Number of UoAs/institution pairs</td>
<td>1,911</td>
</tr>
<tr>
<td>Number of submissions (papers)</td>
<td>190,628</td>
</tr>
<tr>
<td>Number of submissions (papers) in MAG</td>
<td>145,415</td>
</tr>
<tr>
<td>Number of citations</td>
<td>6,959,629</td>
</tr>
</tbody>
</table>
The full dataset used in these experiments and all results can be downloaded from Figshare\textsuperscript{1} [Pride, 2018].

3.3 How well do peer review judgments correlate with citation data at the institutional level?

Once the full dataset was assembled, we wanted to test a range of citation based metrics and the strength of any correlation with REF GPA scores. We compiled the mean, median and total citations for all papers from all institutes for each UoA. This allowed for direct comparison between each institute’s GPA score and their citation data for each UoA. These data were then used to test the correlation between citation data and REF GPA rankings for all outputs for every institution in every UoA. The remainder of this study focuses on the following metrics; mean citations in December 2017 ($mn_{2017}$), median citations in December 2017 ($med_{2017}$), mean citations at the time of the REF exercise ($mn_{2014}$), and

\footnote{https://figshare.com/s/69199811238dcb4ca987}
median citations at the same point \( (med_{2014}) \) as these were the most strongly correlated results. Pearson’s correlation coefficient was used for all of these tests. The ten highest- and ten lowest-measured correlations by UoA are shown in Table 3.5. The citation data \( (cd) \) column denotes whether the REF expert panels considered citation data in their deliberations. Attempts to run correlations with other similar aggregate functions are not shown in this table, as they had far lower correlations with REF GPA scores.

Strong positive correlations can be observed at the discipline level for a large proportion of the UoAs, particularly for median citation count in 2017. Whilst the correlation was most often stronger for those UoAs that had used citation data in the REF peer review process, this was not always the case. Aeronautical and Mechanical Engineering and Social Work and Policy are two disciplines which did not use citation data, yet show very strong correlations with GPA results. At the lower end of the scale, there was little correlation between GPA ranking and citation data, notably for those subjects covered by REF panels C
Table 3.5: Pearson Correlation Coefficient between REF GPA output rankings and citation data. For all results $p < 0.001$

and D [HEFCE, 2015]. Lack of coverage in many of these areas is, however, understandable, as these are disciplines that do not always produce journal articles, conference proceedings, and other digitally-published and highly-citable artifacts as their main type of output. There is, however, clear delineation between the more-highly correlated UoAs ($r > 0.353, n = 37, p < 0.001$) and
those more-weakly correlated \((r \leq 0.159, n = 37, p < 0.001)\). The low correlation for mean citations for Biological Sciences is explained by a single paper, which was the most-highly-cited paper in the UoA. This paper received 4,626 citations, 58\% more than next cited paper and 9 times as many citations as all other submissions for that institution combined. Further, this paper came from second-lowest-ranked institution (by GPA) of 44 institutions. Had this paper been discounted from the correlations, the prediction results would have been far more clearly aligned with the other UoAs \((mn2017=0.782, mn2014=0.766)\).

The variance of citation data coverage across UoAs led us to explore whether there could be a relationship between the strength of the correlations GPA, and citation data correlation with the coverage of citation in a given UoA. Figure 3.2 plots this for both the UoAs that used citation data, and those that did not.

Figure 3.2 plots the median citations for each UoA in MAG against the correlation between median citations and GPA ranking for each UoA. An ‘o’ represents a UoA that did not use citation data, whilst an ‘x’ denotes a UoA that did. It can be seen that
the highly cited domains in MAG are overall the highly correlated domains.

Figure 3.2: Plot of REF GPA / citation data correlation against MAG coverage for each UoA

While the graph confirms that the highly-cited UoAs in MAG are those that used citation data, it indicates that a few UoAs that did not use citation data also exhibit strong correlations. Unsurprisingly, the plot suggests that there might be a small bias exhibited by extra correlation strength in UoAs that used citation data. However, given the small number of UoAs, this is not
3.4 How well can citation data predict peer review based institutional rankings?

Table 3.6 shows the top five institutions for Chemistry, Biological Sciences, and Aeronautical and Mechanical Engineering as ranked in the REF by GPA, and predictions of ranking using $med_{2017}$ and $med_{2014}$ respectively. $mc_{2017}$ and $mc_{2014}$ show the median citation count for that institution. $R_{diff}$ shows the rank difference when ranked by a particular citation metric. The prediction performance indicated in these tables is not unique; in four of the five top UoAs by correlation strength, the highest-ranked institution is predicted correctly by both $med_{2014}$ and $med_{2017}$.

Table 3.7 demonstrates the effectiveness of predictions based on $med_{2014}$ for the 10 most-highly-cited UoAs. To compare the prediction error, expressed by $rdiff$, across UoAs, the mean rank difference was calculated, normalised by number of institutions ($nrdiff$). Mean Average Precision (MAP) was used to express
Table 3.6: Rankings by GPA and predictions produced using \(med_{2017}\) and \(med_{2014}\) respectively for the three most highly correlated UoAs overall prediction accuracy. The HEI column denotes the number of institutions submitting to that UoA. The parameter \(rt\) denotes the prediction rank tolerance. For example, \(rt = 3\) indicates that a prediction within 3 positions of the original assessment result will be considered as correct. Given the simplicity of the prediction method, this is a strong indication of the power of citation data in this task. One could reasonably expect that further improve-
ments can be made by employing more sophisticated indicators. However, as the predictions are not as good for UoAs that have lower-than-average mean-citations-per-paper, the use of citation data unaccompanied by peer review assessments in those UoAs is not recommended.

<table>
<thead>
<tr>
<th>UoA</th>
<th>HEIs</th>
<th>rdiff</th>
<th>nrdiff</th>
<th>MAP rt=3</th>
<th>MAP rt=5</th>
<th>MAP rt=10</th>
<th>MAP rt=10%</th>
<th>MAP rt=20%</th>
<th>MAP rt=30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Sci.</td>
<td>89</td>
<td>12.39</td>
<td>0.139</td>
<td>0.19</td>
<td>0.32</td>
<td>0.50</td>
<td>0.46</td>
<td>0.75</td>
<td>0.87</td>
</tr>
<tr>
<td>Ag. Vet.</td>
<td>29</td>
<td>4.02</td>
<td>0.139</td>
<td>0.45</td>
<td>0.65</td>
<td>0.86</td>
<td>0.45</td>
<td>0.68</td>
<td>0.86</td>
</tr>
<tr>
<td>Clin. Med.</td>
<td>31</td>
<td>4.38</td>
<td>0.141</td>
<td>0.51</td>
<td>0.70</td>
<td>0.93</td>
<td>0.51</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>Allied H.</td>
<td>83</td>
<td>12.03</td>
<td>0.145</td>
<td>0.20</td>
<td>0.30</td>
<td>0.55</td>
<td>0.43</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td>Economics</td>
<td>28</td>
<td>4.07</td>
<td>0.145</td>
<td>0.57</td>
<td>0.71</td>
<td>0.92</td>
<td>0.57</td>
<td>0.78</td>
<td>0.92</td>
</tr>
<tr>
<td>Chemistry</td>
<td>37</td>
<td>5.51</td>
<td>0.149</td>
<td>0.54</td>
<td>0.56</td>
<td>0.83</td>
<td>0.54</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td>Earth Sys.</td>
<td>45</td>
<td>7.24</td>
<td>0.161</td>
<td>0.40</td>
<td>0.51</td>
<td>0.77</td>
<td>0.51</td>
<td>0.68</td>
<td>0.84</td>
</tr>
<tr>
<td>Pub. Health</td>
<td>32</td>
<td>5.18</td>
<td>0.162</td>
<td>0.50</td>
<td>0.62</td>
<td>0.84</td>
<td>0.50</td>
<td>0.68</td>
<td>0.84</td>
</tr>
<tr>
<td>Bio. Science</td>
<td>44</td>
<td>7.59</td>
<td>0.173</td>
<td>0.34</td>
<td>0.52</td>
<td>0.72</td>
<td>0.52</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>Physics</td>
<td>41</td>
<td>7.36</td>
<td>0.180</td>
<td>0.36</td>
<td>0.53</td>
<td>0.78</td>
<td>0.43</td>
<td>0.73</td>
<td>0.80</td>
</tr>
<tr>
<td>All (mean)</td>
<td>45</td>
<td>6.98</td>
<td>0.153</td>
<td>0.41</td>
<td>0.54</td>
<td>0.77</td>
<td>0.49</td>
<td>0.72</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 3.7: Rank prediction quality for the top 10 UoAs with the highest mean-citations-per-paper
Table 3.8: Comparison of the prediction performance of this study with Mryglod et al. [2015]

The prediction performance of this study was compared to that of Mryglod et al. Mryglod et al. [2015]. In order to conduct a fair and exact comparison, it was necessary to remove a number of institutions from our input data. Mryglod et al. reported that they were unable to obtain citation indicators for all institutions in a given UoA; their study covered three of the top-ten highly-cited UoAs. Table 3.8 shows that the predictions in this study are significantly better than Mryglod et al.’s across all categories.
3.5 Discussion

This study, and those by Mingers et al. [2017], Anderson et al. [2013], and Smith [2008], demonstrate that citation counts can offer a surprisingly high degree of accuracy when used at the institutional or departmental level. The work reported in this chapter has been conducted on a significantly larger dataset, and its prediction accuracy is higher than shown in previous studies, despite deliberately using fairly simplistic indicators. Several studies, including The Metric Tide [HEFCE, 2015], The Stern Report [Stern, 2016], and the HEFCE pilot study [HEFCE, 2009], all state that metrics should be used as an additional component in research evaluation, with peer review remaining as the central pillar. Yet peer review has been shown by Hojat et al. [2003], Lee et al. [2013], and Smith [2006] amongst others to exhibit many forms of bias, including institutional bias, gender- / age-related bias, and bias against interdisciplinary research. In an examination of publication bias, Emerson et al. [2010] noted that reviewers were much more likely to recommend papers demonstrating positive results
over those that demonstrated null or negative results.

All of the above biases exist even when peer review is carried out to the highest international standards. There were close to 1,000 peer review experts recruited by the REF; however the sheer volume of outputs requiring review calls into question the exactitude of the whole process. For example, the REF panel for UoA 9, Physics, consisted of 20 members. The total number of outputs submitted for this UoA was 6,446. Each paper is required to be read by 2 referees, doubling the reading requirement to 12,892 instances. Therefore, each panel member was required to review, to international standards, an average of 644 papers in a little over 10 months. If every panel member worked every day for 10 months, each member would need to read and review 2.14 papers per day to complete the work on time. This is, of course, in addition to the panelist’s usual full-time work load. Moreover, Physics is not an unusual example, and many other UoAs tell a similar story in terms of the average number of papers each panel member was expected to review: Business and Management Studies (1,017 papers), General Engineering (868 papers), Clinical Medicine (765
papers). The burden placed on the expert reviewers during the REF process was onerous in the extreme. Coles [2013] calculated a very similar figure of 2 papers per day, based on an estimate before the data we now have was available. “It is blindingly obvious,” he concluded, “that whatever the panels do will not be a thorough peer review of each paper, equivalent to refereeing it for publication in a journal.” Sayer [2014] was equally disparaging in regard to the volume of papers each reviewer was required to read, and also expressed some doubts about the level of expertise and diversity within the review panels themselves.

In addition to the potential pitfalls in the current methodologies, there is also the enormous cost to be considered. This was estimated to be £66m for the UK’s 2008 Research Assessment Exercise (RAE - the predecessor to the REF). This rose markedly to £246m for the 2014 Research Excellence [Technopolis, 2015]. This is comprised of £232M in costs to the higher education institutions, and around £14M in costs for the four UK higher education funding bodies. The cost to the institutions was approximately £212M for preparing the REF submissions for the three areas (out-
puts, impact, and environment), with the cost for preparing the outputs accounting for the majority share of this amount. Additionally, there were costs of around £19M for panelists’ time. This study does not set out to answer whether peer review or bibliometric indicators are best placed for assessing research in national scale exercises such as the REF but the results show that citation data and peer review results may be more closely aligned at the institutional / discipline level than previously shown.

The Metric Tide report (Wilsdon [2016]) noted however that *There is large variation in the availability of metrics data across the REF submission, with particular issues with coverage in units of assessment (UoAs) in REF Main Panel D.* This panel covers many disciplines in the Arts and Humanities. Outputs from these domains are not necessarily ‘citable articles’. Books, stage productions, music, TV and film productions and art installations all fall into this category. Although impactful in their own right, none of these examples can be assessed accurately or fairly with standard bibliometrics and our results show, unsurprisingly, that there is little correlation between peer review decisions and citation data.
Table 3.9: REF UoA Sub-panels using citation data in REF2021

<table>
<thead>
<tr>
<th>Sub-panel</th>
<th>Subject Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-panel 1</td>
<td>Clinical Medicine</td>
</tr>
<tr>
<td>Sub-panel 2</td>
<td>Public Health, Health Services and Primary Care</td>
</tr>
<tr>
<td>Sub-panel 3</td>
<td>Allied Health Professions, Dentistry, Nursing and Pharmacy</td>
</tr>
<tr>
<td>Sub-panel 4</td>
<td>Psychology, Psychiatry and Neuroscience</td>
</tr>
<tr>
<td>Sub-panel 5</td>
<td>Biological Sciences</td>
</tr>
<tr>
<td>Sub-panel 6</td>
<td>Agriculture, Food and Veterinary Sciences</td>
</tr>
<tr>
<td>Sub-panel 7</td>
<td>Earth Systems and Environmental Sciences</td>
</tr>
<tr>
<td>Sub-panel 8</td>
<td>Chemistry</td>
</tr>
<tr>
<td>Sub-panel 9</td>
<td>Physics</td>
</tr>
<tr>
<td>Sub-panel 11</td>
<td>Computer Science and Informatics</td>
</tr>
<tr>
<td>Sub-panel 16</td>
<td>Economics and Econometrics</td>
</tr>
</tbody>
</table>

Looking forward, for the current REF2021 exercise there is little substantive change to the peer review process and Research England are continuing the usage of bibliometric data. As can be seen from Table 3.9, the exact same 11 panels will again use citation data to inform their peer review decisions.

3.6 Summary of findings and conclusion

This work constitutes the largest quantitative analysis of the relationship between peer reviews (190,628 paper submissions) and citation data (6.9m citation pairs) at an institutional level. The
datasets produced by this study were made available via Figshare Pride [2018] and the results were subsequently confirmed in later studies by both Harzing and Alakangas [2016] and Traag and Waltman [2019].

First, the results show that citation data exhibit strong correlations with peer review judgments, when considered at the institutional level and within a given discipline. Specifically, there was a surprisingly strong correlation for many Units of Assessment (UoAs) between an institution’s REF Grade Point Average (GPA) ranking for its research outputs, and citation data. These correlations tend to be higher in disciplines with high mean-citations-per-paper. The results demonstrate that this makes it possible to predict institutional rankings with a degree of accuracy in highly-cited disciplines using basic bibliometric indicators.

Second, the results demonstrate that citation data can be used to predict top-ranked institutions with a surprisingly high precision. For the ten UoAs with the highest number of mean-citations-per-paper in REF 2014, this study achieved 0.77 MAP with prediction rank tolerance 10 with respect to the REF 2014 results.
In four out of five top UoAs by correlation strength, the highest-ranked institution in the REF results was predicted correctly. It is important to note that these predictions are based on citation data that were available at the time of the REF exercise.

While this analysis does not answer whether using citation-based indicators can predict institutional rankings better than relying on a peer-review system, the results provide evidence that the REF peer-review process led to results highly-similar to those that could have been predicted automatically using citation data. The 11 REF UoAs with the highest mean-citations-per-paper in MAG are the identical UoAs in which the peer review panels used citation data to inform their decisions. If peer-review is conducted in the way it was conducted in the REF, then it may have been more cost effective to carry out the institutional evaluation purely using citation data, particularly in UoAs with high mean-citations-per-paper, hence saving a significant proportion of the £246m spent on organising the review process [Technopolis, 2015].

It can be seen that aggregate peer review results and biblio-
metric indicators are highly-correlated at the institution level in many disciplines. This can be observed most conclusively in cases in which peer review panels had access to citation data. It may therefore be argued that countries whose PRFS have a peer review component should consider carefully how the peer review process is conducted, and whether to allow peer review panels to inform their decisions using citation data.

The results of this study have potential implications for PRFS globally; citation data is already being used in global PRFS, not just in the UK. Furthermore, several countries use PRFS that are based entirely on bibliometric data.

In the next stage of this work we begin to look at how much more information could be gained from citations themselves rather than relying on metrics that are built on simple counts. However, the accurate and automatic identification of either the type of citation, or how influential it was, is not a simple task. The next chapter examines these difficulties in depth and investigates methods for improving results in this domain.
Chapter 4

Effective Features for Identifying Citation Influence
4.1 Introduction

Several previous studies, including Zhu et al. [2015], Zhao et al. [2017], and Valenzuela et al. [2015], argue that citations contain far more information than is revealed by a single aggregate digit, and that not all citations should be treated equally. This is particularly true for citations that dispute or refute earlier works. As will be shown in this chapter, citations contain far more information than is exposed by current bibliometric measures. As discussed in Chapter 1, authors cite prior works for a wide range of reasons, e.g.: using either methodology or data from previous works, citing as background, comparing and contrasting results, or extending earlier works are some examples of very different citation types. In each of these cases, the context in which the citation appears offers a wealth of information in terms of identifying the original authors’ reason for the citation.

Studies by Valenzuela et al. [2015] and Zhu et al. [2015] have most recently begun to apply natural language processing techniques to identify the importance of a citation in the context of
a particular research paper, that is identifying the citations that were highly influential to the paper’s author.

As discussed in Chapters 1 and 2, there are a myriad reasons why an author may cite a previous work. There are, however, some caveats that should be considered. Researchers may cite papers for many non-essential reasons. These somewhat superfluous citations can potentially exaggerate a paper’s true importance. This study focuses on finding features to aid the classification of influential citations, a rigorous definition is therefore required. As the study in this chapter builds primarily on the work of Valenzuela et al. [2015] we utilise their description here. Influential citations are those that give clear indication that the cited work is used or extended in the new study.

Hou et al. [2011] first suggested the idea of using an internal citation count based on the full text of a research paper, rather than just the bibliography, to determine how influential a particular citation was. They demonstrated a positive correlation between the number of times a citation occurs and its overall influence on the citing paper. Zhu et al. [2015] combined these
earlier approaches and suggested a range of 40 classification features, including both semantic and metric features, to determine influence. Most recently, Valenzuela et al. [2015] made significant efforts to construct a reference set which was publicly released and from which this chapter draws. Valenzuela et al. suggested a range of 12 features, many of which are similar to those of Zhu et al.

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. Further, these studies identify three essential feature types: 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, a deeper understanding is needed of which of the previously-suggested features are most effective at this task, as reflected in the second research question:

**RQ2. What are the key features that indicate a citation**
is influential, rather than just incidental?

The growing availability of full texts of publications is 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\(^1\) are allowing researchers to use 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 [Knoth and Herrmannova, 2014], which attest that the full text of a publication is required to ascertain its impact effectively.

This chapter addresses the problem of identifying influential citations based on publications’ full text. It discusses the approaches for detecting influential citation in Section 4.2, providing a critical analysis of features previously applied in this task, and selecting a set of three key features for further analysis. Section 4.3 presents a comparative study of the identified features, together with the challenges inherent in this task. The chapter is based

\(^{1}\)http://www.core.ac.uk

4.2 Methodology

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

1. Extracting the full text of the manuscript.

2. Parsing the full text to detect the document structure, such as the document metadata, references, citation markers, and sections.

3. Extracting the features from the document structure, possibly with an enrichment step for features based on external evidence.

4. Applying a classifier trained using supervised machine learn-
ing approaches.

The rest of this section describes this workflow, concentrating on the selection of features used in the citation-type classification task.

4.2.1 Extracting the full text and parsing

Unless a paper is available in a structured format, such as XML, the original PDF file must be converted 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 [Lawrence et al., 1999]. Some tools for inferring the document structure, such as ParsCit [Councill et al., 2008], require initial conversion to plain text. Others, such as GROBID [Lopez, 2009], operate directly on the PDF file.

4.2.2 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.

Consider the features presented in the two most recent studies. Zhu et al. [2015] used a rich range of features in five categories that moved 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] took a similar approach, producing a list of 12 features (Table 4.1).

4.2.3 Selection of features for experiments and comparison

We analysed the 40 features presented by Zhu et al. [2015] and the 12 features presented by Valenzuela et al. [2015]. Of the 40
features, a combination of just 4 features resulted in the best performance of Zhu et al.’s model. Adding features beyond this actually lowered the performance. Of these 4 features, we could not reliably replicate Zhu et al.’s results for one feature (countsInPaperSecNum). Of Valenzuela et al.’s 12 features (Table 4.1), we found the results for 3 features irreproducible (F3, F5, F12); we were unable to reliably replicate the results for 2 features due to PDF extraction issues (F2, F6), and we elected not to use 2 features as they rely on external and potentially-changing evidence (F10, F11). 2 features we tested (F7, F8) did not produce any

---

2We were unable to replicate the results for this feature, because Valenzuela et al.’s dictionary of cue words was not available.
significant correlation with the gold standard.

Of Valenzuela et al.’s three remaining features, 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 et al., the worst feature of Valenzuela et al. (F9-simTitleCore) and a third for which the opinion regarding the usefulness of this feature was divided between the two studies (F4-aux_SelfCite). The sections that follow will provide a direct cross-comparison of these features on a single dataset:

- **Number of direct citations** (Integer): This feature was labeled by Valenzuela et al. as ‘F1 - Direct Citations’ and by Zhu et al. as ‘countsInPaperwhole.’ Both of these studies, and the earlier study by Hou et al. [2011], found 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 Valen-
zuela et al. study. Whilst Zhu et al. 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. also listed this as the 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 Valenzuela et al., and as ‘auxselfCite’ by Zhu et al.. The two studies differ markedly in their opinion of the value of this feature. While Zhu et al. found little correlation between author overlap and influence, Valenzuela et al. listed author overlap as their third-best-performing feature. It was therefore selected for further investigation.

<table>
<thead>
<tr>
<th>Valenzuela et al.</th>
<th>Zhu et al.</th>
<th>This Study</th>
<th>Feature description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author overlap</td>
<td>auxselfCite</td>
<td>Self-citation</td>
<td>Same author/s in citing and cited paper.</td>
</tr>
<tr>
<td>Direct citations</td>
<td>countsInPaperwhole</td>
<td>Number of citations</td>
<td>Number of times cited paper appears in text of citing paper.</td>
</tr>
<tr>
<td>Similarity between abstracts</td>
<td>N/A</td>
<td>Abstract similarity</td>
<td>Measured as cosine similarity between citing and cited paper abstracts.</td>
</tr>
</tbody>
</table>

Table 4.2: Description of features used for comparison with earlier studies.
4.2.4 Classification

Using the above features, a binary incidental / influential classification was performed. WEKA 3 [Witten et al., 2016] was selected as the machine-learning tool-set for this study. Weka is a collection of machine learning algorithms for data mining tasks developed at the University of Waikato. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. For these experiments the dataset was imported to the Weka Data Explorer which was used for text pre-processing, classification and for generating the results graphs.

4.3 Analysis and results

4.3.1 Dataset

The dataset released by Valenzuela et al. [2015] 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 / influential) label were provided. Using the author’s binary classification, 396 citation pairs were ranked as incidental citations, and 69 (14.3%) were ranked as influential citations.

It is interesting to note that studies which employed human annotators to judge citation influence (Zhao et al. [2017], Zhu et al. [2015], Valenzuela et al. [2015]) reported a broadly similar ratio of positive examples. This ranged from 10.3% by Zhu et al., through 14.3% by Valenzuela et al., to 17.9% by Teufel et al... to 16% found by Zhao et al. [2017] This is an important finding, as it gives a clear indication that only a relatively-small percentage of all citations can be considered influential. All of the studies found that the majority of citations were perfunctory at best. Negative citations were extremely rare, and this in itself further increased 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 at
first be imagined.

To obtain a clean dataset for the experiments, first the PDF files of the citing and cited papers used by Valenzuela et al. were collected from the ACL Anthology. These papers were processed using pdf2txt [Ma, 2013] to extract metadata, citations, the full text, and other document structure information. Any papers for which the extraction was not possible, or the abstract was not available, were then removed. This left a dataset of 415 pairs, with 355 citation pairs marked as incidental and 60 citation pairs (14.45%) marked as influential. 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, this dataset was considered to be sufficiently representative for the experiments. The XML files were then processed using ParsCit, and calculations were applied to extract features for each example.
4.3.2 Analysis and comparison of selected features

These experiments tested a range of features and their efficacy as predictors of citation influence using a Random Forest Classifier. The model was tested using bagging with 100 iterations and a base learner, using a 10-fold cross-validation methodology. The WEKA toolset [Witten et al., 2016] was used then also used to generate P/R curves for each of the individual features, as well as the combination of all the features (Table 4.3).

<table>
<thead>
<tr>
<th>Feature</th>
<th>P@R=0.05</th>
<th>P@R=0.1</th>
<th>P@R=0.3</th>
<th>P@R=0.5</th>
<th>P@R=0.7</th>
<th>P@R=0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Citations</td>
<td>0.4</td>
<td>0.34</td>
<td>0.33</td>
<td>0.3</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>Author Overlap</td>
<td>0.27</td>
<td>0.35</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Abstract Sim.</td>
<td>0.46</td>
<td>0.49</td>
<td>0.21</td>
<td>0.2</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>All</td>
<td>0.5</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.29</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 4.3: Interpolated precision at different recall levels for all features for the Random Forest Classifier

The correlation between each of the individual features and the classification given by the human annotators was also measured. Valenzuela et al. presented their results in terms of P/R values for each feature, whereas Zhu et al. showed the Pearson correlation with their gold standard. The results of these experiments are therefore presented in both formats to allow for accurate com-
parison. This work confirms the earlier findings reported in Zhu et al. and Valenzuela et al. that the number of direct instances of a citation within a paper is a clear indicator of citation influence. It also finds that author overlap, or self-citation, does have value as a classification feature. Contrary to the results reported by Valenzuela et al., this work found 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 tested. 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 4.3.2 shows the precision of the Random Forest Classifier at various recall levels. It can been seen from these results that the classifier

<table>
<thead>
<tr>
<th>Feature</th>
<th>Valenzuela et al.</th>
<th>this study</th>
<th>Zhu et al.</th>
<th>this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Citations</td>
<td>0.30</td>
<td>0.21</td>
<td>0.330</td>
<td>0.281</td>
</tr>
<tr>
<td>Abstract Similarity</td>
<td>0.14</td>
<td>0.14</td>
<td>N/A</td>
<td>0.373</td>
</tr>
<tr>
<td>Author Overlap</td>
<td>0.22</td>
<td>0.16</td>
<td>0.020</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Table 4.4: Comparison of results by feature
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.

4.3.3 Results for Individual Features

Number of Direct Citations - F1:

This feature is rated as the highest value in terms of predictive ability by Zhu et al. [2015] and the second highest by Valenzuela et al. [2015]. The latter shows P0.30 at R0.90; however the results of this study demonstrate a slightly lower P value: P0.21 at R0.90.

Zhu et al. listed the equivalent ‘countsinPaper_Whole’ as the most significant feature of their classifier, with a Pearson correlation coefficient of P0.35. We found 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. These 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 Valenzuela et al.’s results, this is the third-ranked ‘most significant feature’, with P0.22 for R0.90. We found slightly-less precision than Valenzuela et al. for this feature: P0.16 at R0.90. Zhu et al.’s results showed little correlation with their gold standard for the similar feature aux_selfCite (Pearson 0.02). In-
Interestingly, despite the low correlation, this feature was the fourth one selected by their model and did indeed improve the performance of the classifier, albeit only slightly. The experiments with our dataset show a far stronger positive correlation, $P_{0.132}$ (significant at the 0.01 level, 2-tailed), than that found by Zhu et al..

**Abstract Similarity - F9**

Whilst Zhu et al. generated many similarity-based features, they did not compare citing abstract and cited abstract. This is somewhat surprising, as it is 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 contains only valid data, i.e., the abstract is available for both citing and cited paper, a direct comparison can be made for this feature with Valenzuela et al., who ranked this as the lowest of their twelve features: $P_{0.14}$ at $R_{0.90}$.

Here the results of this study are the same as Valenzuela et al.’s, with $P_{0.14}$ at $R_{0.90}$. However, the Pearson correlation with the gold standard dataset for this feature is the highest of the three
features tested in these experiments, which found 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. These results demonstrate that abstract similarity between citing and cited paper is more predictive of citation influence than shown previously.

The value of complex features

It is now over a decade since Teufel et al. [2006a] 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 the complex features have been shown to have
little predictive ability with regard to classifying citation function or influence. Some of the most basic features have been shown to offer the strongest potential in identifying influential citations. This 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. It is essential that the types and values of all features should be provided as part of the research dataset (rather than providing just source prior to feature extraction) to serve as a roadmap for replicating the research. Further, features that rely on external datasets, changing evidence (e.g., citations, downloads, etc.), or sets of rules that are not available to other researchers cannot be replicated. There has now been a decade of research in this area, and the predictive ability of many complex features is still uncertain.
4.3.4 Replication of earlier studies

Despite significant efforts, the results for several of the features used by Valenzuela et al. [2015] or by Zhu et al. [2015] could not be reproduced or validated, largely because the data required to define or calculate a particular feature was not available in the dataset, or was based on a judgement that, again, was not available. As an example, one of the features used by Valenzuela et al. [2015] was `is_useful`. This is a subjective feature based on the opinion of the original annotator and it would therefore not possible to replicate exactly. Zhu et al. [2015] used several contextual features based on ’manually curated word lists’. Whilst some of these word lists are available within the published manuscript several are not.

Where the features in question can successfully be replicated, we find slightly different results compared to earlier studies as the datasets themselves cannot be exactly replicated. The dataset released as a part of the Valenzuela et al. [2015] study was a list of citing / cited paper pairs and an influential / incidental label
showing the relationship. The papers themselves are available as a part of the ACL-ARC collection but we were not able to exactly replicate the process. It is a similar situation with the study of Zhu et al. [2015], a dataset of the annotations and list of papers is available. In attempting to replicate these studies, collation and pre-processing of the papers themselves is required. The issues with extraction from PDF files is covered in the next section. Additionally, as noted in the introduction, features that have an external reliance may not be best suited for this task, as this has direct implications for replicability.

4.3.5 Analysis of PDF extraction

Both Valenzuela et al. [2015] and Zhu et al. [2015] use the ParsCit citation parsing tool, based on Conditional Random Fields (CRF). As this is a critical pre-processing stage, experiments were conducted to determine the efficacy and accuracy of this tool. GrobID [Lopez, 2009] 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 mis-tagged.

These errors unavoidably affect 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 these experiments demonstrate that this
is indeed the case in many instances. To understand the impact of this, the following experiment was conducted. Ten papers were chosen randomly from the Valenzuela et al. 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. Valenzuela et al. and Zhu et al. demonstrated that the number of times a citation appears in the body of the text is a significant indicator of influence. However, depending on the chosen method of parsing, there is a difference in the number of citations identified. The reference count for five of the chosen example papers is shown in Table 4.5.

<table>
<thead>
<tr>
<th>Paper ID</th>
<th>ParsCit count</th>
<th>GrobID count</th>
<th>Actual Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>C00-2140</td>
<td>33</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>W06-0202</td>
<td>17</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>W09-1118</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>E12-1072</td>
<td>31</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>P02-1058</td>
<td>13</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 4.5: Comparison of in-text citations counts by extraction method

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. This demonstrates
a potentially-serious failing in current methodologies that rely on PDF extraction for calculation of number of citations contexts and references.

4.4 Discussion

One of the major limitations of this and previous studies is the size of the publicly-available, annotated, datasets. The study by Valenzuela et al. [2015] used 465 citing / cited paper pairs. Due to the unbalanced split between the incidental and influential classes, the complete dataset used in this research contained only 61 examples of the positive (influential) class. 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; this is addressed in Chapter 5. The construction of a gold standard dataset containing many thousands of annotated citations, rather than a few hundred, is a significant undertaking, but this is a vital
step in improving the abilities of the classification models. This is addressed in the next chapter.

There is a noticeable difference between the datasets used by Zhu et al. and Valenzuela et al. that warrants further study. The Valenzuela et al. dataset annotation was undertaken by two independent annotators and found significant value in using author overlap as a classification feature. However, the Zhu et al. reference set was annotated by the authors themselves, and this study ranked 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 Teufel et al. [2006a] or Valenzuela et al..

Finally, 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.
4.5 Summary of findings and conclusion

This chapter presents an examination of PDF extraction tools and highlights the differences in performance. This has consequences for future studies that use these tools.

The central study undertaken in this chapter confirmed that only a few of the classification features suggested by earlier works offer much in way of predictive ability with regard to identifying influential / incidental citations. The study by Zhu et al. [2015] tested a range of 40 features, and reported that few had predictive ability. Of the features tested, Abstract Similarity showed the strongest positive correlation for predicting citation influence. Number of Direct Citations was also highly predictive, and Author Overlap / Self-Citation was less predictive but still valuable as a classification feature.

This study has identified several challenges to comparing classification features and assessing their performance:

1. Reproducibility: There is scope for further work concerning the efficacy and in particular the reproducibility of results
associated with some of the previously-tested classification features. Many of the earlier studies in this domain present results based on sometimes complex and irreproducible features. 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.

2. *PDF extraction:* These results, in demonstrating different success rates in extraction of full-text from PDF, also highlight 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. This may introduce a level of potential inaccuracy that has not been addressed fully.

3. *Size and nature of dataset:* Whilst it may be a relatively easy task for a human to identify important or influential citations, building a model to automatically classify these citations with a reasonable 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 unbalanced nature of any dataset of citations annotated according to influence or importance. For example, the central difficulty associated with the study by Valenzuela et al. is the size of the dataset used. It consists of just 465 citing / cited paper pairs and the classification label. Within this dataset there are just 61 examples (14.6%) of the influential class. No citation contexts are provided which would allow for the investigation of additional types of features. Further, the previous datasets in the domain have usually been compiled from papers in a single domain. If improvements in the automatic classification of citations according to type are to be made, more data, and data from a range of domains, is necessary. There is therefore scope for the development of a significantly larger dataset than those previously compiled.

4. Annotation: The limiting factor in compiling larger datasets of annotated citations has been the annotation phase itself.
The collection and aggregation of annotated citation data is slow, subjective and often expensive. Prior works have used domain experts, or the annotation has been performed by the studies’ authors themselves. In either case, this places a finite limit on the number of annotations that can be collected efficiently.

Chapter 5 introduces a novel, scalable methodology that enables the collection of a much larger dataset of annotated citations.
Chapter 5

Overcoming the Key Challenges in Capturing the Semantic Aspects of Citations
5.1 Introduction

In the previous chapter we identified and documented some of the key challenges in the automatic identification of citation influence. This chapter explores studies that have attempted the classification of citations according to type. Combining work from the previous chapter with the work presented here addresses the final research question:

**RQ3. What are, and how do we overcome, the key challenges in capturing the semantic aspects of citations?**

In section 2.12, *Defining and identifying reasons for citation*, three seminal studies which have attempted to classify citations according to type were introduced – Teufel et al. [2006a], Jurgens et al. [2016], and Cohan et al. [2019] – and their classification schemes were discussed. Each of these studies used a different classification schema, which increases the complexity of comparing the results. The methodology for the annotation process also differed. One of the limitations of these studies has been the size and class balance of the datasets used; this is a consequence of the
annotation methodology which have used either domain experts or the studies authors’ for the annotation process.

Overall, the limitations of these studies can effectively be summarised thus:

- Lack of cohesive classification schema
- Limited size datasets
- Only single or dual domain-datasets, not multidisciplinary
- Limitations of annotation process - scope / subjectivity

Section 5.2 introduces the citation classification schema used in this study which is compatible with prior works but adds a new layer of granularity for comparing and contrasting citations. Sections 5.3 and 5.4 address the current limitations in the collection of annotated datasets of citations and introduce our new annotation methodology using first authors as annotators, and a new platform for the collection of a multi-disciplinary dataset of annotated citations. Finally, Section 5.6 presents the results of a series of classification experiments using this new dataset and the dataset
produced by Jurgens et al. [2016]. Text classification models from SciKit [Pedregosa et al., 2011] are trained and applied to test the hypothesis that domain specific models perform significantly better than general purpose models trained on multi-disciplinary datasets.

5.2 Selection of a citation classification schema

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. The granularity of the class schema must be such that it is possible for automated techniques to carry out the classification task successfully.

Both the method of collection and the classification schema used will have an impact on the utility of the final dataset. As a starting point, this research used the following statements:

- All citations must have a class.
- No citation mention may be assigned to more than one class.
The first of these requirements is straightforward. The intention here is that there should be no ‘null’ class, and the categories should be broad enough to allow all citations to fit comfortably within a single class. The second point is somewhat more complex, as authors may cite a source several times in one paper. In one sentence, the authors may refer to a particular piece of work as motivation for their own study, and, in another, they may compare their results to those in the cited work. However, it will be demonstrated in the section on citation classification that distinctly different language can often be found within the citing sentence itself for each particular class. Whilst in some instances this is an over-simplification, this assumption holds in the vast majority of cases. This is a key part of the automatic identification of citation types this work builds primarily on the studies of Jurgens et al. [2016], Teufel et al. [2006b] and Valenzuela et al. [2015]. To allow for future cross-study comparison of results, the chosen classification schema is compatible with those of Jurgens et al. [2016] and Teufel et al. [2006b]; however it adds an additional layer to the compare/contrast category; show similarities,
show differences, or show disagreement. Although infrequent, it can be argued that citations that demonstrate differences from or disagreement with earlier works are potentially some of the most important. The final classification schema can be seen in Table 5.1

5.3 Annotation methodology

This study differs from previous work, by employing authors as annotators, rather than using independent annotators. As discussed in Case and Higgins [2000] (Page 636), asking authors may poten-
tially introduce problems of both recall, i.e. will the citing author remember their reasons for citing?, and the “social desirability” of answers, i.e Will they answer honestly?. This is of particular relevance in the case of negative or contradictory citations. However, Case and Higgins concluded:

“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 of citer motivation.”

As Teufel et al. [2006b] (Page 106) noted:

“Citation function is hard to annotate because it in principle requires interpretation of author intentions…”

In Teufel et al. [2006b] the inter-annotator agreement between three annotators (the studies authors) was $K=0.73$. The study conducted by Cohan et al. [2019] took a different approach as
they used volunteers rather than experts for the annotation process. A *confidence score* was assigned to each annotation based on agreement between three annotators. Instances with a confidence score of $\leq 0.70$ were discounted from the final dataset. In Valenzuela et al. [2015], inter-annotator agreement was given as 93.6%. This is however for a binary classification. Jurgens et al. [2016] did not give details of the inter-annotator agreement in their studies. It can been seen from the studies that do report inter-annotator agreement that the agreement level is far from 100%, this is whether the annotation is performed by domain experts, study authors or volunteers. This adds weight to the hypothesis that annotation by third-parties adds a layer of subjectivity to the final outcome.

These observations were highly influential to the approach taken in this research. Annotation by domain experts is slow, expensive, requiring domain expertise, and prone to subjective biases. Whilst authors may have their own biases, we agree with Case and Higgins that asking authors to annotate their own papers is effective. Asking authors to annotate their own citations removes
a layer of interpretation. No previous study has, to our knowledge, undertaken large-scale engagement of authors to annotate citations.

5.4 A new platform for annotating citations

For the annotation process, we developed and deployed a new platform for annotating citations using Python, MySQL, JavaScript and React. The online annotation tool is named the Academic Citation Typing (ACT) platform [Pride et al., 2019]. This displays the full text of the authors’ research paper alongside a point-and-click style classification interface. In-text citation markers are highlighted automatically as a visual prompt for the annotator, and are displayed alongside the cited paper’s title, author name, publication date, and the full sentence containing the citation. This allows authors to assign one of the six class labels to each citation in their paper rapidly and accurately.

The ACT platform was, prior to launch, tested for user experience and reliability with six evaluators who were PhD students
or researchers at The Open University. To ensure that the tests were as realistic as possible, the evaluators were the first authors of the sample papers used in the annotation process. Both our observations and feedback from the evaluators who completed the annotation process conveyed 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. The ACT platform also recorded the time taken by each author to complete the annotation process, which was an average of nine minutes, around 22 seconds per citation.

The six evaluators who tested the ACT platform were interviewed post-annotation. These interviews provided two key observations: authors very often remembered their reason for citing a particular paper, and provision of the title and author name was largely sufficient for the author to remember the reason for the citation. A live demonstration of the platform in use can be viewed on YouTube\footnote{https://youtu.be/8l7frJ-fde8}.

5.5 ACT dataset - collation and comparison

The dataset compiled for the study in Chapter 3 consisted of the metadata for 190,628 papers submitted to REF2014, and the metadata retrieved from MAG for the 6.95-million papers that cited the REF2014 papers.
CORE (Knoth and Zdrahal) is the largest collection of Open Access scientific literature available online and comprised over 24.5m full text papers as of September 2020. A cross-comparison was carried out to identify how many of the 6.95-million citing papers from the previous study were open access, and if the full-text of the document was available in CORE. The total was approximately 103,000 papers. First author names, email addresses, and citing sentences were extracted from these papers using GrobID [López-Cózar et al., 2012]. Any record for which the data was incomplete was discarded. The remaining data comprised 26,652 papers and associated meta-data, and 407,907 citation contexts.

This data was then uploaded to the ACT platform, which automatically generated a unique URL token for each paper. Invitations to take part in the annotation process were then sent via email to first authors. 924 authors (i.e., covering 924 citing papers) responded, representing a 3.4% response rate, and annotated a total of 11,233 citation contexts according to type and influence. Note that this represents publications from 924 discrete first authors; we did not ask for more than one set of annotations
(corresponding to a single paper) from any given author.

Each record in the full Academic Citation Typing (ACT) dataset of annotated citations contains:

- the citing paper title, author, and publication date;
- the cited paper title and author;
- the extracted citing sentence;
- the author-annotated class label;
- the CORE paper ID\textsuperscript{2}, which is the source of the full-text paper; and
- the Microsoft Academic Graph (MAG) ID\textsuperscript{3}, which can be used to retrieve further bibliographic and bibliometric data for each paper.

As shown in 5.3, the ACT dataset is the largest collection of citations annotated according to type and influence, is the only author-annotated dataset and the only multi-disciplinary one. The class breakdown of the final ACT dataset is shown in Table 5.2.

\textsuperscript{2}https://core.ac.uk
\textsuperscript{3}https://academic.microsoft.com/home
### Table 5.2: Breakdown of ACT dataset by class of citation

<table>
<thead>
<tr>
<th>Class of Citation</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACKGROUND</td>
<td>54.61%</td>
</tr>
<tr>
<td>USES</td>
<td>15.51%</td>
</tr>
<tr>
<td>COM/COM</td>
<td>12.05%</td>
</tr>
<tr>
<td>MOTIVATION</td>
<td>9.92%</td>
</tr>
<tr>
<td>EXTENSION</td>
<td>6.22%</td>
</tr>
<tr>
<td>FUTURE</td>
<td>1.70%</td>
</tr>
</tbody>
</table>

#### 5.5.1 Comparison of datasets and collection methodologies

This section looks at how previous studies have collected and annotated data, and compares this to our collection method. Table 5.3 shows the total number of papers, annotators, and citations in this study and in three seminal studies in this domain.

<table>
<thead>
<tr>
<th>Study</th>
<th>Papers</th>
<th>Annotators</th>
<th>Annotated by</th>
<th>Citations</th>
<th>Discipline(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teufel et al.</td>
<td>360</td>
<td>3</td>
<td>Study Authors</td>
<td>2,829</td>
<td>Comp. Linguistics</td>
</tr>
<tr>
<td>Jurgens et al.</td>
<td>185</td>
<td>3</td>
<td>Domain Experts</td>
<td>1,989</td>
<td>Comp. Linguistics</td>
</tr>
<tr>
<td>Pride &amp; Knoth</td>
<td>924</td>
<td>924</td>
<td>Paper Authors</td>
<td>11,233</td>
<td>Multi-disciplinary</td>
</tr>
</tbody>
</table>

Table 5.3: Cross-study comparison of dataset collation and annotation type

Whilst the studies by Teufel et al. [2006b] and Jurgens et al. [2016] used domain experts for the annotation process, the study by Cohan et al. [2019] employed crowd-sourced volunteers. Cohan et al.’s volunteers were trained in the annotation process, guided by a domain expert. Several steps were undertaken to ensure the
accuracy of the crowd-sourced annotations. A confidence score, measuring the level of agreement between multiple annotators, was used. Annotations with a confidence score of \( \leq 0.7 \) were discarded. In addition, a subset of the dataset with 100 samples was re-annotated by a trained, expert annotator to check for quality, and the agreement rate with the crowd-sourced annotators was 86%. The final Sci-Cite dataset produced by Cohan et al. contains 11,020 annotated citations taken from papers in biomedicine and computer science.

5.5.2 Authors or third-party annotators

In selecting to use authors as annotators our novel methodology has one potential caveat, that is the annotations are carried out by a single individual. The approach of the earlier studies necessitated a method to demonstrate sufficient agreement and hence used multiple annotators (whether experts or the study’s authors themselves). For these earlier studies it was therefore necessary to ascertain a degree of confidence in the annotations by measuring the inter-annotator agreement in someway. It can also be seen
that the inter-annotator agreement in these studies is not 100%. This indicates a level of subjectivity in the annotation process with differences of opinions between annotators. We believe this adds credence to our decision to enlist authors as annotators.

We took several steps to ensure the annotations by authors were as accurate and genuine as possible. During testing of the ACT platform, the time taken for each annotation was recorded and this allowed us to gauge how long the process should take. With this information we were able to ensure that the authors taking part in the full data collection survey were actually undertaking the process correctly and not simply clicking randomly. Additionally, we undertook a post-hoc survey of the collected data to check for accuracy. A sample of approximately 150 annotations were checked against the author-assigned labels. We of course acknowledge that this in itself is a subjective process. A sample of the citations and their associated labels can be seen in Table 5.4.
They are typically classified as collaborative filtering approaches, content-based filtering approaches and hybrid approaches. In addition, previous studies from the American and Canadian Arctic have also reported differences in the size of polar cod between inshore and offshore areas. First, we were aware of a small but not unsubstantial body of quantitative evidence on smokers who quit unassisted; and second, in the course of our literature search we had identified a considerable number of qualitative studies on smoking cessation. Age improvements in pooling efficiency would be consistent with childhood brain changes including synaptic pruning and myelination, although their dependence on development of specific cortical or subcortical networks remains to be investigated. Such mechanisms need to be understood at several distinct levels of analysis - for example, following influential formulation, both the algorithms used for sensory processing and their "biophysical" underpinnings need to be understood.

We aimed to estimate the incidence and prevalence of MS in BC, Canada using previously validated case definitions of MS based on health administrative data.

Table 5.4: Examples of citations and labels from ACT dataset.

<table>
<thead>
<tr>
<th>Background</th>
<th>They are typically classified as collaborative filtering approaches and hybrid approaches.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare/Contrast</td>
<td>In addition, previous studies from the American and Canadian Arctic have also reported differences in the size of polar cod between inshore and offshore areas.</td>
</tr>
<tr>
<td>Extension</td>
<td>First, we were aware of a small but not unsubstantial body of quantitative evidence on smokers who quit unassisted; and second, in the course of our literature search we had identified a considerable number of qualitative studies on smoking cessation.</td>
</tr>
<tr>
<td>Future</td>
<td>Age improvements in pooling efficiency would be consistent with childhood brain changes including synaptic pruning and myelination, although their dependence on development of specific cortical or subcortical networks remains to be investigated.</td>
</tr>
<tr>
<td>Motivation</td>
<td>Such mechanisms need to be understood at several distinct levels of analysis - for example, following influential formulation, both the algorithms used for sensory processing and their &quot;biophysical&quot; underpinnings need to be understood.</td>
</tr>
<tr>
<td>Uses</td>
<td>We aimed to estimate the incidence and prevalence of MS in BC, Canada using previously validated case definitions of MS based on health administrative data.</td>
</tr>
</tbody>
</table>

5.5.3 Comparison of class distribution

As this work is most closely aligned with that of Jurgens et al., the six-way classification schema was retained, as previously noted. Figure 5.2 shows the comparative breakdown of the datasets from these two studies into classes of citation. Whilst there are some differences in the distribution of citation types between the two studies, overall there is a strong positive correlation between the two: \( R=0.93, p \leq 0.04, n=6 \).

The largest differences can be observed in the Motivation and Extension classes. Two-tailed z-scores were calculated for the Method class, \( z=6.99, p<0.00001 \), and for the Extension class, \( z=4.31, p<0.00001 \). It is interesting to note that authors are
statistically more likely than independent annotators to regard a piece of work as motivational or as an extension to previous work, rather than simply as background information. There was no statistically significant differences in author and expert opinion across the remaining classes.

It is likely that this type of insight can only be provided by the author of the citing paper. Whilst differences are observed, the overall distribution of the data in the ACT dataset is very similar to datasets annotated by domain experts.
5.5.4 The effects of changes in class schema

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 seen most clearly by comparing the dataset from Cohan et al. to those from Jurgens et al. and this study. Cohan et al. took the six-way classification schema suggested by Jurgens et al. and grouped citations into just three classes: ResultsCompare, Method, and Background. This simplification entailed re-labelling citations from the Extension, Motivation, and Future classes as Background. Cohan et al.’s rationale for this was that these three classes were less informative than the others.

To allow for a direct cross-study comparison, this study followed the same methodology used by Cohan et al. To directly compare the class distribution differences between the datasets used in the three studies, we use the same methodology as Cohan et al. This process entailed collapsing the six classes used in the ACT dataset and the Jurgens et al. dataset to the three classes used by Cohan et al..

When the class granularity is reduced to three overall classes,
it is reasonable to expect some changes in class distribution. The final breakdown across all three studies is presented in Figure 5.3.

![Figure 5.3: Comparison of datasets when class granularity is reduced.](image)

The most notable difference in the dataset produced by Cohan et al., 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., for which the figure is 18.5%, and the ACT dataset at 15.5%. It is known from
personal correspondence with the authors that this difference was produced by the deliberate over-sampling of citations from the less-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 overly-optimistic results when the resulting dataset is then used to train machine learning models. As Vandewiele et al. [2020] (Page 6) noted 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.”

In their 2019 study, Cohan et al. performed a series of classification experiments using a Bidirectional-LSTM (Long/Short Term Memory) classifier [Schuster et al.] and contextualized word embeddings (ELMo, Peters et al. [2018]). They conducted classification experiments with the ACL-ARC dataset as used in the
Jurgens et al. [2016] study and with their own SciCite dataset. At the time of their study the results of Jurgens et al. were SoTA for the six-way citation classification task, with a macro F1 score of 54.6. The model tested by Cohan et al. [2019] on the ACL-ARC dataset improved this result to 67.9. They then ran the classification task using the same model on the SciCite dataset and achieved a significantly improved result, macro F1 84.0. This is indicative of the effects of reducing the number of class labels from six to just three. These results are shown in Table 5.5

<table>
<thead>
<tr>
<th></th>
<th>Jurgens et. al</th>
<th>Cohan et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACL-ARC</td>
<td>54.6</td>
<td>67.9</td>
</tr>
<tr>
<td>Sci-Cite</td>
<td>N/A</td>
<td>84.0</td>
</tr>
</tbody>
</table>

Table 5.5: Comparison of micro F1 results from Cohan et al. [2019]

It can be seen that whilst the model used by Cohan et al. [2019] does offer improvement on previous works, the most significant difference can been seen in the results between the two different datasets. The ACL-ARC dataset (Bird et al. [2008]) is a widely know corpora of papers drawn from the computer science domain. The Sci-Cite dataset was compiled by Cohan et al. for this particular study. The differences in results for the same mod-
els on the two different datasets highlights the effects of both the class distribution and over-sampling during the collation phase.

As these models are being used in a real-world setting, in applications such as Semantic Scholar\textsuperscript{4}, the results have real impact on how scientific literature interpreted and disseminated. It is therefore important to ensure that solutions built on these models actually work as expected when used on real-world data. Further work is required in this area to investigate the accuracy of these systems.

5.6 Citation classification experiments

The previous section discussed the collation of the ACT dataset and the comparison with previous datasets. In addition to compiling the datasets, the studies of Teufel et al. [2006b], Jurgens et al. [2016] and Cohan et al. [2019] have all then used these datasets for training text classifiers that attempt to identify citations automatically according to the types defined in the classification

\textsuperscript{4}https://semanticscholar.org
schema.

5.6.1 Citation classification using Jurgens et al. and ACT datasets

Using text classification tools from SciKit (Pedregosa et al. [2011]), a range of experiments was conducted with the ACT and Jurgens et al. [2016] datasets. The classification task is for the model to assign a single class from the citation schema to each of the citation contexts.

Experiments were conducted with two models that capitalised on word embeddings using Glove (Pennington et al. [2014]) and Word2Vec (Mikolov et al. [2013a]). We implemented an embedding vectorizer, a counterpart of the CountVectorizer and TfidfVectorizer methods available in SciKit. Given a word to vector mapping this vectorizes texts by taking the mean of all the vectors corresponding to individual words. We also ran experiments with three models that used ‘classic’ text classification methods, two based on Naïve Bayes classifiers and a third using a linear kernel Support Vector Machine (SVM). All of the above models
Table 5.6: Classification models used in experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mult_nb</td>
<td>Multinomial Naive Bayes Classifier</td>
</tr>
<tr>
<td>bern_nb</td>
<td>Bernoulli Naive Bayes Classifier</td>
</tr>
<tr>
<td>SVM</td>
<td>Linear kernel Support Vector Machine</td>
</tr>
<tr>
<td>glove_small</td>
<td>ExtraTrees Classifier with 200 trees and vectorizer based on 50 dimensional GloVe embedding trained on 6B tokens</td>
</tr>
<tr>
<td>glove_big</td>
<td>As above but using 300 dimensional GloVe embedding trained on 840B tokens</td>
</tr>
<tr>
<td>w2v</td>
<td>As above but using 100 dimensional word2vec embedding trained on the benchmark data itself (using both training and test examples)</td>
</tr>
</tbody>
</table>

were also tested with tf-idf features. Descriptions of these models can be seen in Table 5.5. Evaluation was performed using 10-fold cross-validation, and the micro F1 score was used to report results.

These experiments were to ascertain if there was a difference in performance when using the same classification models on different datasets and, if so, what the reasons for this difference might be.

Table 5.7 shows the results for the five most effective models for the classification task for each dataset. The results shown are the micro F1 score for each model.

It can be observed that there is a statistically significant difference in the results and all models perform less well when using
Table 5.7: Comparison of results (micro F1 score) of NLP models on Jurgens et al. [2016] and ACT datasets

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>SVM_TfIdf</th>
<th>Multi-NB</th>
<th>GloVe_big</th>
<th>w2v</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jurgens et al.</td>
<td>0.68</td>
<td>0.68</td>
<td>0.62</td>
<td>0.62</td>
<td>0.57</td>
</tr>
<tr>
<td>ACT_2k</td>
<td>0.51</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.5</td>
</tr>
</tbody>
</table>

This experiment helps us to better understand the reasons behind this difference as our approach limits the number of varying parameters compared to previous studies. More specifically, if one examines the two datasets, there are many similarities:

- Identical class schema
- Identical dataset size (1,989 citations from ACT)
- Similar citation context length (avg. 212 vs. 230 characters)
- Obfuscated author names.

The ACT_2k dataset for these experiments was made up of 1,989 citations drawn at random from the full ACT dataset. This was to ensure an exact match with the size of the dataset from Jurgens et al., and the class breakdown of the dataset was also checked for similarity. Additionally, Jurgens et al. use `@@citation` as a
marker within the citation context as a replacement for the author’s name. The ACT dataset uses \#AUTHOR\_TAG for the same purpose. This aims to prevent any algorithm attempting to infer information from the author’s name. Finally, citation length was also checked and found to be similar in the two datasets.

There are, however, some clear differences. Notably:

- Authors vs. expert annotators

- Single vs. multiple domains

The first of these was covered in the previous section, where it was shown that overall there is a strong correlation between annotations by authors and those of domain experts ($R=0.93, n=6$).

The second, however, is more nuanced, and necessitated the collation of subject identifiers for each of the papers in the ACT dataset. The following section details the approach used in solving this problem.
5.6.2 Domain coverage of the ACT dataset

One of the central differences between the new ACT dataset and existing ones is its multi-disciplinary nature. As the experiments in the following section will demonstrate, the composition of the dataset is key when training a classifier to identify citations automatically according to type. Using a technique known as hierarchical topic modelling, Microsoft Academic Graph (MAG) assigns one of 19 top-level domains to each paper within its collection. The MAG API was used to retrieve this top-level domain for 5,000 titles selected at random from the full ACT dataset (hereafter referred to as ACT_5k). Titles were matched using cosine similarity, and any for which the result was less than 0.90 were discarded. In this way, 3,727 titles were matched, representing 74.54% of the total dataset. Figure 5.4 shows the breakdown by domain for the ACT_5k dataset, based on MAG title data.

Whereas the ACT dataset contains citations from 19 different domains, the dataset from Jurgens et al. [2016] contained citations from just one, computer science. In fact, the papers were drawn
Figure 5.4: Domain breakdown of the Academic Citation Typing (ACT) dataset exclusively from the NLP domain and therefore were from a subset of a single domain and all highly-related to each other.

5.6.3 **Does language change across domains?**

If the language used in citations changes markedly across different domains, this may provide an explanation of the difference in classification performance reported in the previous section. The
addition of domain / subject area data to each of the citations in the ACT dataset provides new opportunities to investigate how the language used in citations changes across different domains. It would be logical to expect that specific terminologies would only apply to specific domains; it is less clear how much the language used in citations would differ across domains.

To examine if and how the language changes, the ACT_5K dataset and the associated domain for each citation context were used. Using Python MatPlotLib (Hunter [2007]), a weighted word-cloud was then created for the 100 most-frequent terms for each domain.

Figure 5.5 shows, for illustration purposes, the word-clouds for four of these domains: business, medicine, geography, and physics. There is a visible difference in the overt terminology used in each area. Four examples are shown for brevity; however all domains exhibit similar differences. The citation contexts from the ACT dataset were notably different with respect to content and terminology across each of the 19 domains. However, this in itself does not reveal the extent of the differences in cue phrases and other
linguistic elements.

5.6.4 Citation classification using single and multi-domain datasets

To test the hypotheses that the language used within citations changes across different domains, and that this has an affect on the results of a classifier trained on a single domain, a range of classification experiments was conducted using the same experi-
mental setup as in the previous section. A range of datasets was compiled to allow for accurate cross-comparison of the models. Table 5.8 shows the breakdown of the datasets for these experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT_500</td>
<td>500 random citation contexts</td>
</tr>
<tr>
<td>ACT_1k</td>
<td>1k random citation contexts</td>
</tr>
<tr>
<td>ACT_Psy800</td>
<td>826 citation contexts - Psychology</td>
</tr>
<tr>
<td>ACT_Med800</td>
<td>796 citation contexts - Medicine</td>
</tr>
<tr>
<td>ACT_Bio500</td>
<td>490 citation contexts - Biology</td>
</tr>
</tbody>
</table>

Table 5.8: Composition of single- and multi-domain datasets

In compiling a dataset of citation contexts from a single domain, the total amount of data available for the task was limited. The largest single domain was psychology, with 826 citation contexts (ACT_Psy_800 dataset). The next was medicine with 796 (ACT_Med_800 dataset), followed by biology with 490 (ACT_Bio_500 dataset). For these experiments, these three datasets were then complemented by two others comprising 500 and 1000 citation contexts (ACT_500 and ACT_1K, respectively) drawn at random from the full ACT dataset.

Each model’s classification performance was tested on each of
the five datasets. In all cases, for all datasets, all models performed markedly better when classifying a single-domain dataset.

If domain dimensionality did not exert any influence, one would expect to see the models achieve the same results with the ACT_Bio_500 dataset and the ACT_500 dataset. Figure 5.6 shows that this is not the case. All models achieved considerably better results for the classification task with the single-domain datasets, in comparison to the multi-domain datasets. For the datasets with 500 samples, the increase in performance ranged...
Table 5.9: Comparison of results (F1 score) for multiple models on single- and multi-domain datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM</th>
<th>SVM-TfIdf</th>
<th>Multi-NB</th>
<th>GloVe</th>
<th>w2v</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT_500</td>
<td>0.44</td>
<td>0.49</td>
<td>0.45</td>
<td>0.48</td>
<td>0.42</td>
</tr>
<tr>
<td>ACT_1K</td>
<td>0.49</td>
<td>0.55</td>
<td>0.56</td>
<td>0.55</td>
<td>0.44</td>
</tr>
<tr>
<td>ACT_Psy800</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>ACT_Med800</td>
<td>0.62</td>
<td>0.67</td>
<td>0.68</td>
<td>0.64</td>
<td>0.6</td>
</tr>
<tr>
<td>ACT_Bio500</td>
<td>0.57</td>
<td>0.59</td>
<td>0.58</td>
<td>0.58</td>
<td>0.5</td>
</tr>
</tbody>
</table>

from 19.1% (w2v) to 29.5% (SVM). The results are similar when comparing the ACT_Psy_800 and ACT_Med_800 datasets with the ACT_1K dataset; there is again a significant increase in the classifier’s performance across all models, from 18.1% (GloVe) to 36.8% (SVM). Table 5.9 gives all F1 scores for all models and datasets.

The classification experiments in this chapter demonstrate that the size, the citation class breakdown, and the domain dimensionality all have an effect on the results of the automatic classification of citation contexts. Previous studies, notably Cohan et al. [2019], have attributed performance increases in classification models solely to improvements in the models themselves. Our results, and the earlier comparison shown between Cohan et al. [2019] and Jurgens et al. [2016] in Table 5.5, show that the
composition of the dataset and the class structure is a key component of the classification process and have a statistically significant impact on the results.

5.7 Summary of findings and conclusion

The output from this study is two-fold: the ACT dataset, and the ACT platform. First, the ACT dataset created in this work addresses one of the current problems in the domain, that of the availability of suitably large datasets for classification. The final ACT dataset consists of 11,233 citations annotated according to both type/purpose and influence on citing paper. This is the largest dataset of its type, the only truly multi-disciplinary dataset, and the only one annotated by authors.

Second, the ACT platform itself is fully-scalable and is compatible with all PDF files. A portion of the ACT dataset was released publicly and was used for a shared-task competition to find the best models for citation classification according to type and influence. The shared-task took place at the Workshop On mining
Scientific Publications (WOSP2020) held in July 2020 in conjunction with the Joint Conference on Digital Libraries (JCDL2020).

This study shows that authors are an effective source of valuable additional citation information. Further, it demonstrates that, whilst citation annotations by authors show close correlation to those by independent domain 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 schema also adds the ability to recognise citations that explicitly disagree with or contradict previous works. This information is not available with previous citation classification schema, nor is it part of current citation-based metrics. The importance of having the ability to classify citations according to type has often been overlooked.

The classification experiments in this chapter demonstrate that the size, the citation class breakdown, and the domain dimensionality all have an effect on the results of the automatic classification of citation contexts. The change in results that can be attributed
to the changing language across a wide range of domains is a new finding, as previous studies have focused on citation contexts from one or two domains at most.
Chapter 6

Findings, Implications and Future Work
6.1 Findings and contributions

The overarching goal of this work is to answer the question, *How can we effectively identify and capture the semantic aspects of citations at scale?*. The discussion in Chapter 2 highlighted that there are many reasons why authors cite prior works, and treating all of these citations equally may not be good practice. Citation data are already (and somewhat controversially) being used globally as a proxy for research quality in Performance-based Research Funding Systems (PRFS), such as the REF2014 exercise. Therefore, capturing and leveraging the semantic aspects of citation may lay the groundwork for providing enriched citation data to peer review panels in such exercises. This data may also form the foundations of new bibliometrics that can also use this additional information.

Chapter 3, addressed the question, *To what extent do peer review and current bibliometric indicators correlate at the institutional / discipline level?* and presented the results of the first large-scale analysis of the outputs submitted to the REF2014 ex-
exercise. This study found a strong correlation between the aggregate peer review results for each institutes’ outputs and simple metrics based on citation counts. This correlation was particularly strong in the Units of Assessment (UoAs) whose peer review panels opted to use citation data as a part of the review process. Whilst this may seem unsurprising, it raises the question of how much value the peer review process added in these UoAs, over and above what could have been achieved using citation data alone. Additionally the study showed that it was possible to use citation data to predict an institutes’ GPA ranking position with a degree of accuracy in these same UoAs. The outcome of this study is not the recommendation that peer review be removed from these exercises, however these are important considerations as costs for the peer review portion of the REF exercise was £42m HEFCE [2014c].

Having demonstrated a strong correlation between peer review results and bibliometrics, this work then investigated how much more information may be gained from citations, over and above simple bibliometrics based on citation counts. The study in chap-
ter 4 asked, *What are the key features that indicate a citation is influential, rather than just incidental?*

This chapter investigated the additional information that can be gained from the actual citations themselves and examined a range of features tested by prior studies that have attempted to automatically identify *influential* citations; that is, the citations that are central to the citing paper.

This study found that many of the suggested features from prior works exhibit little in the way of predictive ability for identifying influential citations. Only three features were confirmed in line with earlier work as having some predictive ability. These were the total number of times a citation appeared in a paper (*precision* 0.4, *P@R*=0.05); the similarity between the abstracts of the citing and cited papers (*precision* 0.46, *P@R*=0.05) and overlap between authors (*precision* 0.27, *P@R*=0.05). Importantly, we identified reproducibility problems with several of the features applied in prior studies. This was either due to the unavailability of data which would be required to calculate a particular feature or where the feature was based on a subjective judgement of the
authors. One of the overriding issues with the seminal studies in this area, including Valenzuela et al. [2015], Teufel et al. [2006b] and Jurgens et al. [2016], is the small size of the datasets used. This study, and the results presented in chapter 5, demonstrate that a much larger dataset is required to effectively train machine learning models in the classification task, particularly when the dataset is not balanced across the classes as is naturally the case with datasets of annotated citations.

Additionally, chapter 4 completed a review of two PDF extraction tools, ParsCit (Councill et al.) and Grobid (Lopez) and demonstrated that these tools suffer from some limitations. Neither of the tools were completely accurate in parsing reference and citation data from PDF files. This is an important caveat for future studies that rely on these tools as a part of the pre-processing stage of any experiments.

In addition to studies identifying and capturing influential citations, others have attempted to classify citations according to type. Chapter 5 built on these earlier works and addressed the question; *How can we identify and overcome the key challenges*
in capturing the semantic aspects of citations?

It was shown that the key challenges can be summarised as; a lack of a standardised classification schema, a lack of suitably sized annotated datasets for training effective classification models and the lack of multi-disciplinary datasets.

The novel citation annotation methodology introduced in this work removes the constraints presented in earlier studies where far smaller datasets have been the result of the annotation process. The fully scalable nature of the ACT platform means the collation of annotations is no longer reliant on domain experts for the annotation process. The new ACT dataset of 11,233 author-annotated citations is the largest dataset of its type and is the only multi-disciplinary dataset. This now enables further research into the automatic classification of citations.

Our results also demonstrate that there is strong correlation between expert and author opinion when classifying citations. This study shows, however, that authors are statistically more likely to regard a cited work as motivational, or to regard their work as an extension of previous studies. Additionally, annotation by
authors removes a potential layer of subjectivity when external annotators are used.

Furthermore, the classification experiments in this chapter demonstrate that both the size and domain distribution of the dataset are key factors in the training of any machine learning model. This is a new finding and has implications for research and applications which have been based on models trained on single domain datasets. The improved performance of the classification models when applied on a single domain dataset are an indication that further work is required in this area. We suggest the need for a clear benchmark in this domain and work towards establishing this with the creation of the 3C Shared Task held at the WOSP2020 workshop. The results from these shared task can be used as a benchmark to measure improvements in deep learning models that use the standardised classification schema and ACT dataset.

Finally, the completed ACT platform that was developed and deployed as part of this work can now be utilised by a range of stakeholders. There are strong use cases for adoption of this
technology at the point of publication or deposit with publishers, repositories, journals, and conference systems. It is easy to envisage a scenario in which a publisher or content provider is able to provide enhanced bibliographic information using author-annotated citations.

6.2 Implications and Limitations

The use of citation data as a ‘quick and dirty’ measure of research quality has been shown to be problematic. Metrics are used often specifically as a proxy and offer enticingly quick results when measurement at scale is required. However, there are ramifications when metrics are used as a part of evaluating research and, in particular, if used to evaluate researchers themselves. Investigating whether metrics that use enhanced citation data can be more informative is one of the key implications of this work.

This work and the studies referenced in chapters four and five have focused on moving beyond a citation as a single numeric value. Valenzuela et al. [2015] first framed the notion of influen-
tial citations and it is this concept that underpins much of the research presented here. The studies presented demonstrate that many citations are indeed perfunctory or would fall into the class originally defined by Garfield as *paying homage*. The introduction to this body of work outlined the current usage of citation data in two main areas; information retrieval and research evaluation. Much of the work presented here focused on the latter of these two uses. However, as a researcher in search of relevant literature, the identification of papers considered by the citing author to be central to the work is a useful piece of information that can be used to increase relevancy in academic search results. However, in terms of research evaluation, whilst it is indeed useful to see which citations authors consider to be influential, this binary classification can be improved. It is more useful to identify why a piece of actual research was cited using the types of classification schema introduced by Jurgens et al. [2016] and Cohan et al. [2019]. The detail provided by a broader classification scheme rather than a simple binary influential / incidental classification provides richer information that can then be leveraged further.
Our study demonstrates that simple citation-based metrics and peer review rankings are highly correlated in many disciplines in the UK’s REF2014 exercise, particularly in disciplines where the peer review panels selected to use citation data. Our results also suggest that some disciplines that did not officially use citation data were likely influenced by it. This raises questions around the added-value that the peer review process brought in these disciplines and whether the overwhelming cost of these exercises is justifiable. We do not suggest that peer review is replaced with metrics based on citation data in these exercises. However, if citation data is being used in this way, we argue that citation importance and purpose can be used to simplify the job of REF peer-reviewers, thereby offering the potential to make the evaluation process faster and less costly. However, further work is required in this area. This is increasingly relevant when the burden of the peer review process was shown to be particularly onerous for many of the reviewers themselves. It has also been argued that the peer review process employed in the REF2014 was not equitable with the standards of refereeing required for publication in
an international journal - the REF’s own benchmark. As noted by Gadd [2019], the peer evaluation of REF outputs "defies many academics’ views of expert peer review”.

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.

The accurate and automatic identification of citation type, purpose and influence offers exciting possibilities for gaining new insights into the semantic aspects of citation by better understanding why authors cite certain papers at a macro level. However, this work shows there are significant concerns around the accuracy of models previously developed for this purpose, more needs to be done to develop high performing models generally applicable across scientific domains and there is a pressing need for a solid benchmark (Kunnath et al. [2020]) which would allow direct and fair comparison of methods.

The impact of the changing language across domains on the
performance of classification models had not been considered in previous works. Our experimental results reveal the limitations of previous studies, and also this one. Whilst the ACT dataset is the largest of its type, it is a broad dataset covering 19 different domains. It is therefore, in parts, fairly shallow. The breakdown of the domains in 5.4 shows, for example, that the life sciences are overly represented compared to the humanities. Training deep learning models to accurately classify the least represented classes, from the least represented domains, will require substantially more data. The ACT platform, however, provides the tool and scalable approach which can overcome this challenge.

Applications such as Semantic Scholar \(^1\) and Scite.ai \(^2\) incorporate tools based on the automatic identification of citation type. The work presented in this thesis demonstrates that the models in use are trained on single-domain datasets and the classification accuracy of these models therefore will be reduced when presented with works from a previously unseen domain. This presents a

\(^1\)https://semanticscholar.org  
\(^2\)https://scite.ai
significant ongoing challenge to these types of application and it remains to be seen how valuable these types of applications currently are to end users.

6.3 Future work

The natural first extension to the work presented here is the continued deployment of the ACT platform to gather more data to enable better training of deep learning models. The selected classification schema for this research is compatible with prior works and this standardisation marks significant progress in the domain. Our approach using authors as annotators has been shown to be both effective and scalable. A total of just over 100,000 papers from CORE (Knoth and Zdrahal) were used in the initial collection of data for the annotation process. This is from a total pool of some 24.5m full-text papers so there is no shortage of available literature that can be utilised for this process. Gathering larger datasets presents the opportunity for future experiments with deep learning models. The application of the models used by
Cohan et al. [2019], amongst others, such as BERT (Devlin et al. [2018]) and SciBert (Beltagy et al. [2019]) have not been fully evaluated on multi-disciplinary datasets. The new ACT dataset allows for this. Additionally, the current SotA deep learning models cannot currently achieve the accuracy that would be acceptable for the establishment of metrics that leverage this data so more work is required in this domain.

In a very recent paper Smith and Cumberledge [2020] completed an in-depth study into citation errors in high-impact journals. Their research concluded that in some areas, up to 25% of citations may be erroneously citing prior works. During our research we have asked a significant number of citing authors why they have cited a particular work. The completed ACT dataset provides the opportunity to now ask cited authors if they concur with the stated usage of their work and whether it has been cited correctly.

One underlying hypothesis that has driven this research is that the addition of semantic data to citations, specifically purpose and influence, can offer improvements over the current status quo.
where these metrics are purely quantitative. This is however yet to be tested. It will take further work to define how these metrics should leverage this new semantic information. This opens many future questions around the perceived value of particular types of citation and whether one type or another is more ‘valuable’ in terms of impact. Key questions remain around the development of these new metrics and how they can be successfully adopted in publishing, research evaluation and scholarly communication systems.

The models used for the automatic classification of citations are still far from perfect. Even current state-of-the-art studies are yet to achieve results that can be extrapolated to a wider range of domains. There are numerous directions for the improvement of the current status quo. The development of better deep-learning based models for a range of NLP tasks, including citation classification, is an ongoing effort by several teams. Further, studies are now experimenting with the length of the citation context, that is the amount of information fed to the classifier. All of the studies covered in this thesis use the citing sentence as the input. Recent
studies such as that of Lauscher et al. [2021] which introduced Multi-cite are experimenting with adding additional sentences to the citation context in an effort to improve the performance of the classifier.

If one wishes to consider alternatives to the automatic classification of citations, this thesis provides the groundwork for further research in this domain. If citation type can become part of the standard meta-data collected at the point of citing, or point of publication this removes reliance on less than perfect models. There are two future avenues for work in this area. First, the scalability and ease of use of the ACT platform means incorporation with a publishers’ or repositories infrastructure is technically achievable and would provide a simple, fast to use tool for citation annotation at the point of publication. Second, one could start a stage earlier, at the point of authorship. This can be accomplished with the development of an ACT-LaTeX plugin to add citation purpose information as a part of the standard bibliography formatting within LaTeX itself. Another potential avenue for the ACT platform is direct integration with reference management
systems such as Mendeley or Zotero.

6.4 Conclusion

The goal of this thesis was to progress beyond the state of the art in identifying and capturing the semantic aspects of citations. Citation data is used widely as a proxy for quality in areas such as research evaluation, found in research funding and in information retrieval. If reason for citation can become a part of the standard metadata accompanying citations, this opens many further avenues for study and also provides opportunities for enhancing current bibliometrics to use this semantic dimension.

This research has shown that, contrary to previous expectations, bibliometrics based on simple citation indicators are highly correlated with aggregated peer review results from performance-related funding exercises such as the UK’s REF 2014.

Second, as citation data is already being used as a proxy in research evaluation, this research examined what semantic information is contained in the citations themselves. The work found
that many features previously tested for identifying influential citations have little predictive ability, and identified two features that are effective; the number of times a citation is mentioned and a similarity in abstracts between the citing and cited papers.

Third, we identified that previous studies were conducted on far too small datasets, largely drawn from one or two domains at most. Our experimental results show that the language in citations changes significantly across domains and therefore the previous results cannot be directly applied and scaled up across domains. More specifically, when NLP models or deep learning models are used to classify citations according to type, they will be significantly less accurate when applied to a different domain than the one on which they were trained. Applications are currently being built on such models without consideration that the apparent efficacy of the model arises from how the data was prepared and pre-processed in training, rather than from the performance of the model itself.

Finally, the limitations of existing datasets, and the evidence that citations vary between domains, point to the need for context-
sensitive training of models, and for multi-domain research. Our novel approach, employing first authors as annotators, removes the limitation of using experts. This approach has driven the development of a novel platform for collating annotated citations and resulted in a broad new dataset, covering 19 separate disciplines, that can now be applied to further research in this domain.
References


233


Peter Coles. The apparatus of research assessment is driven by the academic publishing industry, 2013. URL https://blogs.lse.ac.uk/impactofsocialsciences/2013/05/14/the-apparatus-of-research-assessment-is-driven-by-the-academic-publishing-industry.


Nicola de Bellis. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Scarecrow Press, 2009.


HEFCE. Research Excellence Framework 2014: Contextual Citation Data, 2014b. URL http://www.ref.ac.uk/2014/background/subguide/citationdata/contextualdata/.


Anne-Sophie Jannot, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V. Perneger. Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology*, 66(3):296–301, 2013.


Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications,*
Vincent Larivière, Yves Gingras, and Éric Archambault. The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, 60(4):858–862, 2009.


Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for


Technopolis. REF Accountability Review: Costs, Benefits and Burden, 2015. URL http://www.hefce.ac.uk/


Vincent A. Traag and Ludo Waltman. Systematic analysis of agreement between metrics and peer review in the UK REF. Palgrave Communications, 5, 2019.


