Quantifying the Impact of Funded Research Works

Sofia Stamou. Paraskevi Tzekou, Nikos Zotos
Computer Engineering and Informatics Department, Patras University 26500 GREECE
{stamou, tzekou, zotosn}@ceid.upatras.gr

Abstract

A number of scientific publications discuss works that have been financially supported by agencies that invest in research. We refer to those publications as funded research works. In this paper, we try to capture the impact of funded research works, in order to assist funding agencies evaluate their efficacy in identifying the best *research efforts. To accomplish our goal, we firstly need a method that automatically identifies publications that correspond to funded research. Our proposed method leverages NLP techniques in order to identify acknowledgments of financial support in the publications' content. Publications containing acknowledgments of financial support are deemed as funded works. Following the identification of funded research articles, we quantify their impact by considering the number of their citations, their freshness as well as the impact of their publication venue. The application of our method to a number of publications reveals that although funded research articles account to nearly 23% of the scientific publications, their average impact is increased compared to the average impact of non-funded publications. Our findings suggest that investments made on research deliver significant results and that funding agencies are effective in judging the potential impact of research efforts.*

1. Introduction

Citation analysis is the most prominent way for estimating the impact of scientific publications. Citation analysis involves counting the number of times a given research publication is cited in the works of others and operates on the assumption that important works will be cited more frequently than less important ones. Currently, citation analysis serves as a global measure for the publications' impact and does not discriminate between works of individuals and works published under some financial support by agencies or organizations. Although such discrimination is needless in capturing the global impact of scientific publications, nevertheless it is useful for enabling funding agencies assess their efficacy in judging the potential influence of research ideas and thus in identifying the best works in a field.

In this paper, we propose an intuitive method for assisting funding organizations evaluate their effectiveness in judging the potential impact of research ideas. Our method relies on the notion that if we could quantify the impact of funded research publications and compare it to the impact of the non-funded works; we could provide funding agencies with valuable evidence for assessing their performance. Towards realizing our goal we essentially need a method for discriminating between funded and non-funded research works. In this respect, we propose processing the publications' contents in order to identify acknowledgments of financial support. Considering that for most funded research, acknowledgments to the appropriate funding agency are requested, we deem that we can accurately determine funded research works simply by relying on the presence of acknowledgments of financial support within their contents. But, identifying acknowledgments of financial support in the publications' contents is a complex task.

Previous studies [2] on acknowledgment analysis determined the following types of support expressed via acknowledgments: (i) moral, (ii) financial, (iii) editorial, (iv) presentational, (v) technical and (vi) conceptual support. Thus, if we want to identify funded research works, we specifically need to detect acknowledgments of financial support within their contents. Towards this goal, we propose the utilization of NLP techniques for processing the acknowledgments' lexical elements in order to determine which of them represent concepts that denote financial support. Acknowledgments of financial support in a publication's contents indicate that the discussed work corresponds to funded research.

Based on the identified funded research works, we derive their impact by examining the number of their citations, the importance of their publication venue and their age, i.e. freshness. The deployment of our method on a number of publications revealed that although funded research works account to 23% of the scientific publications, their average impact is increased compared to the average impact of non-funded research articles. This implies the efficiency

of funding agencies in investing to cutting edge research. Given that the efficacy of funding agencies is not entirely realized by the impact of their funded works, but also by the educational and career opportunities that they provide to researches, our study focuses on a single evaluation aspect of the agencies' performance; that of identifying the most influential works in a field.

The remainder of the paper is organized as follows. We begin with a brief presentation of related works. In Section 3, we introduce our approach for identifying funded research works and deriving their impact. In Section 4, we present the results of an experimental study in which we capture the impact of funded research works. Finally, we conclude the paper in Section 5.

2. Related Work

Related work falls in two main categories, namely citation analysis and acknowledgment analysis. Citation analysis has attracted the interest of many researchers who try to measure the impact of research works in an objective manner [9], [10]. Since the introduction of the Science Citation Index [3] a number of citation analysis tools have emerged, e.g. CiteSeer [15], Scopus [17], etc. that allow automatic extraction and grouping of citations for online research documents. In addition, there have been works that try to evaluate the accuracy of the citation analysis methods in assessing the true value of research works [6], [8]. Another measure widely used in citation analysis is the impact factor [19] which counts the citations from articles of thousands of journals in order to estimate the journals' quality. The commonality in all existing methods and tools is that they rely on the number of citations a scientific publication has received for quantifying its impact.

Although, authors cite the works of others to express intellectual debt, nevertheless they express their appreciation to any support that contributed to the publication of their works via acknowledgments. Acknowledgments are personal statements that most of the times imply debt to some individual or organization that assisted in the realization and/or dissemination of the published work [5]. Despite the potential of acknowledgment analysis in serving as an indicator of influential contributors to scientific works, acknowledgments have not been exploited by major scientific indices mainly because their extraction and processing are tedious and require manual work. As of today, the only systematic effort for extracting and analyzing acknowledgments is reported in [1]. Results showed that combining citation and acknowledgment analysis yields improved impact measurements to the contribution of researchers and their corresponding works.

Motivated by the existing works on acknowledgment analysis and considering that publications pertaining to funded research almost always acknowledge financial support, we conducted the present study in order to quantify the impact of funded research works.

3. Identifying Funded Research Works

To identify which scientific publications discuss funded research works, we rely on the presence of acknowledgments of financial support in the publications' contents. This is because funding agencies request that the works performed under their financial support acknowledge the agencies' contribution in their published scientific materials. Therefore, we recast the problem of detecting funded research works to a classification task in which publications should be categorized as funded or non-funded ones, depending on the type of acknowledgments they contain (if any). To enable classification, though, we firstly need a robust method that detects and extracts acknowledgments from the publications' contents. Additionally, we need to apply lexico-semantic processing to the extracted acknowledgments' contents in order to identify those denoting financial support.

For tackling the first problem, i.e. extract acknowledgments from research articles, one could employ existing information extraction techniques such as the use of regular expressions [1], named entity extraction methods [12], machine learning algorithms, etc. Considering that the aim of our study is not to implement a new acknowledgment extraction technique, but rather to identify acknowledgments of financial support, we decided to adopt the approach introduced in [1] for deriving acknowledgments from research articles. Our decision was made on the grounds that the method in [1] is the only available acknowledgment extraction technique and also it has good accuracy in identifying acknowledgment passages in research articles.

In brief, the adopted method combines regular expressions and Support Vector Machines (SVM) for identifying acknowledgments within scientific publications as well as for extracting their acknowledged entities. Specifically, acknowledgment sections in the publications' contents are identified via the use of regular expressions by searching lines in the text containing only the term acknowledgment in all variations. Upon the detection of such lines, the text that follows them until the next header is extracted. Furthermore to accommodate the case that acknowledgments appear in unlabeled sections of an article, the method of [1] extracts the lines of text in the first and last page of a publication and uses SVM to identify which of the lines contain acknowledgments. Although this method proceeds with the iden-

tification of the acknowledged entities from the extracted passages, in our study we omit this step since our goal is not to derive who gets acknowledged but to estimate the impact of funded research works.

Based on the above method, we extract acknowledgment passages from the contents of research articles. The next step is to deduce which of the acknowledgments express debt to financial support. To judge that, we propose the exploitation of NLP techniques in order to identify the lexical elements in the acknowledgments' contents whose semantics represent financial concepts. In this respect, we start by processing the extracted acknowledgment passages in order to identify their content terms. As content terms we denote those that have been assigned one of the following Part-of-Speech tags: noun, proper noun, verb, adjective and adverb. Passage processing involves applying tokenization, Part-of-Speech tagging, stop-word removal and lemmatization. Then, we rely on the acknowledgments' content terms, which we map to their corresponding WordNet [18] nodes in order to identify whether their semantics represent financial-relevant concepts. To judge that, we extract all the definitions of the content terms found in WordNet and we use regular expressions to locate within their definitions terms that are variants (morphological or derivational) of the words: money, finance and/or fund. Terms containing in their definitions variants of the above words are deemed to represent concepts that relate to finance. Upon the detection of terms denoting finance-related concepts in the extracted passages, we deem that the corresponding acknowledgments express debt to financial support.

To verify the accuracy of our method in identifying acknowledgments of financial support, we relied on a sample of 200 acknowledgment passages extracted from a total set of 432 publications and we manually assessed which of those passages pertain to financial acknowledgments. Of the 200 examined passages, we have annotated 109 of them as representing acknowledgments of financial support. Then, we applied our method to the 200 passages and compared the acknowledgments it identified as expressive of financial support to the financial acknowledgments that have been manually determined. The comparison of the results revealed that our method achieved 88.9% recall and 97.9% precision, which practically indicates our method's accuracy in identifying acknowledgments of financial support in the articles' contents.

The manual examination of the financial acknowledgments that our method failed to recognize as such reveals that they did not contain terms with a clear financial orientation. For example the phrase "*This work is part of the EC-ICT xxx project*" does not explicitly state that the work has been funded. However, encapsulating the term *project* among the indicators of financial acknowledgements would lead to mis-classification of non-funded works as funded ones, as the following acknowledgment example demonstrates. "*We thank xxx for giving us access to the xxx project results*". Therefore, we decided to focus our financial acknowledgment identification method on concepts with a clear financial orientation that can be verified against a rich semantic resource such as WordNet.

In Table 1, we list the terms extracted from acknowledgment passages upon which our method determined credit to financial support.

Table 1. Acknowledgment terms of financial support.

Term	WordNet definition
Financial	Involving financial matters
Grant	Monetary aid
Scholarship	Financial aid to a student
Expense	Money spent to perform work
Sponsor	Support materially or financially
Funded	Furnish money for

Based on the above steps, we classify research works as funded or non-funded, depending on the presence of acknowledgments of financial support within their contextual elements. Following the identification of funded research articles, we proceed with the estimation of their impact.

3.1. Impact of Funded Research Works

To estimate the impact of funded research publications, we rely on the combination of the following evidence: the number of citations that the publications have received in the works of others, the publications' freshness (i.e. age) and the importance of their publication venue. More specifically, to compute the citations of an article we rely on a citation index (e.g. CiteSeer, Scopus, etc.) from which we obtain the complete bibliographic entry of all the publications that cite the article under examination. Then, we remove self-citations from the complete set of references based on the overlapping author names and/or affiliation details between the examined article and its citations in the *bibtex* entry. The reason for eliminating self-citations from our computations is because self-references are mainly used for emphasizing the authors' personal contribution to a piece of research; therefore they strengthen the authors' credibility and not the publications' impact.

Based on the above, we estimate the citations C of a research publication p as follows:

$$C(p) = |\#\text{of citations}| - |\#\text{of self - citations}|$$

Moreover, to derive the freshness (i.e. age) of an article we rely on its publication date and we subtract it form the current date. The reason for considering

the freshness of a publication for quantifying its impact is because long-existing publications might have more citations compared to recently published ones, without necessarily being more influential than the recent ones.

A final aspect we encapsulate in our impact measurement is the importance of an article's publication venue (i.e. conference or journal). To quantify how important or else prestigious is a publication venue, we rely on two different sources, namely the impact factor [4] when dealing with journal articles and the conference impact ratings when dealing with conference articles. Note that there exist numerous online sources reporting the publication venues impact (e.g. [16]) some of which pertain to specific disciplines (e.g. Computer Science Conference Rating).

Based on the above evidence, we formally quantify the impact of a publication as follows. Our impact metric is inspired by the suggestion in [7].

$$\text{Impact}(p) = \frac{|C(p)|}{|\text{Age}(p)|} \text{Rank}(v, p)$$

Where $C(p)$ indicates the number of citations that p has received in the publications of other, $Age(p)$ indicates the number of years that p exists and $Rank(v, p)$ gives the Ranking (e.g. impact) of the publication venue v where p has been published, as the latter is determined by external resources and by considering the average citation rate in a given period of time for all the articles published in v. At this point we should refer to a relatively recent proposal for measuring the quality of conferences [13] that could replace the value of $Rank(v, p)$ in our formula. We defer the examination of this issue for a future study.

Based on the above formula, we can evaluate the impact of funded research works and assist funding agencies assess their effectiveness in identifying the best research efforts.

4. Discussion

To quantify the impact of funded research works, we applied the acknowledgment extraction algorithm to a number of scientific articles pertaining to the discipline of computer science that we collected from DBLP [14]. Specifically, to gather our dataset of scientific publications we issued the query "web" to DBLP and we selected a set of 17,651 distinct research publications from the 35,123 returned results (as of June 2009).

Our selection took place on the requirement that the research publications should be journal, conference or workshop papers. Based on the set of publications that met the above criterion, we downloaded their full articles (note that for some publications we could not access their contents due to the lack of subscription to their hosting digital libraries) and we converted their files to a processable format (i.e. *rtf*). We then manually assessed their conversion accuracy and retained a subset of 17,651 publications that have been correctly converted from *.pdf* to *.rtf* format. Note that the manual assessment of conversion accuracy did not involve reading through the entire article but rather checking the header/section titles, the number of paragraphs, sections, etc.

Having decided on the research papers the impact of which we would examine, we performed data cleaning in order to retain a single version for every examined publication and eliminate article citations in languages other than English and/or self-citations. Thereafter, we processed the publication contents in order to extract their respective acknowledgment passages. Of all the examined articles, 6,985 were found to contain acknowledgments. Furthermore, the distribution of acknowledgments in the publication contents indicates that 79.36%of the acknowledgments are found in labeled sections that usually appear right before the references or the first appendix, while 20.64% of the acknowledgments are found in unmarked sections, i.e. within footnotes.

Based on the extracted acknowledgment passages, our next step was to parse the passage contents in order to identify which of them express debt to financial support. Based on the method presented in Section 3, we identified 4,035 acknowledgments of financial support, indicating that nearly 23% of the considered publications correspond to funded research works. Another observation from our dataset is that the majority of acknowledgments that publications contain express debt to financial support. Based on the two sets of publications (i.e. 13,616 non-funded and 4,035 funded research articles), we computed their impact as previously described.

Figures 1 and 2 illustrate the impact of the examined non-funded and funded publications respectively. In both figures, the y-axis indicates the publications' impact values and the x-axis indicates the number of publications that share the respective impact values.

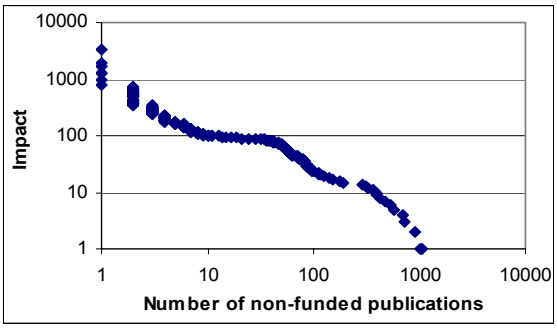

Figure 1. Distribution of impact to the examined non-funded research publications.

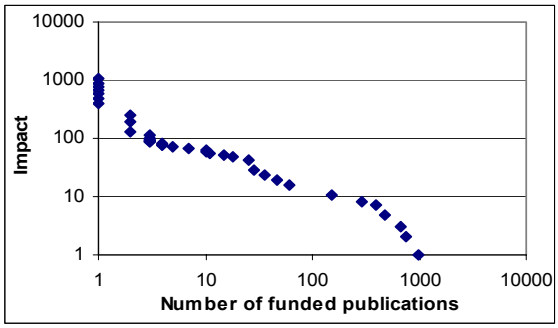

Figure 2. Distribution of impact to the examined funded research publications.

As both figures show only a few papers are deemed as influential by other researchers, while the majority of them are not considered that influential. Our findings confront to previous studies on the distribution of citations [11] and imply that from all scientific publications (either funded or non-funded) a few have a significant impact to the research community, while the majority of them are rarely cited.

However, to assess whether the impact of funded research articles is proportional to their size, we computed their average normalized impact and compared it to the average normalized impact of the non-funded articles. For estimating the average normalized impact of the examined publications *P* (funded and non-funded respectively), we relied on the following formula:

$$\text{Impact}_{\text{avg}}(p) = \sum_{p=1}^{P} \text{Impact}(p) / |P|$$

Table 2 reports the comparison results. Our findings suggest that funded research works have increased average normalized impact compared to the non-funded ones, thus we may conclude that funding agencies are quite effective in assessing the significance of the research efforts they evaluate.

Table 2. Average impact of scientific publications.

Type of Publication	Average Impact
Non-funded	4.63
Funded	5.94

5. Concluding Remarks

We have presented a preliminary study that tries to capture the impact of funded research works in order to assist funding agencies evaluate their efficacy in assessing the contribution of research ideas. The evaluation of our method on a set of scientific publications reveals that although funded research works account nearly to 23% of the scientific publications, still their average impact is increased compared to the average impact of non-funded research articles. The deployment of our method over larger datasets and across disciplines can give useful insight to funding agencies for assessing their efficacy in evaluating the best research efforts.

Some areas for future research involve examining the correlation between the impact of publications and the impact of their publication venues as well as investigating the coherence between the subject of funded research articles and the subject of their supporting grants. In addition, it would be interesting to examine the contextual similarities between publications and their citations in order to identify the impact of the latter as well as the evolution of research disciplines over time. Last but not least, we could apply our publications; impact measure for capturing the productivity and the contribution of researches within and across their fields of study.

6. References

[1] I.G. Councill, C.L. Giles, H. Han, and E. Manavoglu, "Automatic Acknowledgment Indexing: Expanding the Semantics of Contribution in the CiteSeer Digital Library", In *Proceedings of the International Conference on Knowledge Capture*, 2005, pp. 19-26.

[2] B. Cronin, D. Shaw, and K. LaBarre, "A Cast of Thousands: Co-Authorship and Sub-Authorship Collaboration in the 20th Century as Manifested in the Scholarly Journal Literature of Psychology and Philosophy", *Journal of the American Society of Information Science and Technology*, vol.54, 2003, pp. 855-871.

[3] E. Garfield, "Science Citation Index: a New Dimension in Indexing", *Science*, vol. 144, 1964, pp. 649-654.

[4] E. Garfield, "Citation Analysis as a Tool in Journal Evaluation", *Science*, vol. 178, 1972, pp. 471-479.

[5] C.L. Giles, and I.G. Councill, "Who Gets Acknowledged: Measuring Scientific Contributions through Automatic Acknowledgment Indexing", *National Academy of Sciences*, vol. 101, 2004, pp. 17599-17604.

[6] K. Hyland, "Self-Citation and Self-reference: Credibility and Promotion in Academic Publications" *National Academy of Sciences*, vol. 102, 2003, pp. 16569-16572.

[7] Y. Liu, K. Bai, P. Mitra, and C.L. Giles, "TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries", In *Proceedings of the 7th International Joint Conference on Digital Libraries*, 2007, pp. 91-100.

[8] K.L. Meho, "The Rise and Rise of Citation Analysis". In Physics World. CoRR abs/psysics/0701012, 2007.

[9] H.F. Moed, *Citation Analysis in Research Evaluation*. Springer, Dordrecht, the Netherlands, 2007.

[10] E. Rahm, and A. Thor, "Citation Analysis of Database Publications". In *SIGMOD Record*, vol. 34(4), 2005, pp. 48-53.

[11] S.Render, "How Popular is your Paper? An Empirical Study of the Citation Distribution". In *European Physics Journal*, vol.4, 1998, pp. 131-134.

[12] K. Taceuchi, and N. Collier, "Use of Support Vector Machines in Extended Named Entity Recognition", In *Proceedings of the 6th International Conference on Natural Language Learning*, 2002 pp. 1-7.

[13] Z. Zhuang, E. Elmacioglu, D. Lee, and C.L. Giles, "Measuring Conference Quality by Mining Program Committee Characteristics", In *Proceedings of the 7th Joint Conference on Digital Libraries*, 2007, pp. 225-234.

[14] DBLP Data: http://kdl.cs.umass.edu/data/dblp/ dblp-info.html

[15] CiteSeer: http://citeseer.ist.psu.edu/

[16] CiteSeer Impact Ratings: http://citeseer.ist.psu.edu/impact.html

[17] Scopus: http://www.scopus.com

[18] WordNet: http://wordnetweb.princeton.edu

[19] M. Amin, and M. Mabe, *Impact Factor: Use and Abuse*. Perspectives in Publishing, 2000.