


Making sense of citations

Xenia Koulouri¹, Claudia Ifrim², Manolis Wallace¹, and Florin Pop²

¹  Knowledge and Uncertainty Research Laboratory
Department of Informatics and Telecommunications
University of the Peloponnese, Tripolis, Greece 22 131
`xenia.koulouri@uop.gr`, `wallace@uop.gr`
`http://gav.uop.gr`

² Faculty of Automatic Control and Computers
Computer Science Department
University Politehnica of Bucharest, Bucharest, Romania 060042
`claudia.ifrim@hpc.pub.ro`, `florin.pop@cs.pub.ro`
`http://acs.pub.ro/`

Abstract. To this day the analysis of citations has been aimed mainly to the exploration of different ways to count them, such as the total count, the h-index or the s-index, in order to quantify a researcher's overall contribution and impact. In this work we show how the consideration of the structured metadata that accompany citations, such as the publication outlet in which they have appeared, can lead to a considerably more insightful understanding of the ways in which a researcher has impacted the work of others.

Keywords: Research Impact, Citations, Publication Medium

1 Introduction

Academia is competitive by nature. Researchers strive to make the greatest breakthrough, attract the most lucrative funds, obtain the highest awards and achieve the greatest impact. More so now than before, this effort is not related just to excellent accomplishment but even to the very survival of the researcher in academia. Of course, this keep achieving or be ignored, or “publish or perish” as it is usually referred to, approach has its downside. The more a researcher's output is linked to their survival, the more bias they will have in assessing and presenting their work [1].

As a result, we have now reached a point where publications, once the best indicator of the value of a researcher's work [2], need to be examined with a grain of salt [3][4]. Publication records can be skewed in size [5], by publishing multiple similar papers, by splitting one work in multiple incremental publications, by exchanging gratuitous co-authorships, by repeating the same work with different datasets [6] etc, as well as in direction, by carefully selecting titles and masterfully penning abstracts to highlight relevance to one scientific field or another [7].

It is then only normal that we look not only at the publications of researchers, but increasingly also at their impact, as shown by their citations [8]. Of course, citations can also be skewed [9]. In fact, it has already been discussed that the way citations are currently examined is not sufficient [10]. In this paper we look deeper into citations, taking advantage of citations' metadata in order to achieve a better understanding and quantification of researchers' impact. Specifically, we focus on the publication medium in order to best estimate the fields of science that each work impacts.

A paper discussing similar ideas but focusing mainly on the visualization of the results has been presented at the 9th International Workshop on Semantic and Social Media Adaptation and Personalization [7]. A broader paper incorporating some of the ideas of the current work but focusing mainly on the presentation of an integrated working system is currently under consideration for publication in a special issue on "Keyword Search in Big Data" in the LNCS Transactions on Computational Collective Intelligence journal.

The remainder of this paper is organized as follows: In section 2 we discuss existing approaches to the assessment and quantification of scientific impact. Continuing, in section 3 we discuss the types of information that can be mined from citation metadata and in section 4 we present a comprehensive methodology that uses this notion in order to achieve a deeper insight of the way in which each work and each researcher impacts the scientific world. Finally, in section 5 we present and discuss some indicative results from the application of our approach and in section 6 we list our concluding remarks.

2 Counting citations

To this day citations are used to assess scientific impact. There are of course inherent weaknesses [11]; it is possible revolutionary works to go un-noticed due to random shifts of research trends or less deserving works to receive attention simply because of an inspired title [12]. Still, the fact that they are fully quantitative measures that can be computed in an automated manner with little or no human intervention makes them the measure of choice for the estimation of scientific impact.

Thus, a paper's impact is quantified as the count of citations it has received from the day it was published and up to the day of examination. This, of course, favors papers that were published many years ago, as they have been accumulating citations for a longer period of time. This is not necessarily a weakness of the measure; it is only natural that works that have been around for a longer period of time have had the opportunity to have a greater impact on the works of others. Besides, it has been observed that the yearly count of citations received by a paper diminishes after a few years; so, after some time, the advantage of earlier papers is diminished.

Similar ideas are applied towards the evaluation of the scientific value of a publication medium, such as a journal, magazine or conference. There is, though, an important difference originating in the way to use the results of this evalua-

tion. Journals are not evaluated in order to assess which one has had the greatest overall impact on the scientific world. To the contrary, the goal is to assess the probability that an article published in a journal will make an impact in the future; readers consider this evaluation to select the journals to read and more importantly authors consider it in order to select the journals to submit to, thus maximizing the potential of their work. Therefore, the number of years that a journal has been publishing, or even the number of volumes per year or the number of articles per volume cannot be allowed to affect the evaluation.

The impact factor (IF) is the most trusted quantification of a journal's scientific potential. It is computed as the average count of citations articles published in the journal receive in the first two years after their publication; some limitations apply regarding the sources of these citations. It is clear to see that the IF is configured in a way that favors journals that publish carefully selected high quality articles, which is in accordance with the goals of journal evaluation. Of course, the impact factor is also an imperfect measure [13] and efforts are made to improve it [14].

When it comes to researchers, their past impact, and by extension their future potential, is also assessed based on citations. The first, most common and straightforward approach is the consideration of the cumulative number of citations an author has received for the complete list of their published work.

But given the highly competitive nature of the scientific community, it is rather expected that the prime tool to assess and compare researchers has received a lot of attention, both in the form of criticism of its objectivity and in the form of attempts to affect its outcomes. Numerous weaknesses have been identified, related to the number of years of activity, the effect of cooperation networks, self-citations, outlier works, frequency of publication etc.

In order to deal with the weaknesses of the count of citations as a metric, a long list of more elaborate metrics have been proposed, including the average number of citations per paper, the average number of citations per author, the average number of citations per year, the h -index [15] and similar indices [16][17][18], the g -index[19], the e -index[20], the s -index[21], the i -10 index, and more.

3 Citation context

The count of citations, as well as all the other aforementioned measures that are based on it, provide a numerical quantification of impact, without any indication of where that impact has been made. This does not align well with the purpose of assessing a researcher's impact. When researchers are evaluated, for example for an academic position, only relevant publications from their publication record are considered. Still, when it comes to impact, we use the overall citation count without examining which publications they have derived from or which scientific fields they show impact in. Clearly, it would be useful to have access to such information.

In this work we examine the scientific scope of the referencing papers in order to see which fields of science have been affected by a given paper. Our goal is to describe a way to mine more information from citation records, without losing the objectivity of the citation count, i.e. the fact that it is not directly affected by the examined researchers and it is computed with minimal user intervention. The practical question here of course is which of the citations' metadata to use and how in order to identify the scientific scope. In developing our approach we should, of course, also consider the availability of the data that will be examined.

In existing systems papers are indexed by their titles and journals they are published in; authors are indexed by the papers they have published and the keywords they use to characterize their own research interests [22]. But such metadata (titles, journals to submit to, keywords, abstract) are determined by the authors based on a priori preferences and not all are necessarily closely related to the a posteriori information regarding the actual areas that their work has an actual impact on, or even to the content of the work itself. Titles can be misleading; keywords are useful but are not standardized and are not used in all publications; textual analysis is not yet mature enough to guarantee reliable results when applied on abstracts that may be related to literary any given scientific field. More importantly, all of the above can be severely skewed by the authors, especially when they need to build a profile that shows strength in a specific field.

In contrast, the publication medium can provide a good indication of the scientific scope. When a paper is considered, either by a journal or by a conference, thematic relevance is examined together with its scientific quality. Therefore, the editorial process guarantees that, for example, papers published in the IKC conference are additionally related to semantic keyword-based search on structured data sources. Almost all edited publications come with clearly defined scopes and lists of relevant topics, and for those that do not it is relatively easy to produce them manually since this would need to be done only once for each publication medium and not separately for each article. Therefore, the automated and objective (i.e. without considering the subjective opinion of a human expert examining the specific article) consideration of the scientific scope of a given published paper is feasible.

Our approach is to examine each citation's publication medium in order to estimate the scientific field in which it indicates impact and to use this information in order to classify citations to fields and transform the unidirectional citation count - and by extension all similar metrics - into a field by field analysis which will provide much deeper insight in the way a researcher's work has impacted the rest of the scientific world.

4 Methodology

As we have already explained, our analysis is based on the examination of the publication medium. In the next paragraphs we outline the main steps required to put this notion in practice.

4.1 Preparatory steps

The preparatory steps involve the establishment of the knowledge base that is required for the execution of the processing steps, as follows:

1. Develop a list of thematic areas
2. Compile a list of publication media (journals, magazines, conferences)
3. Assign thematic areas to each publication medium

Thematic areas. The list of scientific fields is almost static. Therefore a reasonable first step is to acquire this hierarchy. Existing hierarchies exist that may be considered as a basis, as for example the one found in [23].

Publication media. We can use, for example, DBLP metadata in order to acquire a first list of previous and running journals and conferences, knowing that although this list is long it is far from complete. A comprehensive list of publication media is not easy to establish. Moreover, the list is not static as some conferences disappear whilst new ones appear every year; there are similar changes to the list of journals, but they are less frequent and thus easier to tackle.

Therefore the pre-processing step regarding the acquisition of publication media is not meant to produce a complete and finalized list but rather to facilitate the initiation processing steps by dealing with the problem of cold start.

Medium to area assignments. Although the DBLP metadata are carefully curated, they do not contain semantic information regarding the thematic scope of the included publication media, other than their title. This title is often, but not always, enough to have a rough idea of the thematic coverage.

In order to overcome this a semi-automatic approach is needed.

4.2 Processing steps, for each work

For each considered article, we need to examine the list of citations as follows:

1. Acquire the list of citations
2. Identify the thematic area of each citing work
3. Aggregate findings

List of citations. We use can use Google Scholar or any other similar system to acquire a comprehensive list of citations for each article that we examine. Of course such systems are neither complete nor perfect (they inherently contain false positives, incorrectly assigned fields, damaged titles, repetitions etc). Still, although error rates are high (often exceeding 20%), the deviation is small. Thus citations retrieved from systems such as Google Scholar are a relatively reliable source given that the error rate is similar for different articles and authors [24].

Thematic area of each citation. In earlier sections we have explained that we will use the publication medium to identify the thematic scope, we have developed the lists of publication media and scopes and established the associations between publication media and thematic areas. In the previous paragraph we also saw how the publication medium is acquired as a separate field in an XML document.

Thus, the connection between citing article and its thematic areas is quite straightforward.

Aggregated impact for each work. Conventionally all citations associated with a published work are considered equally and uniformly, and overall impact is given as the count of citations. Given the additional thematic information that now becomes available, a rising question is the validity of considering uniformly references that have been published in a publication medium with an impact on a single science and references whose publication medium influences more than a single science. Our approach is a variable weighting factor for the two cases. In case that the papers influence a single scientific field weight will be equal to 1, whereas the weight will be distributed uniformly when multiple fields are impacted.

The aggregated impact for each work is given as the sum of weights, for each scientific field; as expected the impact is not calculated as a single number but rather as an array of numbers, one per field.

4.3 Processing steps, for each author

For each considered article, we examine the list of public works as follows:

1. Acquire the list of published works
2. Identify the impact of each work
3. Aggregate results

In the conventional approach, an author's citation count is calculated as the sum of citations for all of the author's published works. By extension, in our work we calculate an author's impact in each field as the sum of the impact values for that field for all of the author's works. Thus, the aggregated impact for the author is a vector calculated as the sum of the impact vectors of all of the author's published works, as calculated above.

5 Experimental results

In order to better explain what type of insight we are looking at, in this section we examine what our approach brings to light when applied for three specific researchers, namely Prof. Ioannis Anagnostopoulos, Prof. Costas Vassilakis and Prof. George Lepouras. The results have been produced using an early software implementation of the notions presented earlier herein [25][26].

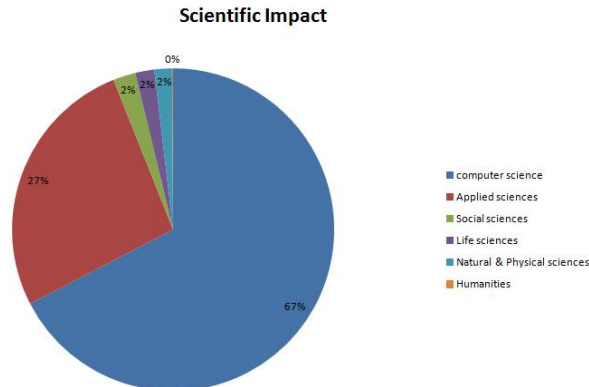


Fig. 1. Anagnostopoulos - Scientific impact

5.1 Ioannis Anagnostopoulos

Ioannis Anagnostopoulos is a member of the Department of Computer Science and Biomedical Informatics, at the University of Thessaly. The position in which he has been elected faculty member, his expertise as presented in his CV and his research interests as presented in his personal web page are focused in the analysis of social networks.

Naturally, one would expect the impact of his work to be in the same area. Still, our analysis finds that 27%, of his research impact does not even lie in the field of computer science.

Looking into the details of the researcher's publications and citations we find that Prof. Anagnostopoulos worked in the fields of neural networks and image processing at the beginning of his career and much of his citation record comes from citations to work of that era. And whilst in examining Prof. Anagnostopoulos's CV we would quickly filter out these publications when evaluating him for his current position, using the conventional approach we would not have been able to similarly filter the 27% of his citations that are not relevant.

5.2 George Lepouras Costas Vassilakis

George Lepouras and Costas Vassilakis are members of the Department of Informatics and Telecommunications, at the University of Peloponnese. Prof. Lepouras's area of research, as indicated by position in which he has been elected faculty member, his expertise as presented in his CV and his personal statement in his personal web page lie in the field of human computer interaction. In similar fashion we can see that Prof. Vasilakis's area of research lies in the field of information systems.

Clearly, the two researchers have quite distinct works. Yet, our impact analysis shows not only that they have impact in the same broader scientific areas

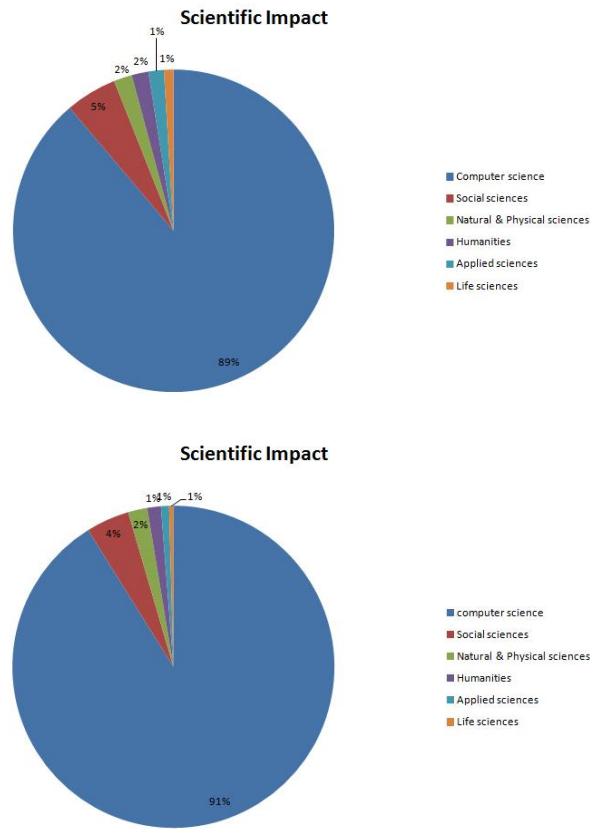


Fig. 2. Lepouras and Vassilakis - Scientific impact

but also that they have very similar impact when examining detailed subfields of computer science. Whilst in the conventional approach we would consider their impact to lie in distinct areas, and more specifically in the areas that they state as their fields of expertise, our closer analysis of their citation records reveals that this would not have been accurate.

6 Conclusions

In this paper we explored the information that can be extracted from citation records' metadata. In order to avoid subjectivity in the estimation and quantification of the impact we have opted to avoid author defined parameters and have instead focused our analysis on the journal or conference where a citing article has been published. This provides an objective and reliable indication

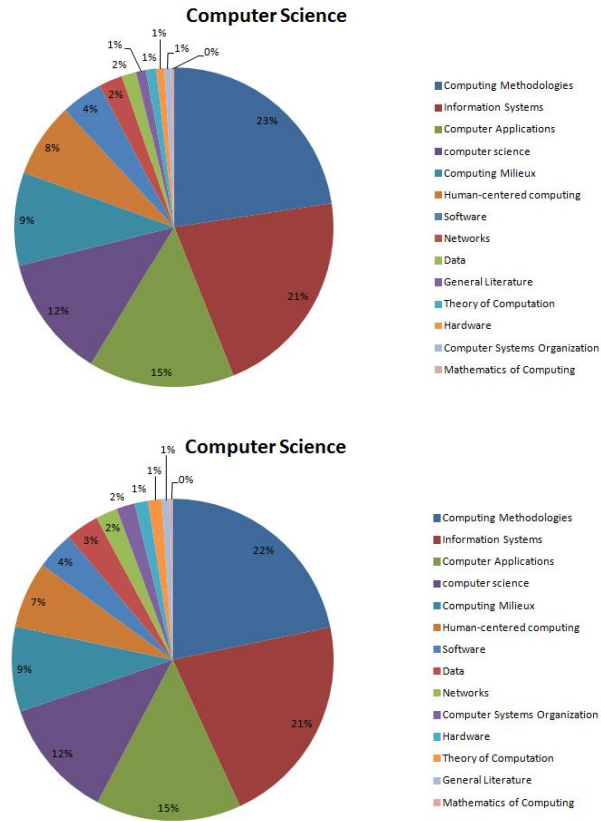


Fig. 3. Lepouras and Vassilakis - Scientific impact in computer science

of thematic scope, which allows us to see, in a semi-automated manner, which scientific areas have been affected by an author’s work.

Through some real life examples we have shown that the approach proposed herein is indeed able to provide a deeper insight into the ways in which a researcher, or even a specific paper, has impacted the scientific work. This can allow for a more fair consideration of citations in the comparative evaluation of researchers, by considering only “in scope” citations, as is also done for publications. Additionally, since our approach effectively partitions citations into thematic areas, any and all conventional citation metrics can still be applied on top of it; for example it is easy to see how to calculate the *h*-index per thematic area.

Of course our work is not complete. We have only just scratched the surface of the treasures hidden in the metadata of citation records. Moving forward, it would be interesting to examine in further detail the different types of impact that can be defined based on the distribution of the areas of impact [7] or to

explore how impact can be redefined or refined by considering not only one but multiple hops in the citation graph.

Acknowledgments

This work has been partially supported by COST Action IC1302: Semantic keyword-based search on structured data sources (KEYSTONE).

References

1. U.S. Neill, *Publish or perish, but at what cost?*, Journal of Clinical Investigations, vol 118(7), pp. 2368-2368, 2008. doi:10.1172/JCI36371
2. R. E. Steinpreis, K. A. Anders, D. Ritzke, *The impact of gender on the review of the curricula vitae of job applicants*, Sex Roles vol 41(7/8), pp.. 509, 1999.
3. D. Fanelli, *Do Pressures to Publish Increase Scientists Bias? An Empirical Support from US States Data*, in E. Scalas (editor) PLoS ONE vol 5(4), 2010. doi:10.1371/journal.pone.0010271.
4. F. Song, S. Parekh, L. Hooper, Y.K. Loke, J. Ryder, A.J. Sutton, C. Hing, C.S. Kwok, C. Pang and I. Harvey, *Dissemination and publication of research findings: An updated review of related biases*, Health technology assessment, vol 14 (8), 2010). doi:10.3310/hta14080. 20181324
5. W. Broad, *The publishing game: Getting more for less* Science, vol 211(4487), pp. 11371139, 1981. doi:10.1126/science.7008199.
6. M. N. Kumar, *A review of the types of scientific misconduct in biomedical research*, Journal of Academic Ethics, vol 6(3), pp. 211228, 2008 doi:10.1007/s10805-008-9068-6.
7. M. Wallace, *Extracting and visualizing research impact semantics*, Proceedings of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization, Corfu, Greece, 2014.
8. M. van Wesel, *Evaluation by Citation: Trends in Publication Behavior, Evaluation Criteria, and the Strive for High Impact Publications* Science and Engineering Ethics, vol 22 (1), pp.199225, 2016. doi:10.1007/s11948-015-9638-0
9. S. M. McNab, *Skewed bibliographic references: Some causes and effects*, CBE Views, vol 22, pp. 183-185, 1999.
10. H. F. Moed, *Citation Analysis in Research Evaluation*, Springer Netherlands, 2005.
11. A. Figa-Talamanca, *Strengths and weaknesses of citation indices and impact factors*, Quality assessment in higher education, pp. 83-88, 2007.
12. A. Letchford, H.S. Moat and T. Preis, *The Advantage of Short Paper Titles*, Royal Society Open Science (The Royal Society, 2015), 150266 <http://dx.doi.org/10.1098/rsos.150266>
13. D. Colquhoun, *emphChallenging the tyranny of impact factors*, Nature, vol 423(6939), 2003, pp.479.
14. R. Mutz, H.-D. Daniel, *Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor*, Journal of Informetrics, vol 6 (2), pp. 169-176, 2012.
15. J. E. Hirsch, *.An index to quantify an individual's scientific research output*, Proceedings of the National Academy of Sciences of the United States of America 102, no. 46, pp. 16569-16572, 2005.

16. P.D. Batista, M.G. Campiteli, O. Konouchi and A.S. Martinez, *Is it possible to compare researchers with different scientific interests?*, *Scientometrics*, vol. 68, no. 1, pp. 179-189, 2006.
17. O. von Bohlen und Halbach, *How to judge a book by its cover? How useful are bibliometric indices for the evaluation of "scientific quality" or "scientific productivity"?*, *Annals of Anatomy* no. 193(3), pp. 191196, 2011
18. L. Bornmann, R. Mutz and HD. Daniel, *The h index research output measurement: Two approaches to enhance its accuracy*, *Journal of Informetrics* no. 4(3), pp. 407-414, 2010.
19. L. Egghe, *Theory and practise of the g-index* *Scientometrics* no. 69(1), pp. 131-152, 2013.
20. C.T. Zhang, *The e-index, complementing the h-index for excess citations*, *PLoS ONE*, no 5(5), 2009.
21. Z.K. Silagadze, *Citation entropy and research impact estimation*, *Acta Physica Polonica*, B41, pp. 23252333, 2009.
22. C. Ifrim, F. Pop, M. Mocanu and V. Cristea, *Agile DBLP: A Search-based Mobile Application for Structured Digital Libraries*, J. Cardoso, G.J. Houben, F. Guerra, A.M. Pinto, Y. Velegrakis (Eds), *Proceedings of the 1st KEYSTONE Conference*, *Lecture Notes in Computer Science*, Springer, 2015
23. W. Glanzel and A. Schubert, *A new classification scheme of science fields and subfields designed for scientometric evaluation purposes*, *Scientometrics*, no.56(3), pp. 357-367, 2003.
24. X. Koulouri, *Estimation of the area of scientific impact through the analysis of citation records*, MSc thesis, Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, 2016
25. N. Babetas, *Automation of the research impact estimation process*, BSc thesis, Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, 2015
26. G. Dimitriou, *Research area estimation and visualization*, BSc thesis, Knowledge and Uncertainty Research Laboratory, University of the Peloponnese, 2015